PROPOSAL FOR TEI-CONFORMANT ENCODING OF BASIC GRAMMATICAL TAGSET

D. Terence Langendoen

3.9.91

This is an attempt to render the tagset proposed for the British National Corpus proposed by Geoffrey Leech as a set of entities representing the feature names and values proposed in TEI AI1 W2, insofar as that is possible. Following each proposed tag, I provide my best guess as to the feature names and values from AI1 W2 to be identified with it. Where extensions are needed to those given in that document, they are specifically noted below. A sample full rendering in feature-structure notation follows, along with some notes about how to construct the necessary entity definitions.

ADJ adjective (unmarked) (e.g. GOOD, OLD)
category=adjective

ADJC comparative adjective (e.g. BETTER, OLDER)
category=adjective, degree=comparative

ADJS superlative adjective (e.g. BEST, OLDEST)
category=adjective, degree=superlative

ADV adverb (unmarked) (e.g. OFTEN, WELL)
category=adverb

ADVC comparative adverb (e.g. OFTENER, LONGER)
category=adverb, degree=comparative

ADVQ wh-adverb (e.g. WHEN, HOW, WHY)
category=adverb, function=(interrogative | relative)

ADVS superlative adverb (e.g. FURTHEST, LONGEST)
category=adverb, degree=superlative

ALPH alphabetical symbol (e.g. A, B, c, d)
symbol=alphabetical [1]

[1] The feature "symbol" and its possible values are extensions of AI1 W2. I assume category information is not relevant for the encoding of alphabetical systems. If one wished to represent these also as nouns, then add: category=noun.

ART article (e.g. THE, AN)

category=article

CONJ subordinating conjunction (e.g. ALTHOUGH, WHEN)

category=subordinator

COORD coordinator (e.g. AND, OR)

category=coordinator

CTHAT the conjunction THAT

category=subordinator, lemma=that [2]

DET determiner (e.g. THESE, SOME)

category=adjective, definiteness=(definite | indefinite) [3] [4]

DETQ wh-determiner (e.g. WHOSE, WHICH)

category=adjective, definiteness=(definite | indefinite) function=(interrogative | relative) [5]

EXIS existential THERE

category=pronoun. lemma=there

GEN the genitive morpheme 'S or '

form=enclitic, lemma=& apostrophe;s [6]

ISOL interjection or other isolate (e.g. OH, YES, MHM)

category=interjection

NEG the negative NOT or N'T

category=adverb, polarity=negative

NOUN noun (neutral for number) (e.g. AIRCRAFT, DATA)

category=noun

NOUPL plural noun (e.g. PENCILS, GEESE)

category=noun, number=plural

[2] The "lemma" feature is an extension to AI1 W2. It can be considered to be a word-level feature (i.e., appropriate for any word of any type). Its value is a canonical spelling of the word or morpheme. To be fully precise, all other subordinating conjunctions should be specified as lemma =/= that (using the f.s.not tag), but that is a nicety that can perhaps be ignored. Similar remarks apply to the encoding of prepositions other than OF.

[3] The "definiteness" and "function" features were inadvertently left out of the list of possible features for adjectives.

[4] Alternatively, the possible values for the feature "category" could be extended to include "determiner", in which case, the result would be: category=determiner.

[5] If an appropriate feature-structure declaration were present, we could replace these (and other) disjunctions by "any".

[6] No category information is needed here.

NOUSG singular noun (e.g. PENCIL, GOOSE)

category=noun, number=singular

NUM cardinal numeral (e.g. 3, FIFTY-FIVE, 6609) (excluding ONE)

category=(noun | adjective), numeral=cardinal lemma =/=one

OF the preposition OF

category=preposition, lemma=of

ONE the word ONE (including numeral and non-numeral uses)

category=(pronoun | adjective), lemma=one

ORD ordinal (e.g. SIXTH, 77TH, LAST)

category=adjective, numeral=ordinal

PART adverb particle (e.g. UP, OFF, OUT)

category=particle

PERS personal pronoun (e.g. YOU, THEM)

category=pronoun

PNOUN proper noun (e.g. LONDON, MICHAEL, MARS)

category=noun, proper=+

POSS possessive form (e.g. YOUR, THEIRS)

category=pronoun, possessive=+

PREP preposition (except for OF) (e.g. FOR, ABOVE, TO)

category=preposition

PROI indefinite pronoun (e.g. NONE, EVERYTHING)

category=pronoun, type=indefinite

PROQ wh-pronoun (e.g. WHO, WHOEVER)

category=pronoun, function=(interrogative | relative)

REFL reflexive pronoun (e.g. ITSELF, OURSELVES)

category=pronoun, anaphora=reflexive

TOINF infinitive marker (e.g. TO, IN ORDER TO)

category=preposition, prep-type=infinitive [7]

UNCL "unclassified" items which are not words of the English lexicon or do not belong to any recognized category. E.g.: formulae, such as XX61, MARKn; foreign words; BOTH when correlative with AND; etc.

category=unknown

[7] The feature "prep-type" and its value are extensions.

VBEB the base forms of the verb “BE”, i.e. BE, AM, ARE

category=verb, verb-type=copula

VBED past form of the verb “BE”, i.e. WAS, WERE

category=verb, verb-type=copula, tense=past

VBEG -ing form of the verb “BE”, i.e. BEING

category=verb, verb-type=copula, verb-form=present-participle

VBEN past participle of the verb “BE”, i.e. BEEN

category=verb, verb-type=copula, verb-form=past-participle

VBEZ -s form of the verb “BE”, i.e. IS, ’S

category=verb, verb-type=copula, tense=present

VDOB base form of the verb “DO”, i.e. DO

category=verb, verb-type=auxiliary, lemma=do

VDOD past form of the verb “DO”, i.e. DID

category=verb, verb-type=auxiliary, tense=past, lemma=do

VDOG -ing form of the verb “DO”, i.e. DOING

category=verb, verb-type=auxiliary, verb-form=present-participle, lemma=do

VDON past participle of the verb “DO”, i.e. DONE

category=verb, verb-type=auxiliary, verb-form=past-participle, lemma=do

VDOZ -s form of the verb “DO”, i.e. DOES

category=verb, verb-type=auxiliary, tense=present, lemma=do

VERBB base form of lexical verb (e.g. TAKE, LIVE)

category=verb, verb-type=lexical

VERBD past tense form of lexical verb (e.g. TOOK, LIVED)

category=verb, verb-type=lexical, tense=past

VERBG -ing form of lexical verb (e.g. TAKING, LIVING)

category=verb, verb-type=lexical, verb-form=present-participle

VERBN past participle form of lexical verb (e.g. TAKEN, LIVED)

category=verb, verb-type=lexical, verb-form=past-participle

VERBZ -s form of lexical verb (e.g. TAKES, LIVES)

category=verb, verb-type=lexical, tense=past

VHAVB base form of the verb “HAVE”, i.e. HAVE

category=verb, verb-type=auxiliary, lemma=have

VHAVD past tense form of the verb “HAVE”, i.e. HAD, ’D

category=verb, verb-type=auxiliary, tense=past, lemma=have

VHAVG -ing form of the verb “HAVE”, i.e. HAVING

category=verb, verb-type=auxiliary, verb-form=present-participle, lemma=have

VHAVN past participle of the verb “HAVE”, i.e. HAD

category=verb, verb-type=auxiliary, verb-form=past-participle, lemma=have

VHAVZ -s form of the verb “HAVE”, i.e. HAS, ’S

category=verb, verb-type=auxiliary, tense=present, lemma=have

VMOD modal auxiliary verb (e.g. CAN, COULD, WILL, ’LL)

category=verb, verb-type=modal

First, I give the full feature-structure tagging corresponding to the feature names and values proposed for VDOZ tag.

```
<f.struct>
  <feature name=category>
    <atomic>verb</atomic>
  </feature>
  <feature name=verb-type>
    <atomic>auxiliary</atomic>
  </feature>
  <feature name=tense>
    <atomic>present</atomic>
  </feature>
  <feature name=lemma>
    <atomic>do</atomic>
  </feature>
</f.struct>
```

Next I give the full feature-structure tagging corresponding to the feature names and values proposed for NUM tag. This is somewhat more interesting because it involves both the f.s.or and the f.s.not tags.

```
<f.struct>
  <feature name=category>
    <f.s.or>
      <atomic>noun</atomic>
      <atomic>adjective</atomic>
    </f.s.or>
  </feature>
  <feature name=numeral>
    <atomic>cardinal</atomic>
  </feature>
  <feature name=lemma>
    <f.s.not>
      <atomic>one</atomic>
    </f.s.not>
  </feature>
</f.struct>
```

In the document *Feature-Structure Markup for Presentation at Oxford and Brown Workshops* (this document has been submitted as a working paper of the AI1 group, but has not yet been assigned a number), entity definitions for a subset of the feature name-value pairs listed in AI1 W2 are given, for example, for tense=past, we have:

```
<!ENTITY T-A "<feature name=tense><atomic>past</atomic></feature>">
```

A corresponding set can be created for the feature name-value pairs needed here. Since the ultimate entity names (corresponding to the proposed BNC tagset) are not composed directly from the features, the naming conventions suggested in the above mentioned article do not need to be scrupulously followed. Assuming that & C-V; represents category=verb; & VT-A represents verb-type=auxiliary; & T-A represents tense=past; and & L-DO represents lemma=do, then we have the following definition for the entity VDOZ;:

```
<!ENTITY VDOZ "<f.struct>&C-V;&VT-A;&T-A;&L-DO</f.struct>">
```

A similar, but somewhat more elaborate, definition is needed for the NUM; entity corresponding to the proposed NUM tag.