# LANGUAGE VARIATION, PARSING, AND THE MODELLING OF USERS' LANGUAGE VARIETIES

James Kilbury
Technische Universität Berlin
IAI, CIS, Sekr. FR 5-8
Franklinstr. 28/29
D-1000 Berlin 10

Most natural language computer systems either do not deal with the problem of language variation at all or else simply treat it as a matter of robustness like ill-formed input. The features of input that mark it as belonging to a particular language variety are regarded as noise. Since this information is lost during analysis, it cannot be used to generate system answers tailored in form and content to particular users, or more importantly, to help in the analysis of further input.

There are cases, however, where information about a user's variety is essential for the correct analysis of his input. For example, if the user of a hypothetical airport information system asks

$$\left\{\begin{array}{l}\textit{Do you have}\\ \textit{Have you got}\\ \textit{Have you}\end{array}\right\}\ \textit{a Mrs. Warrick in your passenger list?}\ ,$$

the use of *Do you have ... ?* points to an American speaker, while the question built with *Have you ... ?* suggests a conservative British speaker. Here the linguistic data have been somewhat simplified for expository purposes. Of course, all three forms of the question mean the same. If the user later writes

*It was important that she took the flight at 9 am.* ,

the system can infer that an American means that she actually did take the flight (otherwise he would have used the base verb form *take*). This inference may not be drawn with the conservative British speaker, however, because his variety of English does not allow the use of the form *take* here. So a correct interpretation of the statement depends on information about the user's language variety gained from previous input.

A major insight gained in recent linguistics is that "a great deal of linguistic variation patterns in an implicational manner" (Bailey 1973: 28). One feature of a language variety may imply another, which itself may in turn imply a third, while neither of the latter two features implies the first. Such an implicational scale is found by Johannesson (1983) to apply to the following five features of a form of nonstandard English:

(A) use of *ain't*
*Maybe he sees things that ain't there.*

(B) prefixed progressive
*They're always a-hammering.*

(C) double negation
*I hadn't got nothing on.*

(D) restrictive relative *as*
*It's a pity that folk as talk about fighting the Enemy can't let others do their bit.*

(E) nonstandard subject-verb agreement
*It don't look like a cloud.*

For these features Johannesson establishes this implicational scale (p. 147): $A \supset B \supset C \supset D \supset E$ .

To deal with variation a natural language system could use such data and model linguistic varieties as points along implicational scales. The attestation of feature E in input would allow no inferences to be drawn, while B would imply that features C, D, and E could be expected in further input. Thus, the system could adjust to the variety of the particular user.

This example requires two further comments. First, the consideration of nonstandard language is important for the development of practical systems intended for wide use by normal people. Second, a linguistic feature not only may imply other linguistic features but also may contribute to a picture of the user's education (and thus of the information available to him) as well as his attitudes and beliefs. Consequently, features of a user's language variety may serve to construct a model of the user himself as well as his language variety.

Although the examples of variation provided so far have involved syntax, variation is found at all linguistic levels. Cross-level implications are frequent, so that lexical or morphological features may be inferred from phonological features or vice versa. Variation is least evident in syntax, more so in morphology, and most significant in the lexicon and phonology.

Variation must be viewed as a central problem of speech recognition if researchers seriously hope to develop systems that can cope with a wide range of users in everyday situations who have not been specially trained for the systems. The difficulties of speech recognition are so great even without the factor of variation, however, that the latter question has as yet hardly been raised. Church (1983) discusses phonological processes (modelled in linguistics with phonological rules) that may lead to the neutralization of oppositions between distinctive sounds and thus to the loss of meaningful distinctions. For example, palatalization of the apical stop in *duel* may result in a pronunciation homophonous with that of *jewel*. Church argues for a maximum utilization of phonotactic and allophonic information in phonological parsing so as to minimize the extent of genuine neutralization, but he regards each problem in isolation without reference to information about the speaker's pronunciation variety that could be gained from earlier input.

Language variation is an extremely complex phenomenon whose dimensions include style and tempo as well as geographic and social factors. Even within a single socio-geographic variety, a speaker's pronunciation will vary according to the former factors. For example, the apical stops / t d / are likely to be deleted between consonants in English so that *missed* and *miss* are indistinguishable (with loss of the tense marker) in the sentence

*Some people miss(ed) the point.*

in normal speech. Deletion is more likely in less formal speech and at higher tempos. Moreover, Bailey (1973: 138) has shown that this deletion also depends on the phonological environment according to an implicational pattern: if a speaker deletes in the most marked environment, represented by *piston* , then it may be inferred that he will also delete in the progressively less marked

environments represented by *laundry, westward, missed the, trustworthy, mostly, handbag* , and *moisten.*

Socio-geographic features may be constant for a given variety and provide information of particular importance for a speech recognition system. The rounded vowel of the abstract (lexical) phonological representation of *pot* is unrounded in most varieties of American English and sounds like the vowel in the British pronunciation of *part.* So sentences like

*I need a new pot/part for my oven.*

are ambiguous until the variety of the speaker has been established. Furthermore, American English merges / t d / intervocalically after stress, so that a genuine ambiguity may arise in sentences like

*I've been writing/riding to Paris every week.*

in American English, while British English preserves the distinction (again, the data have been simplified). Once a system has identified a pronunciation as the realization of *pot* or *part* it can construct a partial model of the speaker's variety; this model can then be used to infer the interpretation of *writing* or *riding.*

In order to deal with language variation in a computer system a formal model for the representation of language varieties is needed. The claim by Schuster (1985: 20) that "grammars can serve as user models" is a step in the right direction but is unsatisfactory for our purposes. Schuster explicitly excludes the consideration of variation (p. 21). We require a model of the user's variety rather than of the user himself, and a grammar cannot serve as the former since it is precisely the variability within a grammar that must be modelled.

Although Generalized Phrase Structure Grammar (GPSG) as developed by Gazdar et al. (1985) is a theory of grammar that takes no account of variation, part of the formalism of GPSG can be directly employed for the formal representation of language varieties. Let every rule and lexical entry in the overall grammar G of a language have an identifier which serves as a *feature name.* An ordered pair $\langle f, v \rangle$ consisting of a feature name f and a Boolean *feature value* v (indicating presence or absence of the feature in a variety) constitutes a *feature specification.* In analogy to a syntactic category of GPSG, a *variety model* $\Pi$ is a consistent set of feature specifications. The model $\Pi$ can be regarded as a partial function

$$\Pi : G \rightarrow \{+, -\}$$

mapping the rules and lexical entries of G into {+, -}. The *special grammar* $G_\Pi$ of the language variety $\Pi$ of L(G) is given by

$$G_\Pi = \{\delta^\Pi \mid \delta \in G \wedge \Pi(\delta) = + \} .$$

Variety models, like GPSG categories, may be more or less fully specified. A model $\Pi'$ is an *extension* of a model $\Pi$ iff every feature specification of $\Pi$ is also in $\Pi'$ and $\Pi'$ is consistent. Thus, the notion of *subvariety* can be directly captured in terms of extension.

GPSG expresses redundancies in categories by means of *feature cooccurrence restrictions* (FCRs) and *feature specification defaults* (FSDs), which are Boolean conditions over feature specifications, typically in the form of material implications.

The obligatory implicational relationships captured in Bailey's theory of variation are therefore represented with FCRs. FSDs, in contrast, stipulate unmarked (default) feature specifications which are normally expected, all other factors being equal, but which need not necessarily be present in a variety model.

The formalism for variety models just presented may be implemented and incorporated in a parser to allow a monotonic incremental approximation of a user's language variety $\Pi'$ during parsing. Let $\Pi_i$ be a model with the certain information (i.e. feature specifications resulting from direct attestation and FCRs) after i successful rule applications and lexical retrievals, whereby $\Pi_0 = \{\}$. $\Pi_i'$ is the extension of $\Pi_i$ obtained by applying all FSDs to $\Pi_i$. $\Pi_0'$ is thus the fully unmarked model of the user's variety postulated by the system before any input has been analyzed, while $\Pi_k'$ is the model after k steps of analysis which constitutes an approximation of $\Pi'$.

The series $\Pi_0, \Pi_1, \ldots, \Pi_k$ is monotonic (i.e. for each $\Pi_i$ with $0 \leq i, i+1 = j,$ and $j \leq k$, $\Pi_j$ is an extension of $\Pi_i$) if the model is restricted to linguistic features that are constant for a given variety. In contrast, the series $\Pi_0', \Pi_1', \ldots, \Pi_k'$ obtained with defaults can only be monotonic if $\Pi_k' = \Pi_0'$. Further refinements will be needed in order to handle stylistic shifting within a given variety.

Provision should also be made for contradictory feature specifications in cases where the user's variety is inconsistent. The system could then revert to the unmarked model $\Pi_0'$ or else, more sensibly, explicitly mark the features for which contradictory specifications have been obtained.

The implementation proposed here has not yet been carried out. A major prerequisite will be the description of a suitable body of language data in the variety-model formalism.

References:


Bailey, Charles-James N. (1973): *Variation and Linguistic Theory.* Arlington, Virginia: Center for Applied Linguistics.

Church, Kenneth W. (1983): "A Finite State Parser for Use in Speech Recognition," 91-97, 21st ACL Proceedings.

Gazdar, Gerald / Klein, Ewan / Pullum, Geoffrey / Sag, Ivan (1985): *Generalized Phrase Structure Grammar.* Oxford: Blackwell.

Johannesson, Nils-Lennart (1983): "On the Use of Syntactic Variation in *The Lord of the Flies* ," 135-151, in *Papers from the Second Scandinavian Symposium on Syntactic Variation* , ed. by Sven Jacobson. Stockholm: Almqvist & Wiksell.

Schuster, Ethel (1985): "Grammars as User Models," 20-23, Proceedings of the 9th IJCAI.