Internal Use Only（非公開）

TR－SLT－0043

English Text Representation and Part-of-Speech Tagging Issues

David Carter

2003年6月27日

This report describes an investigation of the English travel (BTEC) corpus and the part-of-speech (POS) tagging scheme currently in use atATR. I first examine the role of POS tagging in a speech understanding system for English, where POS tag ambiguity is relatively high, and suggest that tagging can be very useful for translation but less so as an enhancement to the speech recognizer's language model. Next, attention is given to issues of word representation, principally those of when a multi-word phrase should be treated as a single lexical unit for the purpose of tagging. I then compare the ATR and University of Pennsylvania (Penn) tagging schemes on the basis of a number of criteria, and propose revisions to the ATR scheme to make it more suited to its purpose. Finally, I look at ways in which tagged texts from ATR and elsewhere can be used to speed up the process of manual annotation and correction. The intention is that all the elements of this report can be used as the basis for writing software to enhance the accuracy and usefulness of the POS tags used for BTEC and any future English-language resources developed at ATR.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

©2003（株）国際電気通信基礎技術研究所
©2003 Advanced Telecommunication Research Institute International

Contents

Chapter 1

Introduction

This report describes an investigation carried out between April and June 2003 of the English travel (BTEC) corpus and the part-of-speech (POS) tagging scheme currently in use at ATR. In this introduction, I will first look at the choices that can be made in representing the text itself so as to apportion work optimally between the speech recognizer and POS tagger. I will then outline why POS tagging is important, explain its place in the current ATR English speech understanding system, and point out some of the problems with the current ATR English corpora and tagging scheme, both theoretical and practical. The rest of the report will go into more detail on these subjects, and on what can be done to remedy existing problems.

I hope the critique and suggestions presented here can be used as the basis for a programme of work that will eventually produce tagged versions of BTEC and, in due course, other corpora, that will support accurate and linguistically valid automatic tag assignment for speech translation.

1.1 Word Representation

Before we can assign POS tags to the words in a text, we need the text itself to be in a suitable form. Consistency is obviously desirable: spellings should be American English ones, hyphens and abbreviations should be handled in a uniform way, as should capitalization, and so on. A particularly important issue is that of deciding on lexeme boundaries, where a lexeme is a unit that will receive a single tag. This issue surfaces in two ways. First, we should be consistent about word boundaries: we should always have "air mail", or always "airmail", but never both in the same corpus. Second, we are under no obligation always to make the lexeme the same as the word. If we wish, we can treat multi-word phrases like "New York" as a single lexeme, to be

given a single proper-name tag in just the way that, say, "Chicago" is. A lexeme may also sometimes be smaller than a word: words like "I'll" and "didn't" are syntactically best analysed as two-lexeme compounds, and we may choose to tag them that way.

As well as putting the text in good shape for tagging, we also need to represent it so as to divide the work optimally between the tagger and the speech recognizer. In a speech translation system, this can be done by modifying, in controllable ways, the spellings of certain homophones.

1.2 POS Tagging in Context

Why should we want to assign POS tags to words?

Ideally, we would like to be able to derive a full, correct syntactic analysis for every sentence we encounter, and to identify ungrammatical sentences. This is well beyond the state of the art. A useful step in that direction, however, is to hypothesize the preterminal nodes of the parse tree. This can be done, quickly and with quite good accuracy, on the basis of strictly local information – the identity and possible tags of a window of perhaps two or three words on either side of a target word. In a speech understanding context, such a partial analysis has two main possible uses:

- It can provide information useful to language modeling during speech recognition. To save memory and to overcome data sparseness, the ATR recognizer represents each word, for language modeling purposes, by the *class* to which it belongs. For the classes to be useful, we want the words in them to have similar distributions, and constraining the words in a class all to have the same POS seems like a sensible way of introducing linguistically valid constraints.

- POS tagging can make a start on the task of sentence parsing, or in other words, it can constrain the parser's search space. The parsing process will be speeded up, and if the tagging is accurate, incorrect parses that might otherwise be preferred can be ruled out. Similar remarks apply to example-based translation: if both the sentence to be translated and the examples in the database have POS tags attached, the search space can be reduced, and/or better solutions can be found.

The first of these uses has already been demonstrated at ATR for Japanese, at least in the case of a small (few hundred thousand word) training corpus where data sparseness is at its worst. However, for English, at the time of writing, no such result has been shown.

The reason may be that relatively few Japanese words are POS-tag ambiguous, so if we can successfully recognize a word, we can assign a tag to it with reasonable confidence; or to put it another way, viewing the tag as part of the word introduces relatively few additional homophones. The situation in English is different. Tag ambiguity is very common, and tagging errors are quite common: typically 3 to 5 per cent at best for error-free text, and probably much more for recognizer output. There is therefore a risk that even if we recognize a word correctly from its pronunciation, we may assign the wrong tag to it, which may make the language model's estimates for subsequent words useless. Or, if we again view the tag as part of the word, we introduce large numbers of homophones: "airmail" can be a noun, a verb or an adverb, and it is difficult to see how a speech recognizer, with incomplete and possibly incorrect information about word identities, can make a sensible choice.[1]

However, the second use of POS tagging, as a first stage in syntactic analysis or other classification or structure assignment, seems more plausible, if tags can be assigned accurately and if they reflect linguistic realities. It is on this second use that I will concentrate in this report.

1.3 Current Problems and Report Overview

Thus in brief, the problems with the current situation are these:

- The text itself is inconsistent: we can find "air mail" and "airmail", "colour" and "color", and so on. We may also want to adjust certain spellings so that certain decisions are allocated to ASR or to tagging as seems most appropriate.

- The tagging scheme (tag set and rules for assigning tags) fails to reflect many generally-accepted analyses of English syntax, making the tags themselves of doubtful value.

- The tagging scheme is in practice not very well-defined. Partly as a result of this, the error rate in the BTEC corpus is fairly high. (The current rather high error rate is also because the tags have been manually checked on only part of the corpus - the rest has simply been automatically tagged).

[1] A less ambitious role for POS tagging in speech recognition might be as a rescoring phase. If recognition is done without reference to tags and the recognizer outputs an N-best list of sentence candidates, each could be POS tagged, and a probability estimate derived for the tag sequence.

This report devotes one chapter to each of these three issues, as follows.

Text-representation issues, particularly those of homophone spelling and word compounding, are addressed in Chapter 2.

Chapter 3 then gives a detailed comparison of the ATR scheme with the best-known English tagging scheme, that developed at the University of Pennsylvania (Penn) and used in the various Linguistic Data Consortium (LDC) tagged corpora. This comparison is useful for two reasons. First, identifying the similarities and differences between the two schemes can help in identifying where each is lacking in some way, and hence how an improved hybrid scheme can be developed. Accordingly, suggestions are made for revisions to the ATR scheme. Second, if the two schemes can be brought into harmony with each other, it should be possible to use the tags assigned by an automatic tagger trained on the extensive and relatively error-free LDC corpora to identify and perhaps even correct errors in the tag assignments in ATR corpora.

Chapter 4 suggests some ways in which this harmonization and error detection could be done, by developing rules for automatically converting from one tagging scheme to another, and by comparing the existing manual annotations with automatically assigned tags. A rule formalism and a development plan are suggested.

Throughout this document, I will be assuming that POS tags first make an appearance in the system *after* speech recognition is complete. At present in the Japanese recognizer, POS tags are associated with words in the lexicon, so the output of recognition is a tagged sentence. However, I do not think this is advisable for English, because tag assignment is a much harder task in English, and there is too much risk of the recognizer, which is exploring a large search space, assigning the wrong tag and leading not only the tagging but the word recognition process astray. If tags are to be associated with words in the recognition lexicon in order to constrain the clustering of words into classes, it might be better to use fixed tags: the "tag" of a word for this purpose could be its most frequent tag, or a symbol formed its most frequent two tags. This removes from the recognizer the requirement to do any POS tagging as such, though of course, it means that proper (re-)tagging is required after recognition is complete.

Chapter 2

Text Representation and Word Compounding for Part-of-Speech Tagging

This chapter deals with text normalization issues. Some of these are straightforward matters of text cleaning: we need to eliminate multiple (e.g. American/British) spellings of the same word, and have a consistent policy for hyphenation, the use of dots in abbreviations, and so on. Most of this is donkey work, and much of the time, no solution is inherently better than another – it is just important that whatever solution is adopted, is applied consistently. I will not have any more to say about those problems.

However, there are two areas where things are more complex, and more interesting. Firstly, there are different ways in which various homograph/homophone decisions may best be tackled, and we will see how this can be approached. Secondly, we need to think about what items should be classed as words, and whether part-of-speech tags should always be assigned to single words, or sometimes to part-words or multiple-word units.

2.1 Dealing with homophones

The important mapping in a speech translation system is the one from word pronunciations to word meanings, which are combined to produce translations. It is of course usually convenient to represent words by their conventional spellings, but since these are not the main output of the system, we are free to spell words in whatever way we wish.

¿From this point of view, the ambiguity problem is when two words are pronounced the same but mean something different. Conventionally, if these

words happen to be spelled differently, we call them homophones, while if they happen to be spelled the same (as well as pronounced the same) we refer to lexical ambiguity. A related phenomenon, that of homographs (two words with the same spelling but different meanings and different pronunciations), is only a real issue when processing written text. In speech processing, we can avoid it altogether by writing the two items differently.

A system that performs POS tagging as a first step towards sentence interpretation is effectively dividing the word ambiguity task into two stages. Ambiguities that involve a difference in POS, such as "book" being a noun or a verb, are resolved by the tagger. Other ambiguities, such as "bank" meaning a financial institution or the side of a river (a noun in both cases) must be resolved after POS tagging.[1]

This gives eight combinations altogether: same or different pronunciation, spelling and POS. Examples of each are shown in Table 2.1.[2] The "Choice" column shows the component of a speech understanding system that would (with conventional spellings) have to resolve the ambiguity: the speech recognizer and its language model (ASR/LM), the tagger, or some later component, which for convenience I will call the interpreter.

ASR/LM ends up having to make the choice in four of the eight cases. Where the pronunciations differ (cases 7 and 8) this is fine, although sometimes it may be tricky if the words are phonetically similar. But ASR is also asked to make the choice in cases 3 and 4, just because, for arbitrary historical reasons, these words are spelled differently in English. Conversely, in cases 5 and 6, ASR is not asked to make the choice, even though it is best placed to do so, because the pronunciations differ. Clearly, responsibilities not being allocated sensibly between the different components.

Notice, though, that if we choose to spell words differently, we can transform one case into another: we can swap between cases 1 and 3, between 2 and 4, between 5 and 7, and between 6 and 8. If we eliminate a spelling difference (writing "plaice" as "place", or "meat" as "meet") we move the choice out of ASR into the tagger or interpreter, and if we introduce one (writing *sake1* for the wine sense and *sake2* for the benefit one, for example), we move the choice from another component into ASR. If we think the ASR and/or LM will be better at making a particular distinction, we can make the spellings distinct. If we think the tagger or (in case of shared POS tags) the interpreter will be better, we can make the spellings the same. We can

[1]Of course, if we did semantic tagging rather than POS tagging, we might resolve "bank" during tagging.

[2]In fact many words participate in several patterns at once. For example, "place" in case 3 could also be used as an example for case 1, because it can be a verb as well as a noun.

Pattern				Example			Choice
Type	Pronunc'n	Spelling	Tags	Word	POS	Meaning	
1	Same	Same	Same	bank bank	noun noun	company riverside	Interp.
2	Same	Same	Diff	book book	noun verb	pages reserve	Tagger
3	Same	Diff	Same	place plaice	noun noun	location fish	ASR/LM
4	Same	Diff	Diff	meat meet	noun verb	food encounter	ASR/LM
5	Diff	Same	Same	sake sake	noun noun	rice wine benefit	Interp.
6	Diff	Same	Diff	wind wind	noun verb	moving air coil up	Tagger
7	Diff	Diff	Same	hat hut	noun noun	headgear building	ASR/LM
8	Diff	Diff	Diff	hat hot	noun adj	headgear very warm	ASR/LM

Table 2.1: Examples of pronunciation, spelling and POS differences

recover the original spellings at some point downstream – after tagging if the POS tags are not shared, or after interpretation if they are.

So what factors might influence our decision to introduce or remove a spelling distinction?

If two differently-spelled words are pronounced the same, we should only leave the spellings different if we think the ASR language model will do a better job of distinguishing them than the tagger (if the tags are different – case 4) or interpreter (if they are not – case 3). Whether this is the case depends largely on the particular technology used and on the distributional properties of the words concerned. A rule of thumb might be that we should usually merge spellings for case 3 (transforming it to case 1), since this is the most difficult case – the words are the same both phonologically and syntactically. However, the language model might sometimes be able to cope: if "plaice" only ever occurs in the phrase "plaice and chips", then a suitable trigram score should do the job. Case 4 is easier, as both the language model and the tagger have a good chance of making the right choice (for similar, distributional, reasons). Which of them should be allowed to make it depends on the specific case.

However, one subcase where we probably will want to merge spellings is where case 4 is due to inflectional differences, because if we don't merge, we risk having an overlarge recognition lexicon which will compromise language model accuracy. The primary case of this in English is the "s" suffix which can signify either plural, singular possessive, or plural possessive (e.g. *books, book's, books'*), and can be used with (virtually) any noun or proper name – probably the majority of words in the lexicon. Currently, both the possessive endings are split off and tagged separately, so we would need to introduce four new tags, e.g. `CN-POSS, CN-PL-POSS PROPN-POSS, PROPN-PL-POSS`, but I think it would be worth it, as otherwise it will be very difficult to get possessive words right at all.

In general, though, if we're thinking of moving a particular problem from ASR to the tagger by writing two words in the same way, we should be reluctant to introduce new tags. It might be nice to rewrite the phrase "a long" (as in "a long time") as "along", because they're homophones, but then we need to give a tag to "a long", and it isn't a constituent so there isn't a suitable tag.

When might we want to introduce a spelling distinction where none conventionally exists? I imagine we will usually want to do so in cases 5 and 6 (same spelling, different pronunciation), at least when the phonemes are different, because if we don't, we are throwing away useful information when we move from ASR to tagging. Whether we also do so when only the stress patterns differ, as in noun/verb pairs like *convert*, depends on whether ASR

detects stress and on how much reduction we expect to encounter (*convert* as a verb can have the first vowel reduced to schwa, but as a noun, it cannot).

We might in principle want to introduce spelling distinctions even when the pronunciations are the same (cases 1 and 2), in case the two meanings have very different distributions and we think the language model will be able to make the choice. However, it is not clear there would be any great advantage to this in practice.

The decisions on what spelling changes to make can be made late, after the relevant corpora are cleaned up and tagged. Once the changes have been decided, applying them is a simple automatic global edit. It should therefore be possible to experiment with different strategies, training the ASR/LM and the tagger (and perhaps the interpreter) afresh each time. It is probably only in this way that we will be able to discover what sorts of changes really do help.

2.2 Issues in word compounding

A prerequisite for a tagging scheme is a set of rules for deciding on lexeme boundaries in the text to be tagged. By "lexeme" I mean a unit that receives a tag. There a number of sources of variation in the spellings and boundaries of lexemes.

- Free variation in the way words and phrases are written in the original text. We may have "non smoking", "non-smoking" and "nonsmoking", and "air mail" or "airmail". All are correct; they just follow different conventions.

- Variation determined by syntactic role. Here, there are two forms, but only one form is correct in a given situation:

  I go swimming *every day.* (Adverbial phrase)
  It is an *everyday* occurrence. (Adjective)

- For tagging, we can choose to split off clitics to reflect syntactic structure better: "I'm" and "John's" could be split at the apostrophe, and "don't" and "mustn't" before the "n".

- We can also choose to join multiword units that behave syntactically or semantically as a single unit and so deserve a single tag: "Los Angeles", "excuse me".

First some terminology:

- A *compound* is multiword phrase that is treated as a single lexeme gets a single tag. I will write it with an underscore: "air_mail".
- A *separated* phrase is a multi-lexeme phrase, one lexeme per word: "air mail".
- A *joined* word is two or more component words appended without any intervening character: "airmail". It receives a single tag.
- A *hyphenated* word contains components separated by hyphens: "air-mail". However, hyphens have been replaced by spaces in the current version of BTEC, so hyphenated words are mapped onto their separated equivalents.
- A *cliticized* word is like a joined word orthographically, but like a separated phrase syntactically: "mustn't" is written without a space, but is syntactically a two-word phrase, each word having its own tag.

Just as with word spellings (discussed in Section 2.1), we are free to choose any compounding (lexeme boundary definition) scheme we like. In a speech translation system, no-one sees the word boundaries, so "correctness" only matters in so far as it affects compatibility between different system components. For example, the lexicon and grammar used by a later parsing stage might assume that the "correct" forms are used. However, we always need to maintain consistency between boundary decisions at the same level of representation (we don't want a mixture of "airmail" and "air mail" in the same corpus) and in component compatibility (we don't want to train the tagger on "air mail" if the speech recognizer is going to output "airmail").

What factors affect whether we might want to treat two or more items as a compound (or a joined word, if appropriate) rather than as a separated phrase? There are four main things that tend to happen, at any stage of processing, not just part-of-speech tagging, when we decide to join two items, or in the case of cliticized words, to avoid splitting them.

1. We are making an assumption that those items "belong together" at the current level. Doing so may be providing extra information. When this information is right, it can help later processing, but if it is wrong, it may cause errors.
2. We are increasing the "reach" of any system component that looks at the input as a sequence. This component could be a trigram language model or a tagger with a fixed window size, for example.

3. We are introducing a new item into the lexicon, which may worsen data sparseness issues.

4. There may be work implied by the decision to compound. For example, some kind of definition may need to be written for the compound, or existing data resources may need to be converted.

Thus, there are benefits and costs associated with any decision to make a compound, and there is no reason why the same tradeoffs should apply at different stages of processing. We might make two words into a compound in speech recognition but not in tagging, or (perhaps) in tagging but not in parsing.

Let's look at each of the above four considerations in turn.

1. **Items belonging together.** When items do belong together at the level in question, there is a benefit. At the level of pronunciation, we can say that items belong together if the combination is frequent, and/or there is significant coarticulation between the words. Thus "can be" might be treated as a compound at this level. At the tagging and parsing levels, "Los Angeles" is a sensible compound, because both syntactically and semantically it is a single unit.

   However, the same sequence of words may be a sensible compound in some contexts, but not in others. We might want the sequence "how do you do" to be treated as a compound when it occurs on its own, or just before a name, but not when it occurs in the sentence "how do you do that?". Essentially, if we treat a sequence as a unit at parsing time (or if we do so at tagging time and don't undo the decision afterwards) then we are saying it is a syntactic constituent, or at least that it behaves syntactically in a predictable way.[3]

2. **Increasing the reach of a component.** This is straightforward. If we treat "Los Angeles" as a single unit, then a trigram language model becomes a four-gram model for any trigram containing "Los Angeles".

[3]Another example, not from the travel domain, is the phrase "harmful insect extermination", the name of a process to be carried out at ATR on July 6th, 2003. It would be tempting to specify "harmful_insect" as a compound in anticipation of translation to the single Japanese word *gaichuu*, but the bracketing implied by this compound might not always be the correct one. According to an e-mail notice issued by General Affairs on June 16th, the process involves filling the whole of the inside of the building with an unspecified insecticide, for which the only promise made is that it is "odorless and doesn't adversely affect the machines". Since General Affairs have also previously instructed us never to open the windows, it is far from clear to me which word "harmful" modifies in this case. Fortunately perhaps, I will not be here to find out.

This is a good thing. Furthermore, for a class-based language model, or for any tagger (viewing parts of speech as class names), classifying "Los Angeles" as a single proper name allows us to substitute it for other, single-word proper names and make better predictions.

3. **Data sparseness.** The more items we add to our lexicon, the more parameters we add to our model, and the sparser our data gets. How bad this effect is depends on how closely associated the items are. If "Angeles" only ever occurs in "Los Angeles", we are not adding any parameters by treating "Los Angeles" as a compound, because we no longer need an entry for "Angeles". In the extreme case, neither word may ever appear without the other, in which case we are reducing the lexicon size by joining them. But such cases are quite rare in practice.

4. **Extra work.** If we introduce a compound into a recognizer lexicon, we will (to get the full benefit) need to write one or more new pronunciations showing the coarticulation. If we introduce one into a tagged corpus, we need to decide what tag to give it. In either case, a global data set may need to be reprocessed, e.g. the language model or tagging rules rebuilt. It may also be non-trivial to maintain consistency by keeping track of what is meant to be a compound and what isn't.

Thus, when deciding whether or not to make a compound for a given stage of processing, we need to trade all these factors off against each other.

These issues as they affect lexicon design for speech recognition have already been dealt with by the work of Padmanabhan and others, and discussed here at ATR. From here on, I will concentrate on compounding issues as they appear in the tagging task.

2.3 Compounding for Tagged Corpora

2.3.1 Achieving consistency

The current version of BTEC contains 4507 different compounds out of a total vocabulary of 26711. Of these compounds, 39% only appear once, and 61% appear three times or fewer. Many appear in more than one form:

- 20% occur as separate words as well as compounded (e.g. "see/V you/PRON" occurs as well as "see_you/INTERJ').
- 3.6% occur as single words as well as compounded (e.g. "airmail/CN" as well as "air_mail/CN"), while 1.1% occur in all three forms.

- The corpus also contains four or five hundred cases of words which occur as single words and word sequences but not as compounds, e.g. "bathtub/CN" and "bath/CN tub/CN" but not "bath_tub/CN". It is not possible to count these cases automatically, because some are coincidental ("car at" is not an alternative spelling of "carat") and others are sometimes variants of each other and sometimes not ("some time" vs "sometime").

- There are also quite a few examples in BTEC where a word sequence is represented as a compound, but the words in it are not used in the way that is assumed by the tag assigned to it: "*I_see*/INTERJ my/DET friend/CN every/DET week/CN".

- Additionally, there are variations due to British/American spelling differences (colour/color, centre/center) and different conventions for periods (U.S.A./USA/U S A/U. S. A.).

This is a rather confused state of affairs which needs to be sorted out. In particular, we need consistency in the forms used for each word or phrase. The main rule is:

> **Rule R1:** Regardless of the tags used, if a phrase occurs as a compound, it should never occur separated. For example, if "see_you" occurs, we should never encounter "see you".

Linguistically it seems reasonable to have "see you" as a compound when it is an interjection but not when it is a verb phrase, since the first use is idiomatic and the second is not. However, for tagging purposes, this is not an option: we need decide which word sequences to make into compounds *before* the tagger is invoked, so we do not know what the tags will be.

Note that this rule does not quite imply that whenever the word "see" is followed by the word "you" in the original version of the corpus, it will be represented as "see_you". This is because compounds can overlap: if "I see" is also a defined compound, then we can only represent "I see you" as "I_see you", "I see_you" or "I see you". For consistency, we need some rule to decide which phrases to compound. The simplest rule is a "greedy" one, which says that at each point in the sentence, we compound as much as possible, then move on to after the compounded phrase. This would give the representation "I_see you" rather than "I see_you".

Note also that we need to consider how rule R1 impacts on any spelling changes (Section 2.1) we may have decided on, because we can only make spelling changes between single lexemes. For example we can only make "I see" and "icy" share a single spelling if the former is treated as a compound.

Otherwise, we are faced with just the same "one lexeme or two" dilemma that we would face if we allowed the corpus to contain both "I_see/INTERJ" and "I/PRON see/V". A more pertinent example involves possessives: in Section 2.1 we considered omitting the apostrophe in singular and plural possessives so that ASR does not have to represent *books, book's* and *books'* separately. However, this will not work if the tagger expects possessive markers, but not simple plurals, already to be split off.

It would be nice if we could have a similar rule to R1 to prevent co-occurrence of spaced phrases (or compound words) such as "air mail" (or "air_mail") and joined words such as "airmail". However, this problem is more complicated than the compounding-spacing one. We can distinguish four cases:

1. Free variation: "air mail" and "airmail" are both correct; we should consistently choose one, but which one is better depends on the four considerations given in Section 2.2.

2. Spacing errors: "understand" is right, "under stand" is wrong, and we should consistently choose the right one.

3. Tag-dependent variation: "headache" and "head ache" can both be right, but only one is right in a given syntactic context: "I have a headache", but "It makes my head ache".

4. Unrelated items: "car at" and "carat" can both occur, but they have nothing to do with each other.

The first cases is straightforward: we just make one choice consistently. Note that one cliticized word, "cannot", comes into this category, as "can not" is virtually equivalent,[4] and also correct.

The second case is also straightforward; we make the correct choice of the two consistently. Cliticized words (other than "cannot") can be treated as if they were spacing errors, replacing "mustn't" by "must n't", etc.

The third case is more difficult: the correct choice depends on the tag(s), but we need to make the choice before we know what the tags are. Sometimes, the ASR language model may be able to do this, in which case we can leave things as they are. If not, we will have to make an arbitrary (but consistent) decision, and join or split after tagging if it turns out to be wrong.

The fourth case is in principle like the third, but in practice we can usually leave both forms in the corpus and in the recognizer lexicon, because, being

[4]Not completely equivalent: "He can not [or 'cannot'] only sing, but dance" is fine, but "He can't only sing, but dance" is wrong.

unrelated, they will often be pronounced differently, and nearly always have different distributions in the language model. Thus we can formulate a second rule:

> **Rule R2:** Regardless of the tags used, if a phrase occurs as a single word, it should never occur as a multi-word phrase unless the pronunciations differ enough for the speech recognizer to reliably make the choice on the basis of pronunciations or language model.

If, in the case of tag-dependent variation, this rule conflicts with correct usage, then the problem needs to be corrected after tagging on the basis of the tags that have been assigned.

2.3.2 Assigning tags to compounds

If phrases are selected for compounding, or if a joined form like "headache" is selected in the case of tag-dependent variation, we need to be sure that it is possible to assign a tag to all uses of the word or phrase. For example, "see you" as in "Can I see you?" can reasonably be tagged as a verb (`V`), because it has the same distribution as some verbs (syntactically it will nearly always be a verb phrase, and intransitive verbs can also be verb phrases). However, "I see" as in "Can I see you?" and "head ache" as in "It makes my head ache" do not correspond to existing tags in this way and do not usually form syntactic constituents.

In such cases, we have a choice. We can avoid the problem and use a separate-word representation after all. Alternatively, if we do use the compound form, we can define a new tag for it. These factors will influence the decision:

1. A separate-word representation decreases the reach of the tagger on sentences containing the phrase, which is a shame if the phrase is a common one like "I see", especially if the syntactically problematic reading is in the minority.

2. A separate-word representation can also lose information about higher-level structure. In BTEC, "how are you" can either be a compound, tagged as `INTERJ`, or a separated phrase, tagged `WHADV BEV-2 PRON`, used in phrases like "how are you going to do that?". But if we separate the compound, the correct tags are also `WHADV BEV-2 PRON`, so a distinction affecting translation has been lost. Similarly "a bit" as a separated phrase can only be tagged "a/`DET` bit/`CN`", but if we make

it a compound, we can distinguish different tags: "a_bit/PREADV slow", "slow down a_bit/ADV", "a_bit/CN of progress".

3. Introducing a compound will always increase data sparseness slightly, unless at least one of the component words *only* appears in the compound.

4. Inventing a new tag for the compound, if this is necessary, will increase data sparseness significantly. This is undesirable, although it can still be justified if it is likely to be frequent in the corpus. Note that a tag does not have to correspond to a valid part of speech for an individual word, nor even to a single syntactic constituent, but can instead be viewed more generally as an instruction for assigning tags (that do correspond in that way) to its individual words. Thus if we invent a tag PRON_V to describe phrases like "I see", we can if we wish, during the post-tagging phase that is necessitated anyway by pairs like headache/head ache, easily recover the word-level tags.

Another approach would be to define a whole new set of tags (called "ditto tags" in Ezra Black's previous work at ATR) to represent compounds. For example, instead of "I_see/INTERJ", we would have "I/INTERJ-1 see/INTERJ-2". Then it would be no problem to have "I/PRON see/V" in sentences like "Can I see the menu?". This will be good from the point of view of points 2 and 3 in the above list, but bad from the point of view of points 1 and, particularly, 4: introducing multiple new tags per existing one will increase the size of the tag set markedly. The approach is worth trying, but depending on the amount of training data available, it may be not be a good solution because of all the additional parameters.

2.4 Practicalities

How might we convert the BTEC corpus to a form that is better suited to training a part-of-speech tagger?

2.4.1 Conversion Rules

Rather than making a firm decision in advance on exactly which forms should be compounded, I think it would be best to develop a rule formalism and a set of rules which would allow decisions to be easily altered in order to determine which combinations worked best for tagging and interfaced best with the speech recognition and translation components. The rules could, I think, be

fairly simple; all they would have to do is specify an equivalence between noncompound and compound forms with certain tags, and a direction of application. For example, a rule such as

air/CN mail/CN → airmail/CN

would force all (suitably tagged) occurrences of "air mail" to be replaced by "airmail". If at a later stage it was decided that "air mail" was a better representation after all, the arrow could simply be reversed and the mapping redone.

It would be possible to generate a first set of rules automatically from the existing corpus. "Air mail" (with or without the space) occurs 515 times in BTEC at present: 222 times as air_mail/CN, 285 times as airmail/CN and eight times as airmail/V (as in "I would like to airmail this"). Some suggested rules could be generated automatically:

air/CN mail/CN ↔ airmail/CN
air/CN mail/CN ↔ airmail/V
airmail/CN ↔ airmail/V

The last two rules would cover the possibility of airmail/V being a mistagging. In fact, six of the eight cases of airmail/V in BTEC should actually be airmail/ADV (as in "I'd like to send this letter airmail to Japan"), one should be airmail/CN, and one is correct. Thus if we decided "airmail" was the desired spelling, we make the arrow in the first rule point to the right, delete or comment out the second rule, and either edit the third to read

airmail/ADV ← airmail/V

or correct the airmail/V cases by hand.

2.4.2 Some examples from BTEC

Tables 2.2 and 2.3 show the twenty most common (defined using an entropy measure) compounding variations in the BTEC corpus. The leftmost column shows the phrase in question. The other three give counts and tags for each of the three possible forms, compounded, separated-words and joined-word. In the "Separated" column, the tags are actually sequences; thus PRON_V for "I see" means "I/PRON see/V". All but a few rare, erroneous tags are shown.

Of these twenty examples, three are in the "unrelated" category from Section 2.3.1: along/a long, mean/me an, and area/are a. These can be left as they are. The last two should not pose a problem because they involve a pronunciation difference. The first does not, and will depend on the ASR

Phrase	**Compound**	**Separated**	**Joined**
Example	*rest_room*	*rest room*	*restroom*
I see	811 INTERJ	1334 PRON_V	
a little	1630 PREADV, PRONADV, DETADJ	403 DET_ADJ, DET_PREADV	
see you	706 INTERJ	727 V_PRON	
what time	3075 WHPRON, WHADJ	230 WHADJ_CN, WHPRON_CN,	
thank you	5232 INTERJ	183 V_PRON	
a long		463 DET_ADJ	347 PREP, PREPADV
how much	4581 WHPRON, WHADJ, HOWADV	100 HOWADV_ADJ, HOWADV_PREADV, WHADV_ADJ	
me an		152 PRON_DET	1022 V, CN, ADJ
a lot of	504 DETADJ	218 DET_CN_PREP	
get to	1104 V	134 V_PREP	

Table 2.2: Most common compounding variations in BTEC: Part 1

Phrase	**Compound**	**Separated**	**Joined**
Example	*rest_room*	*rest room*	*restroom*
a lot	251 PRONADV, PREADV, DETADJ	341 DET_CN	
may be		181 AUXV_BEV	475 ADV
how are you	510 INTERJ	167 WHADV_BEV-2_PRON	
a few	342 DETADJ, PREADV	208 DET_ADJ	
out of	608 PREP	142 ADJ_PREP, PREPADV_PREP, PREADV_PREP	
air mail	222 CN		293 CN, V
are a		130 BEV-2_DET, BEV-1P_DET, BEV-3P_DET	524 CN
go ahead	343 V	159 V_ADV	
rest room	201 CN	2 CN_CN	232 CN
kind of	115 PREADV, DETADJ	436 CN_PREP, ADJ_PREP	

Table 2.3: Most common compounding variations in BTEC: Part 2

language model making the right choice. One example, "may be" is syntactically determined, like "headache"; we can either trust the language model to get it right, or select one of the forms and sort out any spacing problems after tagging (though it is not clear what single tag, if any, would be appropriate for "may be").

The other sixteen examples are all cases where we can make a free choice between two or more of the compound, separated or joined forms. In only two cases, "I see" and "how are you", is there a problem with assigning a tag to the compound form. However, in five cases – "a little", "how much", "a lot", "a few" and "kind of" – if the compound form is chosen as the standard, it will be necessary to inspect the context of each separated instance to work out which tag to assign, since the existing compounded form has more than one valid tag.

It is worth looking at these cases in more individual detail, to give a flavour of the kind of decisions that will need to be made for all the 1659 phrases in BTEC that occur in more than one form.

1. **I see**. In the more common non-interjection use, this is not a syntactic constituent, so it's probably best to replace "I_see/INTERJ" by "I/PRON see/V". This loses some information, but in practice most of the INTERJ cases are stand-alone or occur only with another INTERJ, so they are easy to distinguish.

2. **a little**. Should be a compound, because most of the occurrences are constituents – those currently represented as separate words ("a/DET little/ADJ", "a/DET little/PREADV") are mostly errors and should (under the current ATR tagging scheme) be "a_little/DETADJ" and "a_little/PREADV" respectively. The exceptions are phrases like "a little while" and "a little rest", where "little" is just an adjective like "short". These could be labelled DETADJ too to avoid introducing another tag.

   Note that there are many errors in all taggings of "a little" in the corpus. I think this is related to the DETADJ tag being misconceived. See Section 3.6.1.

3. **see you**. Keep the *INTERJ* version, and replace "see/V you/PRON" by "see_you/V", since this form of "see you" is a transitive verb with an object, which syntactically behaves like an intransitive verb.

4. **what time**. All occurrences should be "what_time/WHPRON". This can theoretically be wrong ("Einstein explained what time and space really are") but we don't expect such examples in the travel domain.

5. **thank you**. This is exactly parallel to "see you". "thank/V you/PRON" should be replaced by "thank_you/V".

6. **a long, along**. These are unrelated. Although the pronunciations are similar or identical, the ASR language model should be able to tell them apart, so no change is needed.

7. **how much**. The relatively few separated examples should all be replaced by the compound "how_much/WHPRON".

8. **me an, mean**. Unrelated and pronounced differently, so leave alone. Note that nearly all the instances of "mean/CN" should really be "mean/V" or "mean/ADJ"!

9. **a lot of**. The separated forms should be compounded to "a_lot_of/DETADJ" (though again, see the objections to DETADJ in Section 3.6.1).

10. **get to**. The separated forms are errors, and should be compounded to "get_to/V", though note that there a minority of examples (both compounded and separated at present) where the "to" is an infinitive, as in "get to know you", and in this case, "get_to" does not quite behave like any existing tag because like a modal (AUXV) it can be followed by a verb base form, but like a non-modal, it can be preceded by a modal.

11. **a lot**. The (majority) separated forms should be changed to "a_lot/PRONADV", or when followed by an ADJ-ER, to "a_lot/PREADV" Note that when "a lot" is followed by "of", we have a case of "a_lot_of" (see above), not "a_lot".

12. **may be, maybe**. These are (functionally, though not historically) unrelated, so should be left as they are. Like "a long/along", the pronunciations are very similar, but the language model should be able to choose.

13. **how are you**. I think the **INTERJ** form should be changed to the separated form, because the latter (as in "how are you doing") is a non-constituent and does not correspond to any existing tag. It is easy to recover the distinction by looking to see whether there is a V-INGP as the right neighbour.

14. **a few**. As for "a lot", the separated examples should be changed to "a_few/PRONADV", or when followed by an ADJ-ER, to "a_few/PREADV".

15. **out of**. The separated cases should be changed to "out_of/PREP".

16. **air mail/airmail**. Either form is fine, but we need to have one of them consistently. The joined version might be better for ASR. Note that only one of the eight current examples of "airmail/V" is actually correct – most of them should be ADV.

17. **are a/area**. Totally unrelated, leave alone.

18. **go ahead**. Replace separated form by compounded.

19. **rest room/restroom**. Separated, joined and compound forms all occur. Joined is probably best, as it is the most frequent, and should be better for ASR.

20. **kind of**. Most of the current examples are separated, and there is no current tag that would be appropriate for compounding the current separated ones, particularly "kind/ADJ of/PREP" as in "It's very kind of you". So I suggest the compounded examples be separated to "kind/CN of/PREP", though this does involve some loss of information. No solution seems really satisfactory here, unfortunately.

An examination of a further twenty compounding variations (not shown here), selected this time at random rather than by entropy, suggests that the pattern of the twenty in Tables 2.2 and 2.3 is fairly typical, except that there are fewer "unrelated" cases in the random set, and, crucially, there are no cases like the five listed above where manual retagging may be required. This means that although some attention to each of the 1659 cases of variation in the corpus will be required, only a limited amount of manual re-tagging of individual sentences will be needed.

Even if some compounds – especially the rarer ones – are dispensed with, it worth be worth applying Padmanabhan's algorithm to see what new ones would be worth creating. In general, the better-scoring word-tag sequences would be worth making into compounds as long as a suitable existing tag could be assigned to them. It would be worth writing a few grammar rules to automate this process; rules such as CN $\rightarrow$ ADJ CN and V $\rightarrow$ V PREP would be typical.

Chapter 3

A comparison of the Penn and ATR tagging schemes

This chapter is a thorough comparison of the ATR tagging scheme used in the LDB and BTEC English corpora, with the University of Pennsylvania ("Penn"; Marcus, Santorini and Marcinkiewicz, 1993; Santorini, 1995) scheme, the best-known scheme for English, which is used in LDC resources. Where the taggings used in the two schemes are not equivalent, I have tried to evaluate their relative merits, and have suggested changes to the ATR scheme where necessary.

3.1 Criteria for a Tagging Scheme

What makes one tagging scheme "better" than another? In principle, different types of application could lead to a wide range of "best" schemes. In practice, however, theoretical and software-engineering criteria tend to coincide to make certain characteristics highly desirable. So let's look at what those characteristics might be.

3.1.1 Reflecting the grammar of the language

The most obvious criterion for a (syntactic) tagging scheme is that words should be assigned tags which reflect their generally-recognized grammatical roles in sentences.

Although a wide variety of different grammatical theories exist, there is generally very close agreement between different (categorial rather than dependency-based) schemes on how word uses should be characterised. Where differences exist, they tend to involve one theory making a distinction that

another fails to make, rather than inconsistent analyses being made.

A related point is that the tagging scheme should reflect only the grammar of the language being tagged. If the purpose of tagging is to support translation into Japanese, it is tempting to make some distinctions on the basis of what form is used in translation. For example, one might classify colour and nationality adjectives as if they were nouns, on the basis that "brown" translates to "cha-iro no" and "Japanese" to "Nihon no". However, such a tagging scheme will not generalize to other applications. It may also compromise modularity, in that it is asking a relatively early stage of processing – the tagger – to make decisions that really belong in a later stage such as transfer.

Similarly, it is tempting not to bother to classify nouns for singular and plural, on the grounds that the distinction will not be reflected in a Japanese translation. However, this distinction may well improve tagging accuracy because of subject-verb agreement. A transition from a singular noun to a singular verb, or from plural to plural, is more likely than one involving a change of number, and so if an apparent change of number is detected, alternative parts of speech might be more likely. An example could be the word pair "musical sound". This could either be `ADJ` followed by `CN`, or `CN` followed by `V`. However, in the latter case, if "sound" is finite then "musical" cannot be its subject because of the number mismatch:

I heard a musical/`ADJ` sound/`CN`.
Does that musical/`CN` sound/`V` interesting to you?
*That musical/`CN` sound/`V` interesting to me.

3.1.2 Decidability and computability

In practice, another vital feature of any tagging scheme is that an automatic tagger should be able to assign tags correctly. Taggers treat sentences as linear sequences, and do not look for the underlying hierarchical structures. They also tend only to look within a window of at most three words either side of the word to be tagged. If a tagger is applied to speech recognizer output, the usable window size may be even more limited, because the wider the window, the more likely it is to contain a misrecognized word which can lead the tagger astray.

Because of this, there are a number of valid syntactic distinctions that are best not included in a tagging scheme. For example, one might wish to distinguish prepositions governing noun phrases from prepositions (or one might call them subordinating conjunctions) governing clauses: "I left before John" versus "I left before John arrived". The problem here is that the

choice of tags depends on information that might be arbitrarily far away (the noun phrase represented by "John" could be of any length) and arbitrarily hard to compute: one needs to parse the whole sentence to know whether the noun phrase after the preposition is or is not the subject of any following verb phrase. In some corpora, this might often be an impossible decision for a tagger. However, in BTEC, it turns out that for the word "before", the clausal argument ("before John arrived") case nearly always begins with a pronoun, and the nominal case hardly ever does; so the distinction is computable for that corpus. Whether we include it therefore depends on how much we wish the tagset to be corpus-independent.

Some distinctions are not decidable even for humans, in the absence of extrasentential context which (we assume) is not available. For example, most uses of the pronoun "you" are undecidable with regard to number, so even if a number distinction is introduced for other personal pronouns, it should not be imposed for "you".

Other distinctions may be easy for native speakers of the language to make when hand-tagging a corpus, but rather harder for non-native speakers, even fluent ones. For example, native speakers can easily distinguish present participles (ATR `V-INGP`) from gerunds (ATR `V-INGG`) from the fact that, in a neutral, non-contrastive context, gerunds take the main stress but present participles do not:

a *smo*king seat. (gerund: the seat is not smoking)

a smoking *gun.* (participle: the gun is smoking)

The Penn tagging guidelines (Santorini) make heavy use of tests involving the acceptability of variants of the sentence in question, such as whether particular types of modifiers are allowed. These too tend to be harder for non-native speakers to judge.

It is possible to some extent to replace such intuitions by objective tests which can be applied more easily by non-native speakers, but care must be taken. Criteria based simply on the presence of dictionary entries with particular parts of speech can be seriously misleading, because those parts of speech may be for different meanings of the word. For example, section 3.2.1-3-b (p24) of the ATR tagging guide, *Eigo Keitaigo Shiyousho* (February, 2001; henceforth just "Shiyou") says that where the "-ing" form of a verb is also defined in READERS as an adjective, the adjective tag should be used. However, when applied to the word "becoming", this gives the wrong result.

"Becoming" as an adjective means "attractive" or "smart", and has nothing to do with the verb "become". Of the 57 occurrences of "becoming" in BTEC, 32 are tagged ADJ, but of these, only nine are actually the adjectival sense; the other 23 are present participles (V-INGP). It is therefore necessary to check that the dictionary entry is for the right sense of the word.

Similarly, tests based on whether a certain word *is* (rather than *can be*) modified in a certain way often lead to inconsistent results. For example, in the ATR scheme, if a word can be either a noun or an adjective, it is taken to be a noun unless one of a number of criteria applies. One of these criteria is that the word is "modified by a degree adverb like 'very' or 'much"' (or, presumably "more"). This means that in BTEC, the sentence "we should be more alert" has "alert" correctly tagged as ADJ, but in "she is alert", it is tagged CN, as it would presumably also be in "we should be alert". In other words, the presence of a degree modifier changes the tag on the modified word.

Sometimes, criteria are consistent and linguistically sensible, but are difficult to apply. Such criteria are usually easy to detect in tagged corpora. Sometimes, however clearly a set of rules is written, there is just not a clear binary distinction between two usages. A major example of this is words ending in "-ing", which are sometimes nouns and sometimes verb forms. In some contexts it is hard to make the decision, and both the Penn and ATR corpora are full of inconsistencies and errors in this respect.

One quite major problem with the Shiyou document (and to a much lesser extent with Santorini) is that the rules for choosing between tags are often not formulated as decision procedures. The typical format is something like "Use tag A when condition A1, A2 or A3 holds; use tag B when condition B1 or B2 holds; and use tag C when condition C1 or C2 holds". This format does not tell the reader what to do if, say, conditions A2, B2 and C1 all hold. The wording needs to be changed to a set of if-then-else rules in which it is clear which tests should be applied in which order.[1]

3.1.3 Redundancy and lexical recoverability

As Marcus *et al* point out, the Penn tagging scheme involves far fewer tags than most of the schemes previously proposed. It also has a lot fewer tags than the ATR scheme if all the subtags are included for the latter.

The developers of the Penn tag set justify this (Marcus *et al* 1993, section 2.2.1) on the grounds of the "stochastic orientation of the Penn Treebank and

[1]In at least one case, the order of application is specified in Shiyou, but the reverse order is followed in practice.

the resulting concern with sparse data". Because of this, they adopted the strategy of eliminating redundancy by taking into account both lexical and syntactic information. Thus for example, the Brown corpus has a scheme similar to the Penn one for tagging most verbs, but has separate sets for "be", "do" and "have". Penn eliminates these special tags, on the grounds that they can easily be re-inferred from the general tags and the particular word used if desired (thus, Penn "was/VBD" will always be Brown "was/BEDZ"). Marcus *et al* (p315) also say that "...since one of the main roles of the tagged version of the Penn Treebank corpus is to serve as the basis for a bracketed version of the corpus" – a function similar to supporting parsing – "we encode a word's syntactic function in its POS tag whenever possible. Thus, *one* is tagged as NN (singular common noun) rather than as CD (cardinal number) when it is the head of a noun phrase" – in contrast to the Brown corpus scheme, where it is always tagged as number. This is another important organizing principle; if we always give a certain word the same tag, based on its form, we are not adding any new information, but if it has a number of functions and we identify its function on each occasion of use, we are doing useful work.

Marcus *et al*'s point about sparse data is well made, but it is not always true that a smaller tag set will lead to better predictions. If a (lexically-based) distinction is made between two sets of words that have very different distributions, then the tagger may be able to use it to capture useful generalizations that exist at an intermediate level between those of individual words and fully general tags. Fortunately, distinctions that are fully lexically recoverable can be introduced and eliminated automatically, so there is no need to commit to a particular level of representation before creating a tagged corpus – it can be done later, during automatic tagging experiments, and can even be varied between domains and tasks.

Note that not all distinctions with a lexical component are fully lexically recoverable. The ATR tag set has a special set of tags for "have", but only for when it is used as an auxiliary verb. Collapsing these tags onto the main tag set for verbs would lose useful information.

3.1.4 Redundancy and syntactic recoverability

Marcus *et al* also argue for the elimination of certain tag distinction that are recoverable on the basis of syntactic structure, which the Penn treebank provides for much of the material it offers. Subject/object distinctions are a typical example. Since there are (as far as I know) no plans to tag sentences for syntactic structure at ATR, this argument does not apply directly to us. However, if a distinction depends on sentence-level syntactic structure, it is

unlikely to be computable by a tagger, as already argued in section 3.1.2 above. Therefore it should indeed be eliminated, though for different reasons from those that Marcus *et al* offer.

3.1.5 Conformity to existing schemes

Finally, it is desirable for tagging schemes to be *isomorphic*: it should be possible to specify a (simple) set of rules for converting accurately and fully automatically from one scheme to the other. To put it another way, any distinctions made by one scheme and not by the other should be lexically recoverable (see section 3.1.3 above).

The reason this is desirable is that it allows reuse in one scheme of any corpora already tagged in the other. In particular, any tagging scheme that is isomorphic to the Penn scheme can take full advantage of all the tagged material released by LDC. This can greatly increase the amount of training data available and hence, if the domains match closely enough, the accuracy of taggers.

This criterion also acts as a brake on radical alterations to any existing scheme: if an alteration requires manual re-tagging of large amounts of existing text, we should think twice about it. Of course, this only applies when the words being re-tagged have been accurately tagged in the first place, which is often not the case in the BTEC corpus. Indeed, part of the motivation behind proposing some of the changes to the ATR scheme is that the existing rules have apparently not proved easy to apply consistently.

3.1.6 Summary of tagging scheme criteria

To summarize, a tagging scheme should as far as possible:

- reflect the grammar of the language in question;
- reflect syntactic functions (roles of words) rather than forms;
- only define distinctions that automatic taggers – and, of course, human annotators – are likely to be able to make accurately;
- be isomorphic to any other tagging scheme for which large amounts of relevant accurately-tagged text are available, including previous versions of itself.

In what follows, I am therefore going to propose changing the ATR scheme in ways that make it better reflect the grammar of English as I understand it, are likely to be computable for taggers and annotators, and make it more

similar to the Penn scheme where the latter seems well thought-out, while avoiding wherever possible changes that will require manual re-annotation of accurately-tagged material. I will take each major part of speech in turn; then look at a particularly tricky set of issues involving adjectives, nouns and certain verb forms; and then summarize the proposals.

Throughout the document, I will make use of the following reference sources.

- I have taken Quirk, Greenbaum, Leech and Svartvik (1985; henceforth QGLS) as the reference standard for English grammar. This is as theory-neutral as it can be, and is generally recognized as the most comprehensive book on the subject. I know of no phenomenon of English grammar that it fails to cover.
- The ATR tagging scheme is defined in the (anonymous) Eigo Keitaigo Shiyousho (February, 2001; henceforth just "Shiyou"). Where English translations are given, they are by Saori Mine, except where otherwise stated.
- The equivalent Penn document is Santorini (1995).

3.2 Adjectives

The ATR adjective tags are `ADJ`, `ADJ-ER`, `ADJ-EST`, and `ADJTO`, though the last is only used for compound forms. The Penn tags are `JJ`, `JJR`, `JJS`.

Basically, `ADJ` = `JJ`, `ADJ-ER` = `JJR`, and `ADJ-EST` = `JJS`. See Santorini p12, paragraphs "JJ or JJR" and "JJ or JJS". The correlation between the Penn and ATR schemes is good in both theory and practice, though note that the words "elder", "greater" and "lesser" are always tagged `ADJ` in BTEC, despite being `JJR` in Penn. Also, "eldest" is `ADJ`, even though "greatest" is correctly `ADJ-EST`.

There are far more complex issues in connection with the difference between adjectives, nouns, and present and past participle verb forms. The latter set of issues are complex, and are dealt with in section 3.8 below.

3.3 Adverbs

The ATR tags for adverbs are `INTERJ`, `ADV`, `ADV-ER`, `ADV-EST`, `NOT`, `PNOM`, `PREADV`, `PREPADV`, `PRONADV`, `CONJADV`. The Penn tags are: `UH`, `RB`, `RBR`, `RBS`, `DT`, `RP`. The interrelationships are quite complex. QGLS (p438) remark that "Because of its great heterogeneity, the adverb class is the most

nebulous and puzzling of the traditional word classes. Indeed, it is tempting to say simply that the adverb is an item that does not fit the definition for other word classes."

3.3.1 Distinguishing interjections from adverbs

Adverbs are also often hard to distinguish from interjections; we will look at this issue first.

The interjection tags, ATR `INTERJ` and Penn `UH`, should match exactly. However, in practice, ATR often uses `ADV` when Penn uses `UH`. By far the most common example is "please", but others include some uses of "sure", "really", "certainly" and "right".

QGLS (sections 11.54 and 11.55, p852-3) describe *formulae* and *interjections*, which both have the key characteristic that they do not enter into syntactic relations. Both would seem to correspond to the `INTERJ/UH` tag. On this definition, "please" is an adverb (`ADV/RB`), i.e. the Penn tagging is wrong, because when it is used alone, it has the feeling of an elliptical version of something longer such as "Please do". Historically, "please" is indeed derived from an adverbial subjunctive phrase "if it please you". On the other hand, some of the other words, such as "well" (in the sense of "let me see") "sure" (meaning "yes"), and "right" (meaning "OK then") are interjections, so the ATR tag should be changed. The words ending "-ly" are less clear; QGLS list some of them as formulae, but to me they seem more like elliptical forms of longer sentences: e.g. after "Would you like to come along?", "Definitely" seems short for "I definitely would". I suggest we continue to treat these as adverbs at ATR.

Thus, roughly speaking, in cases of interjection/adverb uncertainty, "please" and "-ly" words should be classed as adverbs, and others as interjections. However, this rule of thumb is an epiphenomenon, not a definition.

3.3.2 The adverb "not"

The word "not" is always tagged `NOT` at ATR, but as an adverb `RB`, in Penn. Penn has a philosophy of not making tag distinctions that are lexically recoverable (see Section 3.1.3 above), which seems sensible to me. I would therefore prefer to abolish the `NOT` tag and tag "not" as `RB`, but it makes very little practical difference because the change would be automatic.

3.3.3 The PNOM tag

This is used for postnominal adverbial modifiers such as "o'clock", "ago" and "each". They are tagged RB in Penn, except for "each" which (dubiously in my view, since no other determiner appears in this position) is tagged DT.

This ATR tag seems fine to me. However, I notice that "a.m." and "p.m." are tagged CN; I think they should be PNOM as well, since they are syntactically identical and semantically very similar to "o'clock". The year modifiers "A.D." and "B.C." should be tagged PNOM for similar reasons.

3.3.4 The PREADV tag

PREADV is defined as "adverb which modifies verb, adjective and adverb" (Shiyou, 3.7.2), as opposed to ADV which is "adverb which modifies predicate" (Shiyou, 3.7.1). In addition, "-ly" words are not supposed to be PREADV (3.7.2). This initially is a little mystifying.

PREADV tends to correspond to RB in Penn (most uses) or DT (some uses of determiners such as "a little", "some", "any"). The Japanese term *settou-fukushi* for PREADV suggests it stands for "prefix-adverb".

The rules (Shiyou 3.7.1(3), p31, my translation) for distinguishing PREADV from ADV clarify the reason for having PREADV. For an adverb to be PREADV, it must:

a modify the immediately-following word (absolute condition);

b not like being modified (*shuushoku sarenikui*) (not take comparative or superlative);

c express degree.

In terms of QGLS's (7.46ff, p438ff) analysis of adverbs, PREADV seems to cover unmodifiable *intensifiers* (7.56) and *emphasizers* (7.57). It appears to be a fact about English, which I had not noticed before and which as far as I can see QGLS do not mention either, that "-ly" adverbs are nearly always modifiable, [2] while non "-ly" adverbs are less often modifiable (note the rather vague wording of item (b) above). For example, "almost" and "nearly" are virtually synonymous in English, but only "nearly" can be modified:

[2]One exception is "really" in its use as an intensifier: one cannot say "This coffee is very really hot", and "I'm more really exhausted" is marginal (it can only mean something like "I'm more genuinely exhausted [than you]"). One might argue that "only" is another exception, but it is not derived by "-ly" suffixation from an adjective, so it probably does not count.

I have nearly finished.
I have very nearly finished.
I have more nearly finished than you have.

I have almost finished.
*I have very almost finished.
*I have more almost finished than you have.

Thus, "almost" is tagged PREADV, and "nearly" is tagged ADV. However, the prohibition against PREADVs being modifiable is not absolute: one can say "It is much too/PREADV expensive", where "much" must be modifying "too", not "expensive", because one cannot say "It is much expensive".

The PREADV/ADV distinction is only one of many that could be made in the complex realm of adverbs. Having said that, the distinction is probably quite a good one for a tag set: it should be computable, because it is reflected both in word form and in word position (not modified, modifies following word). It can be made more exact as follows: for a word to be a PREADV rather than an ADV, it must

a modify the immediately-following word; and

b not take a comparative or superlative modifier (more, most, less, least);

c express degree; and

d not end in "-ly", except "only", "really" and a few other cases that may be detected.

A few problems seem to arise in the way PREADV is used in BTEC. Of the list of possible PREADV words given in Shiyou (3.7.2, p33),

- "a few" and "some" are (when tagged PREADV) actually determiners (in "some more coffee", "some" modifies "coffee", just as it does in "some coffee");
- "exact" and "plenty" are likewise never adverbs;
- "only" finishes in "-ly", but perhaps we can ignore that, as it's not derived by suffixation.

The other words in the list are valid PREADVs, but most of them are determiners too, and BTEC has a large number of cases where the determiner uses are tagged as PREADV. For example, "this" and "that" are valid PREADVs in phrases like "this flashy" or "that fancy" (Japanese *konna-ni* and *sonna-ni*),

but in fact 30 of the 41 occurrences of "this" as PREADV are either mistaggings or in ungrammatical sentences.

I think there is an argument for distinguishing PREADV from ADV. The way these two classes are translated into Japanese tends to be different, and it should be possible to make the decision automatically with reasonable accuracy. However, the condition for PREADV should be as suggested above, with the "-ly" rule either being dropped or being open to an expanding list of exceptions.

3.3.5 The PREPADV tag

We now look at three ATR tags that appear to represent an attempt (sometimes justified) to avoid or postpone making an important decision on whether a word is an adverb or not. These tags are PREPADV, for words that may be particles or adverbs; PRONADV, for words that may be pronouns or adverbs; and CONJADV, for words that may be conjunctions or adverbs.

In this section, we examine PREPADV, which corresponds to the Penn RP (particle) tag and some cases of the Penn RB tag. [3]

According to the Penn analysis (Santorini p21, "RB or RP"), some of the "particle" components of phrasal verbs are actually adverbs (RB) rather than particles (RP). The former allow modification but the latter do not. For example, "Bring the girls up" can either means "Educate the girls" or "Bring the girls here" (up the stairs). In the former case, "up" is a particle; in the latter case, it is an adverb. "Bring the girls right up", in which "right" modifies "up", can only have the "up the stairs" meaning.[4]

As this example makes clear, the distinction affects the meaning and is therefore important for translation. However, it is hard to imagine that a tagger could make the correct decision, so it is probably best not to try.

In any case, I'm not sure what the justification is for distinguishing PREPADV from ADV, since it is automatically recoverable (by looking at the lexicon entry for a word to see if it also has a PREP entry). However, for the same reason, it is harmless to leave it the way it is. See section 3.3.2 above for a similar point.

[3]Note that PREPADV is not used for any occurrences of words as prepositions, so the name is a bit confusing. A preposition has a dependent noun phrase or clause, while a particle does not. The confusion arises because words that can be particles can generally also be prepositions.

[4]This argument is unfortunately obscured by two typos in the Santorini paper.

3.3.6 The PRONADV tag

Our second "either/or" tag is PRONADV, which is used for "morpheme which functions as pronoun/adverb". It corresponds to some cases of RB, JJ and JJR in Penn (the Penn pronoun tags are rather restricted so do not correspond to any cases of PRONADV).

Examples of the RB case are place words like "somewhere", "someplace", "anywhere else" etc, and quantity words like "a lot", "much" and "a bit".

The JJ and JJR cases are for "much", "more" and "less", though these words can also be RB or RBR in Penn, as we'll see below.

This tag seems to be used for words that sometimes function adverbially and sometimes as direct objects. It appears to be a device for avoiding making a difficult decision over particular cases. For example, "Can you find me somewhere?" probably means "Can you find me a place?" and so "somewhere" is a direct object. But "Can you meet me somewhere?" must mean "Can you meet me in some place?", so "somewhere" must be adverbial. Interestingly, time words like "sometime" and "anytime" are classified as ADV not PRONADV; although they are exactly parallel to "somewhere" and "anywhere" in their adverbial use, they do not have a direct object use (the direct object involves breaking the word in two: "Can you spare me some time?").

Note that Penn does attempt to make this distinction. Santorini (p26, paragraph on "more") makes it clear that "more" should be JJR when it is the direct object (or subject?) of a verb, but RBR when it is used adverbially. Thus in "You should eat more", the tag would be JJR if the meaning is "more food", but RBR if the meaning is "more often". (We can see there is a real distinction here by passivizing the sentence: "More should be eaten by you" can only have the "more food" meaning).

Linguistically, PRONADV cannot really be justified, but in practice, it seems like a good means of postponing (until after tagging) a decision that cannot be made accurately in tagging. So I propose it should be kept.

3.3.7 The CONJADV tag

The third "either/or" adverb tag is CONJADV, which is a subset of Penn RB. It is defined as "morpheme which functions as conjunction/adverb".

Like PREPADV and PRONADV, it appears to be intended as a means of avoiding possibly difficult decisions, in this case between CONJ and ADV. It is used for five words in LDB and BTEC: "so", "though", "just in case", "however" and "otherwise". Most of these have the property that they can occur either as conjunctions or as adverbs, e.g.: "I'd like to go though I'm not

sure if I can", "I'd like to go, though". The exception is "however", which cannot occur as a conjunction. There are other words, such as "before", "since", "once", "although" and "ever since" in LDB and BTEC, which can occur both as conjunctions and as adverbs, but are never tagged CONJADV.

Whether the decision between CONJ and ADV is difficult depends mostly on how good sentence boundary detection is. If multiple-sentence utterances are likely to be common and their transition points are hard to detect, it will be hard to make the decision, and there is justification for keeping this tag. Otherwise, I see no justification for it.

If the tag is kept, it should be used for the right set of words, excluding "however" and including the extra words listed above.

3.3.8 Existential and locative adverbs

Both ATR and Penn distinguish existential and locative uses of the "there". Both tag sets use EX for existential "there"; Penn uses the usual RB tag for locative "there", while ATR uses a special LOCADV tag for locatives, which is also used for "here", "abroad" and phrases like "over there".

The ATR ADV/LOCADV distinction would appear to be lexical. If we need a special tag for locative adverbs, we should presumably also have one for the temporals "now" and "then", but distinguishing temporals from other uses is not easy: in "then let's go", only context can tell us whether "then" means "after that" or "in that case". I therefore propose that LOCADV be merged into ADV, but since it is a lexical distinction, it is easily changed.

3.4 Nouns, numbers, names and symbols

In this section, we will look at the ATR tags CN, CN-PL, NUM, NUM-ORDINAL, PRENOM, PROPN, PROPN-PL, LETTER and LETTER-PL, which correspond to (some cases of) the Penn tags NN, NNS, NNP, NNPS, CD, JJ and SYM.

The complex issue of distinguishing nouns from adjectives and verb forms is dealt with separately, in section 3.8.1 below.

3.4.1 Noun singular-plural distinction

Singular and plural nouns are represented in ATR by CN and CN-PL respectively, and in Penn by NN and NNS respectively.

These should in theory correspond exactly. However, in practice the singular-plural distinction appears to be a bit different. Basically, Penn concentrates on syntactic agreement, while ATR focuses on morphology.

In Penn, the rules are quite complicated:

- The main criterion is not the noun ending, but the verb agreement that is triggered when the noun is a subject (Santorini p17-18, "NN or NNS"). Thus,

  Linguistics/NN is a difficult field.
  The police/NNS have arrived on the scene.

- It follows from this that "if a noun is semantically plural or collective, but triggers singular agreement, it should be tagged as singular:

  The group/NN has/*have disbanded."

- When a word triggers variable agreement – i.e. it can take either a singular or a plural verb – it is tagged "according to usage in a particular text", or if this cannot be determined (because the noun does not occur as subject of a present-tense verb) it is tagged ambiguously, as NN|NNS.

- "Amount" noun phrases are an exception to the agreement rule: they are NNS even though the verb is singular. "Three years/NNS is a long time".

- Although Santorini does not state this, when collective nouns like "police" are not the subject of a verb that shows number agreement, they tend to be tagged NN, even if they are tagged NNS elsewhere in the same corpus when they do occur as present-tense subjects. Thus, we have

  The police/NNS have arrived on the scene.

  because of the agreement rule above, but

  They called the police/NN.
  The police/NN could check up on you.
  I called the police/NN department.

I do not think these criteria are satisfactory, at least in the way they are applied in practice. As well as being rather inconsistent, they are subject to idiolect differences: in Santorini's example above, "The group have disbanded" is completely acceptable for me, so I would tag "group" as NN|NNS.[5]

In ATR, the criterion appears (from the BTEC data rather than from Shiyou) to be primarily morphological:

[5]I suspect this is a British-American difference. The names of sports teams consistently take singular verbs in the American press, but plural ones in the British press.

- If the noun is the plural form *of a singular noun with the same meaning* (see below), it is CN-PL, otherwise it is CN. The plural form can be irregular: thus "people", "men" and "feet" are all CN-PL.

- If the singular and plural forms are the same, as for "fish" and "sheep", the tag depends on the meaning in the particular context. If several entities are meant, it is CN-PL; if only one, or if it is indeterminate, then CN. Thus, "a fish/CN", "two fish/CN-PL", "the fish/CN".

By the first rule, "police" and "group" are CN. "Linguistics", "news" and "customs" (in the sense of the people who search your bags) are also CN, because although the words "linguistic", "new" and "custom" exist, the first two are not nouns, and the third has nothing to do with searching bags.

Whether the Penn or ATR approach is better depends on the purpose of the tagging. For English-to-Japanese translation, it probably does not matter very much, because of the lack of number distinctions in Japanese nouns. Even for translation to languages that do have number, it is unclear which is better: it would depend on how the target language manages its number system. I therefore propose that the ATR scheme be kept, because it is much simpler to apply.

3.4.2 Nouns and cardinal numbers

As we have seen, nouns are tagged CN or CN-PL in ATR, and cardinal numbers are tagged NUM. The corresponding Penn tags are NN, NNS and CD, respectively.

The distinction can be hard to make, especially (but not only) for the word "one". For Penn, Santorini (p7-8, "CD or NN"), says:

> In general, ["one"] should be tagged as a cardinal number (CD) even when its sense is not clearly that of a numeral: "one/CD of the best reasons". But if it could be pluralized or modified by an adjective in a particular context, it is a common noun (CN): "the only one/NN of its kind".

In the latter case, we can have "the only good one of its kind" or "the only ones of their kind". The impersonal "one" meaning "someone" is also usually tagged as CN, although PRON would be better. The "one" in "another one" is usually tagged NUM when it is followed by a phrase like "of the same kind"; I think it should be CN as with unmodified cases of "another one", because it still allows adjectival modification, as in "another good one of the same kind".

In ATR, similar criteria seem to be applied. The rules (Shiyou, 3.1.1(4), p13, my translation) are that "one" should be NUM except:

- when it's modified by a DET, ADJ or CN, and does not itself modify a noun.
- when it's plural (i.e. the word "ones").
- when it's the head of a relative clause (e.g. "one which...").

Note that once again, the Penn guidelines appeal to native-speaker intuition ("...*could be* pluralized or modified ...") whereas the ATR ones simply depend on what is present ("...when it *is* modified ..."). Thus one would expect "one" to be tagged CN less often in ATR than it is tagged NN in Penn. In fact, the opposite is the case; "one" is CN 31% of the time in BTEC, but only 8% of the time in the Penn Treebank. Of course, the corpora are different, but the difference in these ratios is still striking. It is also striking how many mistakes and inconsistencies there are in both corpora.

I think the whole treatment of "one" in both schemes is misconceived, because it confuses a basic class of words (numbers) with their functions on particular occasions (as nouns, or otherwise). In particular, neither scheme allows numbers other than "one" to be tagged as nouns, although they can function that way.

QGLS (6.54-56, pp386-388; 12.15-16, pp869-870) clarify the situation. They say that "one" has three uses: numerical, substitute, and generic. When it is numerical, it can (like other words such as "some" and "this") appear either as a determiner or a complete noun phrase. When it is a substitute, it can substitute for any part of a noun phrase that includes the head. To these classes, we can add a "pure number" class where a number word is (part of) a phone number, address, product model number, etc, rather than forming part of a noun phrase obeying the normal rules of syntax. Some examples:

- Numerical, determiner: (the) one boy.
- Numerical, NP: one of the boys.
- Substitute: I'd like a drink, but just a small one.
- Generic: One can't be too careful.
- Pure number: My extension is one three seven six, and I am in room one.

Note that "one" can be replaced by another number in all but the generic case:

- Numerical, determiner: (the) two boys.
- Numerical, NP: two of the boys.
- Substitute: Give me a couple of those drinks – just the smaller two.
- Generic: *Two can't be too careful.
- Pure number: My extension is two three seven six, and I am in room two.

Note that substitute "one" can be pluralized to "ones", and the numerical-NP "one" to "some", while the others cannot be pluralized at all.

As a starting point for discussion, I suggest that "one" (and other number words, where applicable) be given a different tag for each of these five cases. Such a tagging would be useful for translation, since each case is (I think) translated differently into Japanese. Most of the distinctions look computable, because the syntactic contexts will be quite different, except that it may be difficult to distinguish generic "one" from numerical-NP "one" (both function as complete NPs). Since generic "one" is formal and rare or absent in BTEC, it may be acceptable to give it the same tag as the numerical-NP case.

Thus we might end up with the following tags:

- Numerical determiner: `NUMDET` . Not `DET`, because it is not interchangeable with ordinary determiners: we can't say "two the boys".
- Numerical NP and generic: `PRON`, by analogy with words like "this", which can be `DET` or `PRON` in a similar way.
- Substitute: `CN`. Syntactically this is how it behaves, and the distinction between numbers and other types of `CN` is lexically recoverable if it is needed.
- Pure number: `NUM`, as at present.

We also need to decide how to handle multiple-word numbers like "thirty six" and "four hundred and ten". The cleanest solution is probably to give each word the same tag as would be given to a single-word number, with the exception of "and" and "point" which would be labelled `CONJ` and `CN` as usual. However, it might be hard for a tagger to do the right thing with long numbers, so alternative would be to give only the last (or first?) word the "right" tag, and tag all the others as `NUM`.

This proposal involves introducing only one new tag, `NUMDET`. However, it would involve retagging every number word in the BTEC corpus, so would be a non-trivial amount of work.

3.4.3 Nouns and ordinal numbers

In ATR, the tag NUM-ORDINAL is used for all singular ordinal numbers, whatever their functions: "June the third", "the third floor", "the third largest hotel", "one third". This essentially means postponing any decision on these words until a later processing stage, which seems rather unambitious.

In Penn, the adjective tag JJ is used for the first two of these cases, the adverb tag RB for the third one, and the noun tag NN for the last. The Penn approach thus involves distinguishing the actual functions of the words, rather than subsuming them all under the same (lexically-determined) tag. I think this is preferable; if we distinguish noun, adjective and adverb uses for most words, and (as proposed) make similar distinctions for cardinal numbers, why should we not do it for ordinals too?

The ATR approach is also inconsistent in another respect. because of the non-ordinal fraction-denominator words "half" and "quarter", and the occurrence of plural ordinals in expressions like "two thirds". In ATR, we have "one third/NUM-ORDINAL" but "one half/CN" and "two thirds/CN-PL". Penn tags these consistently as NN (or NNS for "thirds").

I propose that the Penn approach should be adopted: the NUM-ORDINAL tag should be abolished, and these words tagged CN, CN-PL, ADJ or [PREP]ADV as appropriate.

3.4.4 Nouns and proper names

ATR distinguishes singular (PROPN) and plural (PROPN-PL) proper names. Penn does the same, with the tags NNP and NNPS.

ATR makes an additional distinction with the PRENOM tag, used for titles like "Mr.". This is obviously a good idea, as they need to be treated specially in translation.

The ATR distinction between PROPN and PROPN-PL is a little odd. Plural names that are more usually seen as singulars – "Cokes", "Budweisers", "(the) Browns" – are tagged PROPN-PL, but group names like "the Beatles" and "the Yankees" tend to be PROPN. I think the latter should be PROPN-PL too, since they are plural both morphologically and syntactically, and the singular forms do occur ("Paul McCartney, the former Beatle").

Penn and ATR draw the boundary between nouns and names in slightly different places. For Penn, a proper name is any capitalized word (other than the pronoun "I", symbols such as "X", and ignoring sentence-initial capitalization). For ATR, weekday and month names are ordinary nouns, but other day names like "Christmas" or "Christmas_Day" are still proper names.

I slightly prefer the Penn approach because it is easier to apply: basing the decision on capitalization is a clear criterion. Weekday and month names do behave like common nouns more often than other names do, because they are often modified by determiners: "this December", "every Thursday", so one can see the reason for the ATR approach. On the other hand, it is sometimes difficult to draw the line between names that refer primarily to a period of time and those that refer primarily to an event or activity that occurs at that (same) time every year. Also, even ordinary personal and place names can be used like nouns: "The Paris in France is much bigger than the other Paris in Texas".

3.4.5 Letters and Symbols

The ATR `LETTER` tag and the Penn `SYM` tag partially overlap. The latter is for "mathematical, scientific and technical symbols or expressions that aren't words of English" (Santorini, p5). Often these are letters, especially in the BTEC corpus where expressions like "z company" are often used as variables for any company name. However, not all symbols are single letters, and not all single letters are symbols (occasionally they are names, as in "Malcolm X"). Also, in BTEC, letters are sometimes used as variables (symbols), but sometimes also to stand for themselves: "My name is Yamada, Y A M A D A". In fact, BTEC uses `LETTER` for some (Japanese) syllables like "ma", which is understandable, but also for other terms such as "gg" and "url". These latter are presumably errors.

ATR also has a `LETTER-PL` tag, for plural words like "B's". Penn seems to use `NNS` for this, which seems a little inconsistent with the singular form.

On the other hand, ATR is also inconsistent: when multi-letter variables are used in expressions like "zzz hotel" (exactly parallel to "z company") they are tagged `CN`.

I propose that the ATR `LETTER` and `LETTER-PL` tags be used only for letters that stand for themselves, i.e. in spellings-out and in expressions like "from A to B" or "mind your P's and Q's" which are meant to be spoken as written. For expressions like "Z hotel", where you are supposed to substitute a specific hotel name, we should either introduce a new `SYM` tag, or use the tag that would be appropriate to whatever is actually substituted (presumably `PROPN` in "Z hotel"). The second option is probably easier, and has the advantage of not adding to the tag set.

3.5 Conjunctions and prepositions

ATR has tag CONJ for conjunctions and PREP for prepositions. Penn has CC for conjunctions and IN for prepositions.

However, ATR CONJ has a wider scope than Penn CC. Prepositional words governing clauses are marked CONJ in ATR, but IN in Penn: "I left before/CONJ those two people arrived." Linguistically the ATR distinction is reasonable, but it sets a tagger a difficult task: that of deciding whether the noun phrase following the word is itself the object of the preposition, or is the subject of a clause that is the object. In other words, in the above example, the tagger would have to look past "those two people" to find "arrived" and work out that it is the object of "before". In principle, this is a full parsing problem and should be tackled further downstream. However, in practice it may not be so bad. An analysis of fifty randomly-chosen instances of "before/PREP,CONJ" in BTEC suggested that 96% would yield correct results to the rule "assign CONJ if the following word is a PRON, else assign PREP". Thus the ATR solution may in practice be better for BTEC, because it encodes a useful and (probably) computable distinction, whereas the Penn distinction is purely lexical. However, other corpora may have different properties, as may other words that "before".

The word "to" is a special case, because as well as being a preposition, it is also an infinitive marker. ATR marks it as a PREP whatever its function. Earlier versions of Penn (e.g. Wall Street Journal) also had a single tag (TO) for "to", but later ones (e.g. Switchboard) do make the distinction, using PREP for the prepositional function and TO for the infinitive one.

This distinction is a useful one for translation, and it should nearly always be quite feasible for a tagger to make it, since it just involves looking at whether the material immediately to the right looks like the start of a verb group or the start of a noun phrase. I propose that the TO tag be introduced to the ATR tag set, and used for infinitives.

3.6 Determiners, pronouns and clitics

Here we deal with the ATR tags DET, DETADJ, NO, PRON, PRON-PL, and POSS, which between them correspond to the Penn tags DT, PRP, PRP$ RB, PDT, NN and NNS.

3.6.1 Determiners: DET, DETADJ and NO

These three tags cover the space of determiners,[6] predeterminers and some adverbs, corresponding to some of Penn DT and RB, all of Penn PRP$, and all of Penn PDT. The NO tag is simply the word "no" (in its determiner rather than interjection form).

The definition of DET (Shiyou, p27, my translation) is "article or possessive pronoun", and DETADJ is "article-adjective [*kan-keiyoushi*]; it can be followed by a determiner (DET); it modifies pronouns, PRONADVs and noun phrases". The main examples in LDB and BTEC, with their Penn equivalents, are:

- ATR NO, Penn DT: no.
- ATR DET, Penn DT: a, an, another; the; this, that, these, those; every, each.
- ATR DET, Penn PRP$: my, your, his, her, its, our, their.
- ATR DETADJ, Penn DT: any, some, all, a few, a lot of, a little, a couple of, lots of, plenty of, both.
- ATR DETADJ, Penn RB: about, just, only, around, at least, almost.
- ATR DETADJ, Penn PDT: all, half, such, quite.

The rationale for DETADJ seems to be an attempt to account for a range of, in my view, unrelated phenomena. These phenomena are:

1. adverbs preceding noun phrases: "He is only/DETADJ a boy", "He never reads even/DETADJ magazines."
2. genitive quantifier phrases: "I drank a-lot-of/DETADJ the beer", "I spent hundreds-of/DETADJ dollars", "I spent about/DETADJ a thousand dollars."
3. determiners followed by comparative pronouns (tagged as PRONADV, see section 3.3.6): "I drank some/DETADJ more", "I drank quite/DETADJ a-lot/PRONADV.
4. genuine predeterminers: "I drank all/DETADJ the beer", "Why did you do such/DETADJ a thing?".

[6] I take the word "determiner" to subsume both articles such as "the" and possessive pronouns such as "my". I think this usage is standard.

What all these cases have in common is that the word or phrase tagged DETADJ is immediately followed by something that is viewed in the ATR scheme as a pronoun (PRON), a PRONADV or a full noun phrase. However, they seem to me to have nothing else in common. Taking each in turn:

1. "Only" and "even" are straightforward adverbs modifying the verb. Compare "He is a boy only", and "He never even reads magazines". Even in the utterance "Only a beer", "only" does not modify "a beer" – rather, both "only" and "a beer" can be seen as modifying an elided structure such as "I would like...".

2. Syntactically, phrases like "a lot of beer" are best viewed as a noun phrase ("a lot") modified by a prepositional phrase ("of beer"). It seems bizarre to view "a lot of" as a kind of predeterminer, especially since a very wide range of nouns can take the place of "lot" (variety, range, touch, total, ...) or of "a lot" (millions, loads, heaps, ...).

3. It is not clear to the that the existence of phrases like "some more" and "any more" justifies classifying "some" and "any" with a different tag from, say, "every". "Some" and "any" cannot modify all PRONADVs in this way, just the comparatives "more" and (in the case of "any") "less", so phrases like "some more" are probably best viewed as idiomatic, and tagged neutrally as "some/DET more/PRONADV".

4. This leaves us with the predeterminers, corresponding to the Penn PDT tag.

I'm not sure of the purpose of the NO tag, unless it is to avoid having to make the choice between a determiner ("no books") which would be a DET and an adjective modifier ("no better") which could be a DETADJ (the interjection use should not be too hard to distinguish). This choice is not too hard to make. Also, "any" can occur in exactly the same places, and it doesn't have its own special tag, so why should "no"?

Note that Penn distinguishes DET from PRP$. However, this distinction is lexically determined, so by Penn's own philosophy it should be absent. ATR seems OK in this regard.

I propose that DETADJ be used only for predeterminers, and perhaps be renamed PREDET. Its current RB members should be reclassified as PREADV and its DT members as DET. The phrases such as "a lot [of]" that are currently tagged as DETADJ should be split into their component words, or at least have the "of" detached and be given a more standard tag such as CN or PREADV.

I also propose that NO be abolished and the word "no" treated like any other determiner and classified as DET or PREADV. I propose that adjective

modifiers (as in "any better") should be PREADV, or if desired, have their own special tag.

3.6.2 Personal and possessive pronouns

The ATR scheme for these types of pronoun is as follows:

- POSS is used for "mine", "yours", "ours" and "hers" – what QGLS (6.2, p336) call *independent* genitive pronouns, in contrast to "my", "your", "our" and "her" which are *determinative* ones.
- PRON-PL is used for plural *reflexive* pronouns ("ourselves", "yourselves", "themselves", and for the non-pronoun "others".
- PRON is used for all other pronouns, including plural non-reflexives such as "we" and "them".

The ATR definition of a pronoun is wider than the Penn definition. Firstly, determiner words that can also act as stand-alone NPs (definites "this", "that", "these", "those", and quantifiers "all", "some", "any", "both", "each", "another", "either" "neither" and "a few"), are tagged PRON when they do so, whereas they are always DT in Penn. Secondly, indefinite pronouns of the form "(some/any/no)(thing/one/body)", and also "none" are tagged PRON in ATR but NN in Penn.

The wider scope of the ATR PRON tag seems preferable to me, covering words that are generally viewed as pronouns in the literature (see e.g. QGLS chapter 6, and Halliday and Hasan). The differences are all lexically determined, except that ATR requires the determiner/pronoun choice to be made for "this", "that" etc, which Penn does not. I think ATR is better here, because this is a sensible choice to expect a tagger to be able to make, and it is clearly an important one for translation (e.g. determiner "this" goes to *kono*, pronoun "this" goes to *kore*).

Both ATR and Penn have a tag for possessive pronouns (POSS and PRP$ respectively), but they have different scope. Both include the independent possessives (mine, yours, etc.) but the Penn tag additionally includes the determinative possessives (my, your, etc.) which ATR includes with the non-possessives. The consequence is that Penn forces one to decide whether "her" is genitive or not ("give me her/PRP$ book", "give her/PRP the book"; both PRON in ATR). On the other hand, ATR forces one to decide whether "his" is determinative or independent: "it is his/PRON book", "the book is his/POSS"; both PRP$ in Penn). Both decisions should be possible for a tagger; the latter seems to be rarer, as there appear to be no cases of independent "his" in the BTEC corpus.

It is a bit awkward that ATR makes a singular-plural distinction but that non-reflexive plural pronouns are classed with the singulars. The reason is presumably that it is only with reflexives that one can recognize number distinctions in the second person (yourself vs yourselves). It is in general not possible to know whether "you", "your" and "yours" are singular or plural without knowing the extrasentential context. As far as I can see, the distinction between PRON and PRON-PL is lexically determined anyway, so I don't think it's needed.

I propose that the PRON-PL tag be abolished, and the words tagged with it should be relabelled PRON, apart from "others" which is not a pronoun and should be tagged CN-PL. Keeping PRON-PL initially seems attractive, but it is awkward because we would need to decide to do what with "you" and its variants. A true singular-plural choice would be impossible for a tagger to make, so we would have to make them always PRON or always PRON-PL. The latter would be better, because "you" as subject triggers the plural form of the present-tense verb ("you go"); but "I" also does this for verbs other than "be", and we can hardly call "I" plural.

3.6.3 Adverbial WH words

The ATR tags HOWADV, WHCONJ and WHADV between them correspond to the adverbial-WH-pronoun Penn tag WRB. (Note that "how" is an honorary WH-word; the "w" was accidentally deleted by a medieval scribe and replaced at the end).

HOWADV is used for "how" when it modifies an adjective or adverb, as in "how quickly" or "how good" ("how much" and "how many" are treated as single lexical items and given other tags). It is not used in cases like "how did it go?".

WHCONJ is used for "conjunction which starts with interrogative". Examples are "when" and "whenever" in constructions like "The ground gets wet when[ever] it rains". It should presumably be used for "where" and "wherever" in such contexts too, but there appear to be no such uses of "where" in BTEC. There are three uses of "wherever", but they are tagged WHADV, wrongly I think. Several "whenever"s are also tagged WHADV, which also must be wrong.

WHADV is used for "where", "how", "when" and "why" when they are used as adverbial WH phrases on their own: "where/how/when/why did you go there?".

These all seem like sensible distinctions to make, so I propose the scheme should be unaltered. However, the "wherever/whenever" mistaggings as WHADV should be corrected to WHCONJ.

3.6.4 WH Pronouns and WH Determiners

There is a distinction between WH pronouns proper – ones that form NPs on their own, as in "what are you reading?" – and WH determiners, which modify head nouns: "what book are you reading?".

In Penn, the tags are basically WP for WH pronouns and WDT for WH determiners. However, for reasons that Santorini does not explain, "whose" has the separate tag WH$ for both these uses. Also, it seems strange that Penn makes the pronoun-determiner distinction for WH words, but not for TH words (this, that etc) which can also function in both ways. Another oddity is that "whichever" and "whatever" are always WDT, even when they are WH pronouns.

In ATR, the tags are WHPRON for WH pronouns and WHADJ for WH determiners.[7] So roughly, Penn WHP corresponds to ATR WHPRON, and Penn WDT to ATR WHADJ.

However, relative pronouns are treated differently. The "which" in "The book which I'm reading" is tagged WHPRON in ATR but WDT in Penn. Thus, an ATR tagger needs to distinguish relative pronouns from WH determiners, and a Penn tagger needs to distinguish them from WH pronouns. The latter task is probably harder.

Unlike Penn, ATR (I think) also distinguishes pronoun and determiner uses of "whose", although the only pronoun use I can see in BTEC is actually tagged WHADJ, along with all the determiner ones.

Ideally, we would have a three-way distinction here, with tags WHPRON, WHDET (the current WHADJ – see footnote) and a new WHREL tag for relatives. All three classifications lead to different translations into Japanese, so all distinctions are important. If this is not possible, the current ATR distinctions should be preserved.

Note: BTEC has five examples of "how much a ..." (day, etc.) where "how much" is tagged WHADJ. However, this "how much" does not modify "a day" – the sentence can be viewed as elliptical for "how much [is it for] a day" – so the tag should be WHPRON.

3.6.5 $S

$S is the possessive marker for 's and '. It corresponds exactly to Penn POS.

[7] I think this tag should really be called WHDET rather than WHADJ. We can see that these words stand in for determiners not adjectives because, like determiners, they cannot occur with other determiners in the noun phrase – "which the book?" is ungrammatical – and they must precede any adjectives in the noun phrase – "which red book?" is OK, but "red which book?" is not.

The tag itself is straightforward, but detaching possessive markers and treating them as separate lexemes does not go well with speech recognition, because, as already argued in Section 2.1, it indirectly forces the recognizer to make a choice between the plural, singular-possessive and plural-possessive inflection of every noun and proper name. It was recommended there that words like *book's* and *books'* should be kept as single lexemes, and tagged `CN-POSS`, `CN-PL-POSS` `PROPN-POSS` and `CN-PL-POSS` as appropriate.

3.7 Verbs

I will look at auxiliaries and the special verbs "be", "do" and "have" in a moment, but first let's concentrate on open-class ("standard") verbs.

3.7.1 Standard verbs

Both ATR and Penn distinguish base forms ("go"), third person present singulars ("goes"), -ing forms ("going"), past tense forms ("went") and past participle forms ("gone"). Note that for standard verbs all these distinctions, except (often) the past tense / past participle one, can be made on the grounds of word form alone.

Penn makes one additional distinction ATR does not: it splits base forms between infinitives ("to go") and non-third-person-singular present forms ("they go"). This involves a "real" tagging decision as the word forms are the same.

ATR makes a distinction that Penn does not: it splits -ing forms between present participles ("John is going") and gerunds ("Going is not a good idea at the moment"). This is also a "real" tagging decision, and sometimes quite a tricky one. It is important for translation, though. See Section 3.8 below for details.

I think it would be useful, and feasible, to distinguish the two uses of base forms in the way Penn does, introducing a `V-P` ("plural", but also first and second person present) tag. We would then have these pairings: `V:VB`, `V-P:VBP`, `V-3S:VBZ`, `V-INGG:VBG`, `V-INGP:VBG`, `V-PAST:VBD`, `V-PP:VBN`.

ATR also has a `VTO` tag for spoken-form words such as "wanna" and "gotta". This kind of variability should be handled within the speech recognizer, which should output "want to", "got to", etc. `VTO` should then not be needed.

3.7.2 "Be" verbs

Unlike Penn, ATR has a separate scheme for forms of the verb "be". The root tag is BEV in place of V. Because this verb varies more than standard verbs with number and person in both present and past tense, additional tag distinctions are easy to make, though since they are purely lexical, it's unclear how useful they are.

The infinitive form "be" is just BEV, with the present tense forms distinguished as BEV-1S, BEV-1P, BEV-2, BEV-3S, BEV-3P. Note that no attempt is made to distinguish BEV-2S and BEV-2P for "are"; see Section 3.6.2 for this.

Penn has a special tag BES which is used for the clitic "'s" when it means "is". This corresponds to ATR's BEV-3S.

Past tense forms BEV-PAST-1S, BEV-PAST-1P, BEV-PAST-2, BEV-PAST-3S, BEV-PAST-3P are also distinguished. These are not trivial, because "were" can be any of these forms except BEV-PAST-3S. It requires the subject of the verb to be identified.

Participle forms BEV-INGG, BEV-INGP and BEV-PP exactly parallel the corresponding standard verb tags.

3.7.3 "Have" verbs

Similarly, ATR has a special set of tags for "have", delineating the different forms: have/HAVEAUXV, has/HAVEAUXV-3S, having/HAVEAUXV-INGG (and perhaps -INGP), had/HAVEAUXV-PAST, and had/HAVEAUXV-PP. As the tag name implies, these are only used for the auxiliary function of "have": "I have/HAVEAUXV bought a drink", but "Now I have/V a drink". An ATR tagger, unlike a Penn one, therefore needs to decide whether each occurrence of "have" is auxiliary or not. This is an important decision, and one that should be within the capabilities of a tagger.

Penn has a special tag HVS for "'s" when it is short for "has". This corresponds to ATR HAVEAUXV-3S (only the auxiliary "has" gets contracted like this, at least in my idiolect).

3.7.4 Other auxiliary verbs

ATR has a tag AUXV for other auxiliary verbs, which subsumes all instances of Penn's MD (modal) tag. However, ATR's definition of auxiliaries is functional, while Penn's is (uncharacteristically) morphological: a modal is verb that is the same in third-person present singular and plural (Santorini, p3). This means that auxiliary forms of "do" are AUXV in ATR but V (etc) in Penn.

Therefore, just as with "have", an ATR tagger needs to distinguish auxiliary from standard uses of "do": "They do/AUXV not do/V it".

There are special ATR tags AUXV-3S for auxiliary "does" and "doesn't", and AUXV-PAST for "did". (There are no AUXV-INGG, AUXV-INGP, or AUXV-PP, as auxiliaries do not occur in these forms). The tag AUXV-PAST is also used for "could", "would" and "might" and their negative (-n't) forms. However, only some uses of these modals are actually past versions of "can", "will" and "may" ("he said that he would/AUXV-PAST do it"). They are more usually hypotheticals or counterfactuals: "I would like a drink" is not a past form of "I will like a drink", but carries an implication of something like "...if you were willing to give me one". It is probably too difficult for a tagger to distinguish these uses from true past tenses. I therefore think that AUXV-PAST should be abolished, since it does not necessarily imply "past tense", and since we cannot in practice distinguish past-tense modals from hypotheticals, the distinction becomes a purely lexical one.

3.8 Adjectives, nouns, gerunds and participles

There are multiple differences between Penn and ATR in the way adjectives, nouns and certain verb forms are distinguished when they are used in adjectival positions (either modifying a noun, or after a verb like "be" or "become"). The issues here are interrelated and very complex.

Here are some clear cases, with Penn tags being shown after ATR tags:

Adjective:	That is a good/ADJ/JJ idea.
Noun:	What is your phone/CN/NN number?
Gerund:	Eating/V-INGG/VBG caviar is enjoyable.
Present participle:	He was eating/V-INGP/VBG caviar.
Past participle:	He was arrested/V-PP/VBN by the police.

Confusions can arise because:

- words of any form can be adjectives or nouns/names "terminal/CN type" vs "terminal/ADJ disease"; "German/PROPN lesson" vs "German/ADJ visitor".
- "-ing" words can be adjectives, nouns, gerunds or present participles: "a surprising/JJ result", "a living/V-INGP creature", "the living/CN|V-INGG room".

- "-ed" words can be adjectives or past participles: "decaffeinated/JJ|V-PP coffee".

Let's look at each of these dilemmas in turn, comparing the ATR and Penn recommendations. Note that Penn does not distinguish between gerunds (ATR V-INGG) and present participles (ATR V-INGP), using the same VBG tag for both.

3.8.1 Adjectives and nouns/names

The distinction between nouns (CN; Penn NN) and names (PROPN; Penn NNP) is dealt with elsewhere, in section 3.4.4. Here, we will just look at how these two tags can alternate with the ADJ (Penn JJ) tag.

Santorini (p12, "JJ or NN" paragraph) explains how to distinguish adjectives and nouns in the Penn scheme. The rules are basically:

- Nouns used as modifiers are tagged as nouns. Thus "wool/NN sweater", "terminal/NN type", but "woollen/JJ sweater", "terminal/JJ disease". To judge whether the modifier is a noun or an adjective, one can make a sentence out of the compound: "the sweater is made of wool" but "*the sweater is made of woollen"; "*the type is terminal" but "the disease is terminal".

- Hyphenated modifiers are always adjectives. Thus "income-tax/JJ return" but "income/NN tax/NN return". This seems a little odd to me!

- Gradable prenominal modifiers – ones that can be modified by an adverb, or made comparative/superlative – are adjectives: "a fun/JJ party" because "a really fun party", but "a cocktail/NN party" because "*a really cocktail party".

- Color words are nouns only when they are used as nouns: "That's a nice red/NN", "These plants are a dark green/NN" (contrast "These plants are dark green/JJ).

- Adjectives serving as noun phrase heads are still adjectives if they can be modified by adverbs: "the (very) rich/JJ pay far too few taxes". If they can't be, they are nouns: "Little good/NN will come of it", but "*Very good will come of it". Note that in the first case, the noun phrase refers to all members of the set described by the adjective: "the very rich" means the set of all very rich people, but "little good" does not refer to the set of all (little) good things or people.

The ATR rules for such cases (Shiyousho, 3.1.1 (2)) are based on the parts of speech listed for the words in the READERS dictionary. If only one of CN and ADJ is listed, that tag is used. If both are listed, the procedure is as follows.

- CN is used if the word is for a colour or material; a religion, principle or idea; a country; or is "half", "extra" or "top".
- Otherwise, ADJ is used if one (or more) of the following four conditions applies:
  - it can take an "-er" or "-est" modifier;
  - it is a form derived from a noun – presumably something like "moonlike" or "hopeful";
  - it is (note: "is", not "can be") modified by a degree adverb like "very" or "much";
  - it is one of the words "single" or "double".
- Otherwise, CN is used.

These rules lead to many cases which Penn would treat as adjectives but ATR treats as nouns. For example:

- "*red* shoes": an adjective in Penn because it's gradable ("very red shoes", "redder shoes"), but a noun in ATR because it's a colour word. However "*reddish* shoes" would presumably be an adjective in ATR because "reddish" isn't (or shouldn't be) defined as a noun in READERS.
- "a *terminal* disease": an adjective in Penn because "terminal" meaning "fatal" is always an adjective, but a noun in ATR because of the unrelated noun "terminal" meaning a computer peripheral. The same thing happens with "the *present* emperor", "the *local* food".
- "a *Japanese* attitude": an adjective in Penn because it is gradable (e.g. "a very Japanese attitude"), but a noun in ATR because it's a country word.

However, many cases such as "special", "local" and "right", that according to the ATR rules ought to be nouns, are actually (correctly in my view) tagged as adjectives in LDB and BTEC.

I propose that the Penn scheme should be adopted, so that the above cases become ADJ when they're used adjectivally. I also propose that that the capitalized ones should be PROPN rather than CN when used nominally – see section 3.4.4. Some examples:

He is Japanese/ADJ.
He is a Japanese/PROPN.
He speaks Japanese/PROPN.
This is a Japanese/ADJ book.

3.8.2 "-ing" verb forms and nouns

It can be very difficult to distinguish a gerund (ATR V-INGG, Penn VBG) from a present participle (ATR V-INGP, Penn VBP again) and/or from an adjective or noun.

Santorini (p14, "JJ or VBG" and p19, "NN or VBG") gives guidelines for these cases, although the Penn scheme does not distinguish gerunds from present participles. The ATR Shiyou document also gives guidelines (3.1.1, 3.2.1), but they are not always followed: I looked at a sample of 100 "-ing" word occurrences in BTEC that are tagged V-INGG, V-INGP, ADJ or CN and followed by a CN, and judged 32 of them to be incorrectly tagged.

I will look at nouns and verbal forms first, and then in section 3.8.3 bring adjectives into the picture.

According to QGLS (17.54, p1290) there is a continuum from deverbal nouns (clear CN's like "painting" as a physical object) via verbal nouns (gerunds, roughly V-INGG) to participles (V-INGP). QGLS give 14 examples involving the word "painting" to illustrate a sequence of points along this continuum. These examples are reproduced in Table 3.1, together with QGLS's descriptions of them and the corresponding ATR and Penn tags. QGLS's description of each occurrence of "painting" is given in the second column of the table; "Ger/Part" are cases that QGLS call participles, but they seem to me to be gerunds because they function as the heads of noun phrases, and QGLS's argument implies that they would be gerunds if "Brown" and "deftly" were absent.

QGLS (p1292, note A) in fact dislike the term "gerund" because they say it is hard make a coherent distinction between gerunds and (other) participles – and indeed the Penn scheme does not try to do so. However, it is worth trying to make this distinction where possible, because the Japanese translation will often depend on it (a gerund will generally translate to a nominal form, and other participles to a verbal form).

It can be seen that the Penn NN tag covers QGLS's "pure count noun" and "verbal noun" cases. Cases 5 and 6, which QGLS and Penn agree are not nouns, are CN under the ATR scheme because they are modified by the determiner "Brown's".[8]

[8]I am classing "Brown's" as a determiner because it can be replaced by a possessive

1	Pure count noun	CN	NN	some paintings of Brown's.
2	Pure count noun	CN	NN	Brown's paintings of his daughter.
3	Verbal noun	CN	NN	The painting of Brown is as skilful as that of Gainsborough.
4	Verbal noun	CN	NN	Brown's deft painting of his daughter is a delight to watch.
5	Gerund	CN	VBG	Brown's deftly painting his daughter is a delight to watch.
6	Gerund	CN	VBG	I dislike Brown's painting his daughter
7	Ger/Part	V-INGG	VBG	I dislike Brown painting his daughter
8	Participle	V-INGP	VBG	I watched Brown painting his daughter
9	Ger/Part	V-INGG	VBG	Brown deftly painting his daughter is a delight to watch.
10	Participle	V-INGP	VBG	Painting his daughter, Brown noticed that his hand was shaking.
11	Participle	V-INGP	VBG	Brown painting his daughter that day, I decided to go for a walk.
12	Participle	V-INGP	VBG	The man painting the girl is Brown.
13	Participle	V-INGP	VBG	The silently painting man is Brown.
14	Participle	V-INGP	VBG	Brown is painting his daughter.

Table 3.1: QGLS, ATR and Penn analyses of "-ing" forms

Despite the difficulties, I think it is desirable for the same distinctions to be made distinctions should be the same in the ATR and Penn tag sets, i.e. ADJ or CN should never go with VBG or VBN, and V-INGG/INGP should never go with JJ or NN.

The Penn guidelines seem to me to be consistent with the grammar of English as generally understood. However, there are problems with the ATR guidelines which can be seen from the following examples.

The rules to decide when "-ing" word is CN (rather than V-INGG or V-INGP) are given in section 3.1.1 (6) of Shiyou. Using Mine-san's translation in italics:

When "V -ing" is CN*:*

1. *it is plural.* I agree.

2. *it is modified by determiners, adjectives* (yes, I agree), *adverbs etc, or followed by nouns.* This second part is wrong. An "-ing" word modified by an adverb is a gerund, not a noun. Adverbs never directly modify nouns. See also Santorini (p19, "NN or VBG", bullet point 2).

3. *it is an antecedent* (the example suggests that this means "it is the head noun of a relative clause"). This doesn't seem right. It would mean that in "Brown painting his daughter, which happened yesterday, was a delight to watch", we would have painting/CN, but in the absence of the relative clause, we would have V-INGG. The problem is that gerunds as well as nouns can be modified by relative clauses.

4. *it is "non"+"V -ing" (one word "non...").* "Non-" words are difficult; they are certainly not gerunds, but Penn (section 3.8.1 above) treats them as adjectives when they are hyphenated. I think CN is as good a solution as any here.

5. *when 1-4 don't apply, it must have an entry in READERS and also, it has to be concrete.* This may be problematic in some cases because the READERS entry may not be for the same sense of the word. But the stipulation that it has to be concrete probably corrects the majority of these cases.

pronoun like "his", which is a DET in the ATR scheme, and presumably the status of "painting" here is not affected by pronominalizing "his". However the Japanese term translated "determiner" by Mine-san is "kanshi", which can also be translated as "article", and one cannot replace "Brown's" by an article like "the" in 5 or 6. So perhaps 6 would be tagged V-INGG. However, 5 would still be CN because it is modified by an adverb, "deftly".

If these conditions do not apply, the tag is either V-INGG or V-INGP. The Penn scheme does not help in this particular decision because its VBG tag covers both ATR tags. In Shiyou (3.2.1 (3)), a list of conditions is given for each tag; if none clearly apply, the default tag is V-INGP. Most of the individual conditions seem sensible; however, there are some sentences in which conditions from each list apply, and it is not clear to me what tag should be assigned in that case. For example, the first condition for V-INGG is "it is followed by objects and complements, and it has to form a phrase", and the eighth condition for V-INGP is "it follows go/come/spend/waste/be busy etc". Both are satisfied by

Brown was busy painting his daughter.

The ATR rules seem most consistent with QGLS and my own intuitions if we assume that the V-INGG conditions take priority. But then, logically the V-INGP conditions are redundant.

3.8.3 "-ing" form verbs and adjectives

Now let us consider when "-ing" words should be tagged as adjectives. Santorini (p14, "JJ or VBG") gives the Penn conditions for an -ing word to be an adjective JJ:

1. if it is gradable, that is it can take "very" or "more ... than".
2. if there is corresponding "un-" form with the opposite meaning, e.g. interesting/uninteresting.
3. if it occurs in construction with "be", and the "be" could be replaced by any of *become, feel, look, remain, seem* or *sound.* "The conversation was (became) depressing".
4. "if it precedes a noun, and the corresponding verb is intransitive, or does not have the same meaning". I don't see the logic for the "intransitive" condition, and Santorini's examples are not consistent with it. She has "an appealing/JJ face" because she judges "a face that appeals" as wrong; I think it's OK, and anyway one can have "a face that appeals to people". In contrast, she has "the existing/VBG safeguards" because "safeguards that exist" is acceptable; but "exist" is an intransitive verb, so this should be JJ by her condition. Interestingly, the Penn Treebank-3 corpora have 133 examples of existing/VBG and 26 of existing/JJ; virtually all examples are prenominal so they should have the same tag.

5. if there is no corresponding verb: "a thoroughgoing/JJ investigation". This seems fine to me.

The ATR conditions (Shiyou 3.2.1(3)(b), p24, my translation) are these:[9]

1. If it has an object, etc, it is a V. Example: "I am missing my way".

2. If it is premodified by anything other than a DET, DETADJ, $S or NUM, it is a V-INGG or V-ING. Examples: "English/CN speaking/V-INGP guide", "other/ADJ sporting/INGG events".

3. If it's in READERS as ADJ, treat it is ADJ. Examples given are "interesting", "missing", "following" and "becoming". This is wrong, because it takes no account of the READERS ADJ entry being for a different sense of the word. For example, BTEC contains "it's becoming/ADJ cloudy", but this is not the adjective "becoming", which means "attractive" or "smart". The confusion is evident throughout BTEC; there are 32 examples of becoming/ADJ and 25 of becoming/V-ING(G,P), but only nine of the becoming/ADJ cases are really the adjectival meaning; seven of the others even have an NP object, so they should be caught by the first condition in this list anyway.

4. In ambiguous cases of "-ing" words following "be", if a decision cannot be made even by looking at context, use V-INGP.

Thus neither set of conditions seems to work exactly as given. However, the Penn scheme can be made to work if we reword its fourth condition to say "if it precedes a noun, and there is no corresponding verb with the same meaning". Alternatively, the ATR scheme could be modified by changing the third condition in the list above to say "If it's in READERS as ADJ *with the same sense*, treat it as ADJ". The two sets of conditions would then, I think, be consistent, so they could be merged for clarity.

3.8.4 A statistical analysis

I trained the fnTBL tagger (Florian and Ngai, 2002) on the whole of the Penn Treebank-3 and applied it to the BTEC corpus (with all compounds replaced by their component words). I then analysed the 37832 the cases where a word ended in "-ing", was tagged ADJ, CN, V-INGG or V-INGP in the ATR tagging, and was tagged JJ, NN or VBG by fnTBL. The percentage

[9]The third condition in this list is actually given first in Shiyou, but the first two start with "Regardless of whether it has an ADJ entry in READERS", so they are logically prior to it.

	ADJ	CN	V-INGG	V-INGP	Total
JJ	2.49%	**0.26%**	**0.29%**	**0.83%**	3.87%
NN	**2.61%**	18.58%	**2.91%**	**2.73%**	26.82%
VBG	**6.18%**	**6.76%**	13.86%	42.51%	69.31%
Total	11.28%	25.59%	17.06%	46.07%	100.0%

Table 3.2: Classifications of "-ing" words by ATR and fnTBL

of each combination is given in Table 3.2. Invalid combinations are shown in bold.

Overall, about 77.4% of combinations shown in the table were valid. The large majority of these were probably correct, because for them to be incorrect, both the ATR annotator and the fnTBL tagger, which are completely independent, would have had to make the same mistake.

I then examined 25 randomly-chosen examples of each of the invalid combinations, discounting ungrammatical sentences and cases where I judged neither tag correct. The Penn (fnTBL-assigned) tag was correct about 50% of the time for clashes between NN and either V-INGG or V-INGP. However, on the other cases, Penn was reliably better: from 68% of the JJ/V-INGG and VBG/ADJ cases, up to 92% on the NN/ADJ cases. Overall, Penn is probably better in around 75% of the cases of disagreement. This suggests that looking at such disagreements would be a very efficient way of identifying annotation errors: at least for these tag combinations, three out of four problems identified would actually be an annotation error.

Taking these figures together would suggest an error rate (on this particular set of words, which is certainly more difficult than average) of about 20 to 23% for the ATR annotation, and about 8 to 10% for the Penn-scheme annotation. Because the Penn-derived error rate is lower, a majority of the clashes probably do represent places where a correction to the ATR tag is needed, providing further evidence of how comparisons of the type suggested here could improve matters.

3.8.5 "-ed/-en" form verbs and adjectives

Distinguishing adjectives and past participles (ATR V-PP, Penn VBN) can also be hard; see Santorini p15, "JJ or VBN" and QGLS 7.15-7.19, pp413-416.

Santorini's criteria for this case are more or less consistent with QGLS. They have a number of parallels to those for distinguishing adjectives from present participles. She gives nine conditions altogether, some of them quite

complex.

The ATR rules, in contrast, are just these (Shiyou 3.6(3), p30, my translation):

- If the ADJ and V-PP meanings differ, tag according to which meaning is used.
- If the V-PP usage is old or otherwise not used, tag as ADJ. Examples: renowned, handwritten.
- Words prefixed with "un", "non" or "middle" are also ADJ.
- Otherwise, tag as V-PP.

These criteria are rather simplistic. For example, the word "surprised" is always tagged as V-PP in BTEC (even when it appears as a past tense form, as in "you surprised me"!), and that is indeed what the ATR rules would dictate. However, in the Penn treebanks it is just as often tagged as an adjective, and this fits with Santorini's (p15, "JJ or VBN", point one) that a gradable "-ed/-en" word should be tagged as an adjective, as in "I'm really surprised/JJ". In the BTEC corpus, there are seven cases of "really surprised" and one of "greatly surprised", all tagged as an ADV modifying a V-PP. This is in contrast to the word "surprising", which is always tagged as ADJ, never as V-INGP.

It is not clear to me what should be done here. Both Santorini and QGLS acknowledge that this distinction is hard to make, and the Penn criteria are complex. Effectively, ATR draws the line so as to allow fewer "-ed/-en" adjectives than Penn does, and there is nothing obviously wrong with this. However, it seems strange to do this when nothing similar is done in the "-ing" case. It is interesting to look at the cases where Penn and ATR differ on the adjective-verb decision for both "-ing" words and "-ed/-en" words. When the BTEC corpus is tagged using the Penn scheme (trained on all the Treebank-3 data), 86% of the relevant disagreements on "-ing" words have ATR favouring the adjectival reading, but this is true for less than 2% of the "-ed/-en" cases.

3.9 Summary of proposals

This section summarizes the proposals made so far in this chapter.

- Country names, colour words and other adjectives should be tagged JJ, not CN or PROPN. Section 3.8.1.

- The distinction between interjections and adverbs should be made as suggested in section 3.3.1.
- "Not" should be tagged RB, not NOT. Preference only. Section 3.3.2.
- "a.m." and "p.m." should be tagged PNOM. Section 3.3.3.
- The distinction between ADV and PREADV should be redrawn somewhat. Section 3.3.4.
- Merge PREPADV into ADV. Preference only. Section 3.3.5.
- The CONJADV tag should either be abolished, with words being tagged either CONJ or ADV, or used for the correct set of words. Section 3.3.7.
- The LOCADV tag should be merged into ADV. Preference only. Section 3.3.8.
- The tagging of cardinal numbers should be reorganized, with numbers being tagged NUMDET, PRON, CN or NUM as appropriate. Section 3.4.2.
- The NUM-ORDINAL tag should be abolished, with the relevant words being tagged as nouns, adjectives or adverbs as appropriate. Section 3.4.3.
- Plural proper names like "Beatles" should be consistently tagged PROPN-PL, not PROPN.
- The LETTER tag should be used only for letters, not variables. Variables should either be removed from the corpus or given a new SYM tag. Section 3.4.5.
- Care should be taken that prepositions governing clauses can accurately be tagged CONJ rather than PREP, in BTEC and any other corpora that may need to be processed. Section 3.5.
- "To" should be tagged TO (or V, to avoid introducing a new tag) rather than PREP when it is an infinitive marker. Section 3.5.
- The DETADJ should be used only for predeterminers, and perhaps renamed PREDET, with its other current members being assigned other (existing) tags. Section 3.6.1.
- The NO tag should be abolished, with "no" (as a determiner rather than an interjection) being retagged as DT. Preference only. Section 3.6.1.

- The -PL suffix should not be used in the PRON tag. Section 3.6.2 .
- Taggings of "wherever" and "whenever" should be changed from WHADV to WHCONJ. Section 3.6.3.
- If possible, a new WHREL tag should be introduced for relative pronouns. Preference only. Section 3.6.4.
- Possessive forms should be kept as single lexemes, abolishing the $S tag and introducing CN-POSS, CN-PL-POSS PROPN-POSS and CN-PL-POSS to handle the new lexemes. Section 3.6.5.
- Infinitive and non-third-person-singular present verbs should be separately tagged. Section 3.7.1.
- The VTO tag should be abolished, and the words assigned it rewritten in their full forms. Section 3.7.1.
- Adjectives (ADJ) and nouns (CN) should be distinguished according to the Penn criteria, not the existing ATR ones. Section 3.8.1.
- Nouns (CN) and verbal "-ing" forms (V-INGG, V-INGP) should be distinguished according to the Penn criteria, not the existing ATR ones. The existing ATR distinction between V-INGG and V-INGP should be kept, however. It should be clarified which order the rules are to be applied in. Section 3.8.2.
- The Penn and ATR criteria for distinguish adjectives and verb "-ing" forms should be corrected and merged as detailed in section 3.8.3.
- Further thought is needed on distinguishing past participles from adjectives for "-ed/-en" words. Section 3.8.5.

3.10 Final Recommendations

Tagging a sentence can be seen as a limited first step towards a full syntactic and/or semantic analysis. It is easy to find cases where the correct tag can only be reliably assigned – whether by a human or a machine – on the basis of such an analysis. Such cases are in the minority, but they are definitely not trivial in number. This can be seen by looking at the many cases in BTEC where the human annotator has clearly misunderstood the sentence completely and assigned the wrong structure, or no structure, to it. A typical example is

what/WHPRON does/AUXV-3S it/PRON mean/V when/WHCONJ I/PRON initial/ADJ this/PRON ./. .

It is difficult to imagine what structure or meaning the sentence might have if "initial" is tagged ADJ.

The Penn scheme assumes that the human taggers will be native, or at least very fluent, speakers of the language. In particular, it assumes they will be able to evaluate the acceptability of variants on the sentences they look at, and where variants are acceptable, whether their meaning changes in a certain way. In contrast, the ATR scheme tries to avoid relying on native-speaker intuitions, instead formulating its rules in terms of what is (rather than can be) in the sentence, and on what entries are found in the READERS dictionary.

Unfortunately, there is no way to tag a corpus accurately without having reliable full understanding available at some point in the process, and making use of acceptability tests on variants. This does not mean that only native speakers of a language can assign tags. However, it does mean that a native speaker should be available to check every sentence if required. Non-natives should include uncertainties in their work (for example, by assigning several possible tags to a word, or marking a tag with a question mark), and a native speaker should check every such sentence, as well as a few that are not marked as uncertain, to ensure quality control.

As experience is gained in working with a tag set, it can be used to improve accuracy on new material. The rules can be updated and clarified (though always with reference to a native speaker, who should do their best to think of counterexamples for any proposed changes). Also, reliably-tagged sentences can be used to provide suitably analogies for tagging decisions. Ideally, a record should be kept of such uses so that any mistaken inferences can be spotted and corrected.

The more similar a tagging scheme is to the Penn scheme, the more feasible it becomes to use Penn Treebank sentences (whose reliability is good, though not perfect) as references. It will probably make sense to build up a local, corrected version of the Penn Treebank, because working on difficult decisions often leads one to Penn sentences where a similar problem has been tackled and given the wrong solution.

There is plenty of scope for designing tools for large-scale detection and correction of tagging errors. For example, the BTEC corpus could be tagged under the Penn scheme by a tagger trained on Penn Treebank data, and inconsistent tags flagged up, as in the small experiment described in Section 3.8.4. Let us now look more closely at that task.

Chapter 4

Automatic conversion between tagging schemes

4.1 Introduction

To recap, the current situation at ATR regarding tagged English corpora is as follows.

- We have the LDB and BTEC corpora tagged with the current ATR scheme. BTEC, only part of which has been manually checked, contains quite a lot of tagging errors.
- We have various LDC corpora tagged with the Penn scheme. There are a few errors, but not too many.
- I have proposed a revised ATR scheme which is more or less a hybrid between the current ATR scheme and the Penn scheme.

Correcting tagging errors and converting between schemes both involve a lot of work if done manually. It would be nice if we could automate the work as far as possible. Specifically, we would like to preserve the information in the manually-checked part of BTEC when we convert automatically from the current ATR scheme to the revised one. It would also be useful somehow to use the information in the LDC tagged corpora to detect and correct errors in the BTEC tagging (both manual and automatic parts). In this document, I will set out some ways to do this. In Section 4.2, I will propose a rule formalism (for which a simple interpreter could be written) for specifying how to convert between tag sets. This could be used directly for converting from current-ATR to revised-ATR, for example. Then in Section 4.3.1, I will look at various options for training a tagger on LDC data and then running

it on BTEC data. This would give us an alternative set of tags for BTEC. Discrepancies indicate likely errors which can be corrected manually in much less time than it would take to check the whole of BTEC. It may also turn out to be possible to semi-automate the correction process by specifying common error patterns and their corrections.

As well as using Penn/LDC data to correct BTEC, it would be worth feeding back corrected taggings on parts of the BTEC into the rest of it. That is, as BTEC taggings are corrected (manually, but speeded up by the use of the procedure detailed in this chapter), the correct part of the corpus can be used to train a tagger for the uncorrected part. The results of this can be put alongside the Penn/LDC-derived tags as one more source of error detection.

4.2 A formalism for tagging scheme conversion

Tagging scheme conversion is best viewed as an example of finite state transduction (FST). The input is the word sequence and their initial tags, and the output is a revised tag for each word.

Any existing FST scheme could therefore be used. One possibility is the rule formalism used by the fnTBL tagger (Florian and Ngai, 2002) which I will be recommending for another purpose in Section 4.3.1 below. However, since run-time efficiency is not crucial, I would prefer a variant of the two-level morphological rule formalism (Koskenniemi, 1983). In this formalism, each rule specifies the output value at a given target position, based on the input values at the target position and, if desired, the inputs and outputs at neighbouring positions. For example, in the revised ATR scheme (Section 3.5) it is proposed that when "to" is an infinitive marker, it should be tagged TO rather than PREP. If the word "to" is immediately followed by a verb base form, we can be pretty sure it is an infinite marker. Thus we might have a rule looking something like this:

!to/PREP */V *rightarrow* TO

The asterisk is a place holder which will match any word. The exclamation mark before the "to/PREP" identifies it as the target for the rule, to make sure we end up with "to/TO ... /V" rather than "to/PREP .../TO".

It may be useful (and easy to implement, if Perl is used) to allow regular expressions on the left hand side of the rules, both for words and for tags. In the case of words, we might want to match certain initial casings or word

endings. For tags, it will save some duplication if we can specify alternatives, e.g. if we want our infinitive rule to take account of split infinitives ("to boldly go...") we could say

!to/PREP */V|ADV *rightarrow* TO

which is equivalent to the more long-winded

!to/PREP */V *rightarrow* TO
!to/PREP */ADV *rightarrow* TO

In the above examples, only one immediate neighbour is specified as a constraint; but we should allow constraints to be specified on both sides, and not just single items but any number. Thus we could have:

*/V !to/PREP */ADV */V *rightarrow* TO

The "two-level" in two-level morphology means that there only the input and output levels, and no intermediate representations: the output of one rule in the list does not become the input of later rules. Although this can initially seem a bit awkward, it makes for much easier specification and debugging of a large rule set. Thus if we have (in this order)

!to/PREP */V *rightarrow* TO
to/TO !*/V *rightarrow* V-INF

the second rule will not fire on a sequence like "to/PREP go/V", because the left-context of the input item "go/V" is still the input item "to/PREP", whether or not the first rule has set the output tag to TO.

It might be useful to allow rules to specify output contextual values on their left hand side, perhaps after an additional slash character, e.g.

to/PREP/TO !*/V *rightarrow* V-INF

("change V to V-INF whenever the word to the left is 'to', with input tag PREP and output tag TO"). However, when writing rules that are only intended to run in one direction, it is in practice seldom necessary to do this. In any case, it is best not to allow output tags to be specified on both sides of the target, because this prevents us from writing a simple unidirectional rule interpreter: the applicability of rules starts to depend on each other, with two rules effectively saying to each other "I will if you will...". If we disallow constraints on right-context output values, we can write an interpreter which moves from left to right through a sentence, finding the output for each target in turn and then moving on to the next word.

We need to specify what should happen if two rules both apply to the same target but designate different output values. One possibility would be to say that rule ordering is significant, with rules higher in the list taking priority. However, this creates a temptation to resolve rule conflicts by reordering, which in practice leads to trouble by disturbing other rule relationships (if we swap rules A and B over to give B priority over A, we may unintentionally prevent C from applying because B is now above it when it was previously below it). Two-level morphology generally makes use of a "default" rule, which just copies the input to the output if no other rule applies. However, we are working with tagging schemes that may use completely different symbol sets: if we are mapping from Penn to ATR, the default output for a `NN` (noun) tag when no special circumstances apply is `CN`, not `NN`.

I think the best approach is therefore to have two types of rule: one (henceforth "type one") is the type we have been discussing so far, in which words and contextual values can be specified as constraints, and the other ("type two") is a much simpler kind of rule that just specifies the default value for each tag if no non-default rule applies, e.g.

`NN` → `CN`
`JJ` → `ADJ`

The left hand sides of these rules would be a simple tag symbol, with no words, contextual values or regular expressions. We would expect to have at most one such rule for each tag in the input scheme. The rule interpreter would then move from left to right through the sentence as before. At each point, it would first look for all the type one rules that can apply. If more than one can apply, and the output values are different, an error is flagged. If one applies, or if several apply but they all give the same output, that value is used. If none apply, the type two rule set is tried. If there is exactly one applicable rule, it is used; otherwise, we have an error.

I think it is best not to have a "default-default" rule which copies the input to the output if no rules of either type one or type two apply. Even if this would often be appropriate, as when converting from the current ATR scheme to the revised one, it is better to require the rules to be specified explicitly, in order to catch any unaccounted-for cases. For example, we might want to abolish some tag altogether in the conversion, and we might write type one rules that are intended to account for all of its occurrences. If we intentionally do not provide a type two rule for that tag, and if the rule interpreter runs successfully on some data, then we know that all cases of the tag in question have been explicitly handled by the type one rules.

4.3 Error Detection and Correction

Let's assume that the scheme described above has been applied to the BTEC corpus, so that we have a version tagged with the revised ATR tagset. This version will still contain errors from the original one, because the conversion rules only take account of revisions to the scheme itself; they do not attempt to correct individual mistaggings.

If we train an automatic tagger on the LDC tagged corpora, we can run it on the BTEC data and get an alternative, automatic tagging. We will find some words that have inconsistent tags: for example, a noun from the manual annotation but an adjective from the automatic one. When this happens, then as long as there is no genuine ambiguity (these are fairly rare), either the manual tag or the automatic one – or both – must be wrong. A human annotator can then look at the problem and make the required correction. Because (we hope) only a small proportion of the words will have inconsistent tags, this procedure will require much less effort than inspecting the whole corpus.

The procedure will not detect every error in the corpus, because it is possible for two tags to be consistent with each other but wrong. This is especially likely to occur with ambiguities that can only be resolved by doing a full parse of the sentence or by understanding what is meant by a phrase. However, it is reasonable to hope that a good proportion of errors will be flagged by inconsistencies and so can be flagged up for correction.

4.3.1 Converting between tagging schemes

To carry out this procedure, we need to overcome the problem that the LDC corpora use a different tag set from BTEC. There are several things we can do:

1. We can easily train the tagger on the existing Penn tags in the LDC corpora. It will then assign Penn tags to the BTEC data. We can then compare tags: for example, Penn `NN` is consistent with ATR `CN` but not with ATR `ADJ`.

2. We can write conversion rules, in the formalism already proposed, to map from Penn to (revised) ATR, apply them to the LDC corpora, train a tagger on the result, and run it on BTEC. We will then have ATR tags for BTEC, and can compare them easily with the original BTEC tags: consistency is just equality.

3. We train the tagger on Penn tags, and run it on BTEC, as in 1. We also write conversion rules as in 2, and apply them to the output of the tagger. The end result is again an ATR-tagged BTEC, which we can compare by equality, as in 2.

Each scheme has its pros and cons:

- Scheme one is easy to carry out, but it will miss some errors because the relationship between the tag sets is not one-to-one. For example, Penn uses the same `DT` tag for "this", whether it is acting as a determiner ("show me this book") or a pronoun ("show me this"), while ATR uses `DET` and `PRON` respectively. Therefore, we will not be able to spot cases in BTEC where a this/`DET` is mistagged as a this/`PRON`, or vice versa.

- Scheme two implies some work needs to be done in writing the Penn-to-ATR conversion rules. Also, these rules may introduce some errors when they are applied. However, the last stage, of comparing manually and automatically derived ATR tags on BTEC, is much more straightforward.

- Scheme three has essentially the same advantages and disadvantages as scheme two, because it consists of the same two operations swapped over. However, either or both of the intermediate products – an ATR-tagged LDC corpus for scheme two, and a Penn-tagged BTEC corpus for scheme three – may be of some value in their own right, and either could be hand-corrected if necessary.

In practice, it will be probably be best to carry out all three schemes. The main effort involved will be that of writing Penn-to-ATR conversion rules for schemes two and/or three; once this has been done, everything else is just a matter of running the relevant software in different configurations. Running all three schemes will give three sets of results, which will be very useful because the different patterns of agreement and disagreement between them can be used to estimate the likelihood of an error and to detect problems with the conversion rules.

As mentioned earlier, once the correction process has got under way, these three sets of results can be put alongside another set, arising from a tagger trained on the corrected part of the BTEC corpus. As correction progresses, this stream of results should become more and more reliable. Where BTEC tags have not already been manually corrected, they should be viewed as just one more automatically-derived set of results, without any special status.

It is worth giving serious consideration to the use of fnTBL as the tagger for doing all this. fnTBL is a transformation-based tagger, using the same

principles as Brill's (REF) system but with a much more efficient implementation. It has two main advantages over most other taggers:

- The output of the training phase is not a large body of statistics, but a relatively small number (typically a few hundred) of rules that specify when tags should be changed. Initially in a tagging task, each word is tagged with a default value, usually its most frequent tag. Then the rules use context to make corrections. The advantage of having a small number of explicit rules is that they can be understood and altered easily. In practice, this leads to faster implementation and debugging of tagging tasks.

- fnTBL can be used for a wide range of label assignment tasks. It does not only map words to tags; it is possible to configure it to accept any number of streams of input and produce any number of output streams. So, for example, it could be made to accept as input words and tags assigned from two or more of the taggers suggested above, and to use the information as best it could to output tag values that took advantage of all the inputs.

In connection with the last point, fnTBL can even be used for non-part-of-speech tagging tasks such as noun-phrase bracketing: there is an example configuration in the fnTBL distribution which assigns NP start, continuation and end "tags" based on words and part-of-speech tags. A similar configuration could, perhaps, decide on whether words should be compounded before tagging, thereby avoiding the issues raised in Section 2.3.1, but this might not be very reliable as POS tags would not be available to help in the process.

4.3.2 Compound words

Another factor that needs to be taken into account in this process is word compounding (Chapter 2). We need as far as possible to represent the same words and phrases in the same way between corpora: if we have only "air_mail" in BTEC but only "air mail" and "airmail" in LDC, errors may be introduced when the LDC-trained tagger is run on BTEC. Furthermore, we not only need to convert the words, we also need at some point to assign tags to the new words, using rules like those proposed in Section 2.4.1.

Various decisions need to be made. Should we make LDC conform to BTEC, or BTEC to LDC, or should we alter both to some more neutral scheme? And how should the compound-adjusting step be combined with the tagset-conversion and automatic-tagging steps?

Whatever scheme is adopted, I think it would be best for the end result, on which the comparison between tags is made in order to spot errors, to use the same word forms as BTEC. This will make it easier to correct tags in individual cases and know that the correct versions will be stored explicitly.

Note that word-adjustment and the selection of sensible new tags for words do not have to take place at the same time or even in sequence. We could simply join tags together when we make a compound, so "air/NN mail/NN" might initially be converted to "air_mail/NN_NN" and only later (perhaps after tagset conversion and/or automatic tagging) to "air_mail/NN" or "air_mail/CN". However, this would introduce quite a lot of temporary new tags, which might worsen data sparseness problems and worsen the performance of the automatic tagger. I therefore think it is best to carry out the two steps simultaneously or immediately in sequence.

The LDC corpora as distributed do not contain any ATR-style compounds such as air_mail. However, since they come from different sources, there are many differences within them in the representation of the same word or phrase. For example, "air mail" and "airmail" both occur, and, again to reduce data sparseness, it would be good to normalize to one form.

If we do compounding on the LDC data before tagset conversion, then the compounding rules will need to refer to Penn tags. On the other hand, if it is done after compounding, they will specify ATR tags, and we expect anyway to be developing a set of rules (Section 2.4.1) to adjust compounds using the ATR tag set. Note, however, that Scheme One from Section 4.3.1 above does not involve any tag conversion at all, but rather the direct comparison of Penn and ATR tags to spot errors. Therefore if we are to adopt Scheme One (on its own or in combination with the other schemes) we are going to need some Penn-tagset compounding rules.

In practice, I suspect that it will be quite easy to convert automatically from ATR-tagset compounding rules to Penn ones. This is because the areas where the relationship between the tagsets is complex tend not to occur very often in compounds, which are usually just noun-noun, adjective-noun and verb-particle combinations. So I recommend that a set of compounding rules be developed with both BTEC consistency and Penn conversion in mind, and converted automatically for the latter purpose. Compound conversion would then be the first step in the error-detection process. In other words, each of schemes one to three in Section 4.3.1 would be carried out not on the LDC corpora in their original forms, but on versions which had first undergone compound conversion. We would have to take some care with compound phrases like "a_lot", which (currently) have ATR tags for which there is no real corresponding Penn equivalent; the simplest solution is probably just to use the appropriate ATR tag.

References

Anonymous, February 2001. *Eigo Keitaigo Shiyousho* ("Shiyou"). ATR Spoken Language Translation Laboratory.

Florian, R., and Ngai, G. *Fast Transformation-Based Learning Toolkit.* At `http://nlp.cs.jhu.edu/ rflorian/fntbl/index.html`, checked 17th June 2003.

Halliday, M., and Hasan, R.(1976). Cohesion in English. Longman.

Koskenniemi, K. *Two-Level morphology: A general Computational Model for Word-Form Recognition and Production.* Ph.D. thesis, University of Helsinki.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J., 1995. *A comprehensive grammar of the English Language.* Longman. In class 214 of the ATR library.

Santorini, Beatrice, 1995. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.* 3rd revision, 2nd printing. Available at `ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz` (checked June 4th, 2003; if missing, see the Penn Treebank home page at `http://www.cis.upenn.edu/~treebank/home.html`).