Poplack, Shana. 1993. Variation theory and language contact. In D. Preston (ed.), American dialect research: An anthology celebrating the 100th anniversary of the American Dialect Society, 251-263. Amsterdam: Benjamins.

Variation theory and language contact:[1]

Shana Poplack
University of Ottawa

1.0. Introduction

This paper describes a variationist sociolinguistic approach to the study of language contact phenomena. In what follows we first briefly outline the basic notions informing the variationist framework, describe the key concepts and issues in current language contact research, and then proceed to explore how variationist sociolinguistic concerns may be applied to issues fundamental to the bilingual[2] inquiry. In so doing, we draw on our ongoing work on typologically similar and different language pairs: Spanish/English, French/English, Finnish/English, Tamil/English and Arabic/French in North American contact situations. Our focus is not on the results of these studies, but rather on illustration of 1) the conceptual, methodological and analytical problems arising in the course of these investigations, and 2) some of the solutions we have adopted to overcome them.

2.0. Variation Theory

The branch of empirical linguistics known as *variation theory* (e.g. Labov 1971, 1984; Sankoff 1982, 1988, G. Sankoff 1974, G. Sankoff & Labov 1985, Guy this volume, Wolfram this volume) involves a combination of techniques from linguistics, sociology, anthropology and statistics, among others, to scientifically investigate language use and structure as manifested in natural(istic) context. The variationist viewpoint on language may be characterized by its preoccupation with 1) accounting for grammatical structure in connected discourse, and 2)

explaining the apparent instability therein of linguistic form-function relations (Sankoff 1988:141). In scientifically accounting for the production data contained in a speech sample, variationists seek to discover *patterns* of usage, which pertain to the relative frequency of occurrence or co-occurrence of structures, rather than simply to their existence or grammaticality.

The primary object of description of the variationist is the speech of individuals qua members of a speech community, i.e. informants specifically chosen (through ethnographic or sociological methods) to represent the major axes of community structure. Thus, an important aspect of any study in the variationist framework involves entrée into the speech community, where observation of language use in its socio-cultural setting is carried out. A specific goal of this procedure is to gain access to the *vernacular*, the relatively homogeneous, spontaneous speech reserved for intimate or casual situations. This is taken to reflect the most systematic form of the language acquired by the speaker, prior to any subsequent efforts at (hyper-) correction or style-shifting (themselves imposed by the combined pressures of group membership and the social meaning within that group of the linguistic options available). Since in almost every corpus of production data there are some linguistic elements that do not obey the normal constraints of the system, the analyst must be able to distinguish systematic from unsystematic heterogeneity. Another motivation for analysis of the vernacular is to provide a basis for establishing the nature of the system, against which we can subsequently assess what may be characterized as deviant with regard to it.

The structure of communication in the speech community is seen by variationists as realized through recurrent choices made by speakers at various interactional and grammatical levels (ibid.:151). The choice mechanism entails that given linguistic 'functions' may be realized in different 'forms.' Thus, it is fairly uncontroversial that the Caribbean Spanish plural marker *-s* may be produced as [s], [h] or ∅; the French negative particle as *ne ... pas* or *∅ ... pas*; Vernacular Black English 3rd p. sg. copula as *is*, *-s*, or *∅*, and none of these choices involves differences in referential meaning. In order to account for the variant that was actually selected in a given situation, the variationist

must determine why, where and when it was used, as well as by whom. As becomes apparent from examination of natural discourse collected in any speech community, the answers to these questions are themselves variable. Methods developed for dealing with this variability stem from the recognition that it is *inherent;* i.e. (in contrast to classic cases of 'allophonic' variation, for example) it cannot be factored out, no matter how closely the analyst specifies the context. This does not imply that such variability is unstructured. The variationist adopts quantitative techniques to uncover the systematic differences between speakers, often associated to some extent with one or more of age, sex, ethnicity, educational level, etc. Typically, each speaker will alternate among all the choices, but will manifest an overall *pattern* of variant frequencies consistent with that of other individual members of her group.

In conjunction with extra-linguistic influences, purely internal features of the linguistic environment will also play a role in determining variant choice. The use of multivariate or 'variable rule' analysis (e.g. Sankoff 1979, Rand & Sankoff 1988) enables the analyst to extract regularities and tendencies from the data, and thereby determine how selection of a linguistic structure is influenced by specific configurations of factors that characterize the environment in which it occurs. In this way it is possible to ascertain which features of the (social and linguistic) context favor or disfavor the occurrence of a form when all are considered simultaneously, and how strongly. The use of this methodology has succeeded in overcoming many of the analytical difficulties associated with intuitive judgments and anecdotal reporting used in other paradigms. This is particularly crucial in the study of bilingual and/or minority language situations, where normative pressures inhibit the use of vernacular or non-standard forms, and where 'categorical perception' on the part of the linguist/observer tends to inflate the importance of a form which may have in fact only occurred on a few occasions. In what follows we illustrate how these considerations may be applied to the bilingual context.

3.0. Concepts in Language Contact

Our own program of research on language contact involves the study of the linguistic processes by which forms from two or more languages may be combined as a result of their common use, the linguistic constraints on such combination, and its consequences for the structure of the languages involved. We have also sought to ascertain the social meaning of language choice as exemplified by speaker 1) behavior, 2) attitudes, and 3) perceptions.

We begin by defining our terms. We follow Weinreich (1968:1) in designating the individual as the locus of language contact, with the proviso that that individual be a bona-fide member of a bilingual speech community. Again following Weinreich (ibid.), we define *bilingualism* as the practice of alternately *using* (emphasis ours) two or more languages, and the individuals involved as *bilingual*. The usage requirement ensures that both languages are regularly accessed in normal interaction, and in the stable bilingual communities we have studied, speakers typically make use of both languages with the same interlocutors, in the same domains, and within the same conversational topic. Our focus on *intra-situational* language combination is at least partially motivated by the goal of obtaining data permitting the establishment of linguistic, in addition to other, constraints on its occurrence; situational language switching (as described by Gumperz 1982) may consist entirely of (monolingual) stretches of speech in one language followed by (monolingual) stretches in another, and thus provide no locus to observe the processes of combination which interest us.

Our studies have focused on adult bilinguals whose language repertoire is 'stable' in the sense that neither language acquisition nor attrition is involved in the contact situation, although each of the relevant languages will, of course, continue to manifest internal variability. This focus is not imposed by any theoretical dictate, but simply by the goal of describing the linguistic concomitants of *regular* interaction in two or more languages, to which the more labile behaviors of language learners or losers may ultimately be compared. Our emphasis on stable bilingual communities, as opposed on the one hand to communities undergoing language shift (e.g. Mougeon and Beniak 1991) or lan-

guage death (e.g. Dorian 1981, 1989), and on the other, isolated individuals who happen to know two or more languages, but who are not (necessarily) constrained by group norms of usage (e.g. Woolford 1983, di Sciullo et al. 1986), is similarly intended to establish a baseline for *conventional* bilingual interaction against which other, perhaps idiosyncratic, behavior may be assessed.

The characterization of *bilingual* provided above imposes no *a priori* requirement as to degree of language proficiency required to be so classified (see e.g. Baetens Beardsmore 1982 on the difficulties inherent in such an assessment), and our studies have involved speakers of varying bilingual abilities when such individuals have been ascertained to represent core members of the bilingual speech community. Though level of bilingualism has not constituted a criterion for inclusion in or exclusion from our speaker samples, we regard the speaker's bilingual ability as a key explanatory factor of his actual linguistic performance. We thus take account of this factor by including it as an 'independent variable' in linguistic analyses of bilingual phenomena, as described in section (5.2.1) below.

Sustained contact between two languages may manifest itself linguistically in one or more of the following ways: code-switching, lexical borrowing on the community and individual levels, incomplete L_2 acquisition, interference, grammatical convergence, stylistic reduction, language death. Our understanding of these concepts has basically been informed by the classical and current literature in the field of language contact. Empirical quantitative analysis, however, requires us to operationalize these concepts such that they refer to mutually exclusive phenomena. Observation of their actual manifestations in discourse reveals that along with unambiguous instances of each, there exist other examples whose surface form does not permit ready classification as one or another result of language contact. We return to this issue below. The working definitions provided in what follows are based on unambiguous manifestations of these phenomena.[3]

Code-switching is the *juxtaposition* of sentences or sentence fragments, each of which is internally consistent with the morphological and syntactic (and optionally, phonological) rules of the language of its provenance. Code-switching may occur at various levels of linguis-

tic structure (e.g. *sentential, intrasentential, tag*) and it may be *flagged* or *smooth*. Intrasentential switching may occur at *equivalence sites* (where permissible switch points are constrained by word order homologies between switched constituents), or, more rarely, consist of *constituent insertion* (where word-order constraints *across* switch boundaries need not be respected for eligible constituents). The internal structure of the constituent is determined by the grammar of one language, but its collocation in the sentence is determined by the grammar of the recipient language.

Borrowing is the *adaptation* of lexical material to the morphological and syntactic (and usually, phonological) patterns of the recipient language. We distinguish established *loanwords* (which typically show full linguistic integration, native-language synonym displacement, and widespread diffusion, even among recipient-language monolinguals) from *nonce borrowings* (which though identical to loanwords in linguistic manifestation, need not satisfy the diffusion requirement). Loanwords generally are indistinguishable from native-language material at all but the purely etymological level, fail to be recognized by speakers as being of foreign origin, and do not involve active *borrowing* per se in any but the historical sense, as they are transmitted naturally along with the remainder of the monolingual lexicon. Though nonce borrowings show the same patterns of morphological and syntactic integration as established loanwords (in contrast with code-switches, which remain *unintegrated*), they do require active access to the L_2 lexicon, and in this sense they resemble code-switches.

Convergence also involves the process of borrowing, although we reserve this term for the transfer of grammatical structure (e.g. plural marking, agreement rules, etc.) from one language to another. Unlike lexical borrowing, it does not involve *adaptation* of other-language material to recipient-language grammar, but consists rather of the *introduction* of (unadapted) other-language patterns into the recipient-language system. Also in contrast to lexical borrowing, which generally features an etymologically foreign form, convergence may involve no *visible* other-language material (as in e.g. the transfer of a word order). In fact, convergence need not involve any transfer at all: it may simply consist of the selection and favoring of one of two (or

more) already existing native-language forms which coincides with a counterpart in the contact language (e.g. Klein 1980). (Other types of borrowing which do not involve surface indications of other- language material include *calquing* (e.g. Sp. *rascacielos* based on Eng. *skyscraper*) and *semantic shift* (e.g. Fr. *librairie* based on Eng. *library*)).

Though we have not actively focused on these in our research, we see *incomplete L_2 acquisition* as a (possibly fossilized) state of the language acquired through formal means and not used for normal interactional purposes, and *interference*, as the unpatterned, *idiosyncratic* manifestation of any of the above-mentioned language contact phenomena.

Stylistic reduction is the narrowing of the stylistic repertoire available to the individual, which may or may not be accompanied by concomitant expansion via incorporation of stylistic options from the other language. Stylistic reduction may also affect every level of linguistic structure available for style shifting, and may manifest itself as 1) undue preference for only one of several available variants of a variable, thereby obviating the choice mechanism and depriving that variable of its stylistic connotations (e.g. Lavandera 1978), or 2) continued use of all of the options, but failure to *distribute* them appropriately according to style (e.g. Gal 1984), or 3) preference for one or another member of a stylistically marked lexical doublet without reference to contextual appropriateness (Miller and Poplack, forthcoming). *Language death* is the gradual diminution of domains seen as appropriate to the use of L_2 until such time as none remain, and at different stages of this process, may or may not be accompanied by linguistic change due to contact (e.g. Dorian 1981).

Because code-switching, borrowing, incomplete language acquisition, and interference may result in utterances containing elements of two languages, each of these bilingual behaviors has at one time or another been used as evidence about another. And because convergence, stylistic reduction, and language death need involve no overt elements of the other language, they may remain undetected by any but the most systematic examination, except in cases where the resulting structure is clearly ungrammatical by the standards of one of the two contact languages. (e.g. Fr. *Je **suis** 14 ans* 'I am 14 years old'; as

opposed to J'***ai*** *14 ans*, lit. 'I have 14 years'.). Long-term examination of these issues has led us to conclude that each of these mechanisms for combining material from two grammars within a single utterance results from different processes and is governed by different constraints (see also, e.g., Grosjean 1990). This observation is generally uncontroversial when it comes to unambiguous manifestations of these processes. The problem is that it is often difficult to infer synchronically which mechanism has produced a given utterance. As in the case of (monolingual) syntactic ambiguity, this is because different processes can result in the same surface string. Given present knowledge, it does not seem possible to identify *a priori* every token on a case-by-case basis. In section (5.3) below, we illustrate how variationist methodology, when applied systematically to corpora of bilingual discourse, with special attention to cases where the different mechanisms have *different* manifestations, can contribute to the resolution of this problem.

In ensuing sections we briefly address four of the methodological and analytical tenets associated with the variationist framework, insofar as they can be applied to issues in language contact. These are: 1) the use of appropriate data, 2) the selection of informants to ensure representativeness and the knowledge of what they represent, 3) the principle of accountable reporting, and perhaps most important of all, 4) circumscription of the variable context, or defining the object of study.

4.0. Methods

4.1. Appropriate data and collection procedures

The notion of appropriate data gained importance in variation studies when it became apparent that styles of speech other than the vernacular are often characterized by unsystematic hypercorrection *away* from the speaker's native speech patterns. Thus (monolingual) speakers may not only fail to produce underlying segments in contexts in which they are expected, but when attending to their speech they may also re-insert them non-etymologically (cf. Eng. *tuna-**r**-on toast*, Fr. *huit-z-autres*

'eight others,' Sp. *un sojo* 'an eye'). This behavior is particularly frequent when the variable involved is stigmatized, as the manifestations of language contact have been reported to be in most communities. We are not aware of reports of 'hypercorrect' bilingual behavior per se; what does seem to be the case is that in formal or awkward or other speech styles perceived to be inappropriate, those manifestations subject to conscious control tend to be avoided altogether. As an example, Table 1 shows that in the speech of one Puerto Rican informant, code-switching occurs at least four times as often in informal or vernacular speech situations, providing the interlocutor is also an ingroup member, as opposed to simply a fluent bilingual.

Speech Style	Number of code-switches	Number of conversation minutes	Average number of code-switches per minute
Formal (ingroup)	87	90	1
Informal (nongroup)	107	120	1
Informal (ingroup)	152	30	5
Vernacular (ingroup)	54	15	4

N = 400

Table 1: Average number of code-switches per minute by speech style and group membership (after Poplack 1981)

When the interlocutor does not enable code-switching, for example by fulfilling the conditions of group membership and/or succeeding in establishing an interaction perceived to be appropriate for it, not only does it occur infrequently, but (in this particular case, though not shown in Table 1) the incorporations from English are largely restricted to nouns, and ethnically-loaded or untranslatable nouns at that (Poplack 1981), which are ambiguous as to their status as 'true' code-switches. So while the vernacular/ingroup data show a full gamut of intrasentential, intersentential and tag switching, English incorporations collected by the outgroup member (the author) were extremely limited.

Restricting the object of study to the 'vernacular' has not proved to exclude potentially important data associated solely with other speech styles. For one thing, certain bilingual behaviors (including code-switching, and to an extent, borrowing (Poplack, Sankoff & Miller 1988)) are themselves *hallmarks* of vernacular style. For another, and this has also been our experience with monolingual linguistic variables (with the possible exception of purely lexical ones), the data comprising the bulk of the other styles is *included* in the vernacular materials, while the reverse is not the case (e.g. here, informal styles include some noun incorporations, but formal styles show little or no intrasentential switching).

Perhaps the richest, most copious data on code-switching it has been our privilege to work with were the Puerto Rican Spanish/English materials collected by Pedro Pedraza in the course of nearly seven years of participant observation of a single block in East Harlem, New York. The sheer volume and quality of the data he obtained enabled us not only to detect many instances of rare switch types previously thought to be non-existent or not permissible (e.g. between pronominal subject and verb, between auxiliary and verb, switches of lone determiners, etc. (Poplack 1980, 1981)), but also enabled us to discover that even within a single well-circumscribed community, different *patterns* of code-switching could coexist, differentially employed by different groups of speakers. Since very few of us are permitted the luxury of investing several years in data gathering, we continue to experiment with ways of approximating that situation.

A basic methodological requirement of our studies of bilingual, minority and/or stigmatized language situations is that the raw data be collected by skilled interviewers who not only are, but are also perceived by informants to be, ingroup members, and whose own linguistic repertoires feature the same phenomena we are attempting to elicit. In our experience only interviewers with these characteristics are consistently capable of creating the appropriate interactional conditions to enable linguistic manifestations of language contact that are subject to conscious control.

The elicitation techniques employed within the interview setting do not take the form of direct questioning about the bilingual behavior

in question, but are rather adaptations of the 'sociolinguistic interview' (e.g. Labov 1966, 1984; Labov et al. 1968, Sankoff & Sankoff 1973, Wolfram & Fasold 1974, Poplack 1979, 1989; Baugh 1979): a loosely structured set of topics preselected by the interviewer to mirror current, local and/or individual interests, minimally including childhood games, customs, folklore, recipes and narratives of personal experience. The interviewer is instructed to follow the informant's lead in topic shifting, and only introduces a topic when none appears forthcoming from the informant. The content of each interview will thus vary from informant to informant, but we find that a common core of subject matter generally recurs. Where information is required concerning language attitudes (questions which are by nature more formal), these may be asked at the end of the interview, or at a posterior meeting. The entire conversation is tape-recorded (with the permission of the informant), and constitutes the raw data for all subsequent analyses. As will be obvious from the description of our collection procedures, these interviews contain, in addition to (varying amounts of) data on the language contact phenomena of interest, ample attestation of at least one, if not both, of the (monolingual) codes in contact. In fact, it has been our experience that most bilingual phenomena are as a rule extremely sparse in running discourse (e.g. in our French/English materials, code-switches occur anywhere from not at all to 132 times in an interview, loanwords represent between 0.1% and 2.5% of the total lexicon employed by an individual, unambiguous cases of convergence are exceedingly rare, etc.). It is thus our policy to collect as *much* data as possible (sometimes up to five hours per informant), in the hopes of obtaining a sufficient number of spontaneous attestations of these rare phenomena.

The purely monolingual portions of the interview are also fundamental to the inquiry, as they play a crucial role in establishing whether a given feature is appropriately analyzed as resulting from contact. The codes entering into the contact situation may themselves show regional or non-standard features not found in normative varieties, which may or may not result from prior interlinguistic influence. For example, we would be obliged to consider a borrowed form like *afforder* rendered with a retroflex [ɻ] as failing to show phonological integration into French, if we were not aware that the retroflex

variant had already penetrated the Canadian French phonological system, where it presently co-varies with apical [r] and velar [ʁ], even in French-origin words, and among French monolinguals. Though the retroflex variant may well be due to contact in the *historical* sense, considering it on a par with *synchronic* manifestations is tantamount to classing the voiced palatal fricative realization [ʒ] of *garage* in the speech of a contemporary French/English bilingual as due to influence from French. Admittedly, this is its ultimate source, but not within the lifetime of the speaker.

Communities may also evolve innovative compromise solutions to the problem of reconciling two languages, with no apparent counterpart in either of the monolingual codes. This is the case of double stress assignment to bisyllabic nonce loans in Canadian French: main word stress is assigned according to English rules, shifting stress to the left, while syllable stress is assigned according to French (e.g. *quíèt*). On the one hand, this pattern forms part of the stereotypical 'French Canadian accent' in monolingual English discourse, and so could be considered due to English influence, but on the other, its use in French discourse appears to be restricted to flagging nonce borrowings.[4] These kinds of facts are crucial for the decisions the linguist ultimately makes regarding the identification of a given phenomenon as resulting from language contact.

4.2. Selection of informants

We have been referring to ingroup and outgroup members, implying the existence of some entity one can be a member of, which in turn leads to the question of the optimal informants for a variationist study of language contact phenomena. It is uncontroversial that any speaker with any degree of knowledge of more than one language is theoretically capable of combining them in any way she chooses. There have been ample reports in the literature, usually in the guise of counter-examples to proposed constraints, of the learned use of foreign words and expressions, cross-language punning and other bilingual word-play observed among academics, family or friends. The variationist seeks to deter-

mine the actual *role* of such phenomena in the bilingual repertoire. A key component of the variationist research program (in monolingual as well as bilingual discourse) is to distinguish the isolated, and perhaps idiosyncratic, token from the regular patterns that characterize natural exchanges in the speech community.

It has been observed repeatedly that membership in a social network imposes clear restrictions on the behavior of members (e.g. Labov et al. 1968, Milroy 1980). Our studies of language contact phenomena within this framework have shown that such restrictions are not directly predictable from the typological relationship or other purely linguistic features of the languages in contact, and are often stronger than these would warrant. To cite but one example, in the Puerto Rican community in Harlem, code-switching is copious, transitions between languages are smooth, and it occurs at all possible switch boundaries, of which there are many, given the typological similarities between the languages. Moreover, no special rhetorical effect appears to be accomplished on the *local* level, i.e. by the *individual* switch (Poplack 1980, 1981, Sankoff & Poplack 1981). The situation differs markedly in the French/English bilingual communities in the Ottawa-Hull region of Canada. Here only a very small proportion of the code-switching is genuinely intrasentential. Instead of juxtaposing the two languages smoothly, Ottawa-Hull francophones draw attention to, or 'flag,' their switches, by different discourse devices: metalinguistic commentary, English bracketing, repetition or translation. In fact, just about every switch serves a rhetorical purpose, and to accomplish this purpose it must be flagged, and should not pass unnoticed (Poplack 1985). These differences cannot be ascribed to the linguistic configuration of the contact language pairs, since they are typologically very similar. For reasons detailed elsewhere (ibid.), we conclude that the different code-switching patterns stem from differences in community norms, which must be empirically established on a case-by-case basis.

Much of our work (as indeed, much of the sociolinguistic work in the field of language contact more generally) has been based on small-group studies, using standard social network methodology. As has been described by Milroy (1980, cf. also Poplack 1989), there is a major trade-off between the depth afforded by participant observation

and the scope available from 'survey'-type studies (Labov 1966, Sankoff & Sankoff 1973), where potentially explanatory extralinguistic variables (e.g. age, sex, socioeconomic class, educational level, etc.) may be manipulated in ways not possible in the study of self-selected peer groups. In particular, a recurrent criticism of network studies concerns their possible lack of representativeness. In 1982, we began to confront this problem by supplementing our ethnographically-oriented studies of bilingual behavior with a large-scale study of bilingualism in the adjoining cities of (officially anglophone) Ottawa and (officially francophone) Hull, which together constitute the national capital region of Canada (Poplack 1989)[5]. One hundred and twenty francophone informants were selected using strict random sampling procedures and stratified according to age, sex, and minority vs. majority language status of the French language in their neighborhood of residence. Random sampling ensures that informants meeting predetermined quotas are fully representative of the (francophone) population of the region. Each sample member is also identified according to socioeconomic status, educational attainment, level of bilingual ability, and neighborhood of residence, and each of these factors is regularly incorporated as an independent variable into studies of her linguistic behavior. The inclusion of such factors in our linguistic analyses has enabled us to uncover sometimes unexpected extra-linguistic constraints on bilingual behavior which we could not have intuited, such as the finding that membership in the speech community is more important than bilingual ability in determining borrowing rates (Poplack 1988), or the social class constraint against established loanwords (Poplack, Sankoff & Miller 1988).

5.0. Data Manipulation

5.1. Transcription and handling of primary speech data

The raw data on which all our studies are based consist of tape-recorded naturalistic conversations containing (some) bilingual phenomena

which will vary in type and degree according to the individual informant. The tape-recordings are typically searched exhaustively for a given feature (e.g. loanwords) and *all* instances of that feature are extracted for future analysis, in keeping with variationist analytical methods to be described in more detail in section (5.3). This procedure is then repeated for each subsequent feature under study.

Because the sheer size of the French/English corpus (approximately 3.5 million words) precludes repeated exhaustive searches, we resolved to transform these data into machine-readable form. This involved transcribing, correcting and entering the entire corpus onto computer, an undertaking which took several research assistants approximately three years of full-time work to complete. Space does not permit full explanation of the transcription protocol (see Poplack 1989); suffice it to say here that there is a major conflict between level of transcription detail and subsequent accessibility of the data, and the first crucial decision the analyst/transcriber must make concerns where the materials will be located on the continuum between them. In our French-Canadian data, for example, the word *père* is variously realized with a lowered, raised, or diphthongized [ɛ], and with a velar, apical or deleted [r]: [p ɛʊ], [paᴵr], [per], [peʊ], [pe], etc. Similarly, the loanword *high-rise* was produced as follows: [a: ɹáiz], [ai ráiz], [hái ɹaiz], etc. Since each of these variant realizations may have different social meaning in the community, we initially wished to distinguish them in our transcription.

But accounting orthographically for numerous phonetic realizations of a single lexical item means that in a study involving just one of these words, its occurrences would have to be located under six or seven separate entries. When this is multiplied by the 17,000 or so lexical types occurring in the corpus, the number of sites which must be searched to extract lexically identical forms becomes unmanageable. To facilitate the automated treatment of the data and maximize accessibility we thus adopted a solution of standard orthography for our transcriptions while still preserving much of the pertinent variability. Our overall strategy was to represent variation resulting from the operation of phonetic or phonological processes in standard orthography, regardless of the actual pronunciation of the form (i.e. all of the realizations

listed above were transcribed as *'père,' 'high-rise'*). If, on the other hand, the variant realization affected an entire morpheme (e.g. the variable deletion of [l] in *l'église*, as in (1), these were represented as produced.

(1) Puis j'étais mariée à (Ø< [l]) église catholique puis toute. (091/1147)[6]
'And I was married at the Catholic church and all.'

This transcription protocol extends to English interventions in the text: these are also transcribed according to standard English orthography, even if there is a current French alternative. Dialect orthographies like *bines* 'beans,' *filer* 'to feel' are represented by us as '*beans*,' '*feeler*' in the interest of better accessibility and reduction of homography. Because this is a bilingual corpus, we of course wished to flag interventions from English for purposes of automatic recognition. We initially attempted to distinguish unambiguous code-switches, unambiguous loanwords and intermediate forms. For tagging purposes, a code-switch was provisionally defined as any sequence of two or more English words, other than compound nouns (e.g. *science-fiction*, *real-estate*, *baby-sitter*), whose status must be established using other criteria, and proper nouns (e.g. *Born-again*, *Women's Lib*). Other lone lexical items of English origin known to be widely used in the region were considered for these purposes to be loanwords. Words whose status is doubtful (e.g. single French words calqued on English forms, such as *insulation*, *capabilité*, *déshonnête*, *dépressé*), or nonce loans (e.g. *patroller, expropriétait)* were to be classed in an intermediate category.

Perhaps not surprisingly in retrospect, the tagging procedure failed for all but the unambiguous code-switches. Since the transcribers were (of necessity) native speakers of the dialect(s) under study, it quickly became apparent that in most cases they were incapable of identifying many loanwords as etymologically English. As they were themselves accustomed to designating *sewer* as *sour* [su ʁ], and *beans* as *bines* [bIn], etc., they had no reason to consider them less 'French' than other *canadianismes* like *char* 'car' (an example which, in contrast, was (erroneously) classed as borrowed). Moreover, with

few exceptions, there was no way for the transcribers to determine which potential loanwords were in fact widespread, before having transcribed a few dozen of them. Since months could elapse between two encounters with the same loanword, and since it was not feasible during the transcription phase to keep counts of each of the 20,000 occurrences of borrowed forms (while at the same time applying other aspects of a detailed transcription protocol to thousands of other items), we were forced for the sake of consistency to leave borrowed items unmarked. So while we do in fact have statistics on the frequency and level of diffusion of every borrowed form in the corpus (Poplack, Sankoff & Miller 1988), these were only obtained after first extracting them *manually* by reading through the entire 3.5 million word document.

A number of automated data handling programs were run on the interview files, in particular, the Oxford Concordance Program (Hockey & Marriott 1980). Figures 1 and 2, reproductions of entries in the Ottawa-Hull French Concordance, illustrate the organization of the data in alphabetical order by lexical type, along with the total number of occurrences of each type (or keyword), followed by every instance of its occurrence in the corpus. Each occurrence is preceded and followed by its immediate discourse context and accompanied by an address (speaker number and line number in the complete transcript of his individual interview) to facilitate retrieval of additional contextual information when necessary. The frames presented illustrate, among other things, the occurrence of the noun *pad* and the verb *pack* in the guise of a borrowing (*elle voulait avoir un* ***pad*** 'She wanted to have a pad' (063/1853); ... *rien dans une couple de rangées faut tu* ***packes*** '... you only have to pack in a few rows' (14/354)) and as part of an unambiguous code-switch (*you took my writing* ***pad*** (013/623); *Faut tu* ***pack*** *your own au Basics* 'You have to pack your own at Basics.' (014/356)).

SPEAKER# →	LINE# ON TRANSCRIPT		KEYWORD →	pack 2 ← N OCCURRENCES	
014 355		(F) les affaires de même là? C' est, tu (A)	pack		your own (F) puis à Basics je le sais pas s' il
014 356		s' il faut tu ... Je pense que oui. Faut tu (A)	pack		your own (F) au Basics, oui. Ça va plus vite
		CS FLAG →	packboy 1	← CS FLAG	
075 148		ils en ont une job. Quand même ça serait pa--	packboy		oubedonc livraison, ils- va- va dans les
			pack-boys 2		
014 351 ←		occupé là, bien il y a des fois qu' ils ont des	pack-boys		mais ... quand-qu' il y a pas assez de pack-boys
014 352		pack-boys mais ... quand-qu' il y a pas assez de	pack-boys		tu sais ... (1) Ça doit être long, je sais pas
			packe 1		
005 2197		des Anglais, on a un chauffeur puis le gars qui	packe		les tubs, puis toute le restant c' est toute
			packer 3		
007 1156		de besoin. (007) Vois-tu moi j' étais (A)	packer		and helper (F) moi dans le temps du- de l' armée
007 1168		toute la (068) place. Tu sais, tu appelais ça (A)	packer		and helper (F) dans le temps. (inc) du
031 3482		(031) Ah, j' étais ... Comment-ce tu appelles (A)	packer.		(2) Ouais? Puis ensuite de deça? C' est là vous
			Packers 2		
081 924		un peu d' argent. Quand mon père travaillait à	Packers		là, on- on vivait bien. (2) Mhm. (1) Mhm. (081)
081 926		je pense, trente-quatre. (1) Mhm. (081) ... À	Packers,		tu sais, sontaient maudits dans ce temps là
			packes 2		
014 350		pas au ... Loblaw' s là, non. Non, faut tu ...	packes		ton- tu sais quand c' est bien occupé là, bien il
014 354		est rien dans une (W) couple de rangées faut tu	packes.	← LOAWORD (UNMARKED)	Comme si les (A) express, (F) les affaires de
			packetait 2		
105 1761		buggy, ça brassait un peu. On appelait ça- ça se	packetait		hein tu sais, on- on disait que ça packetait, ça
105 1761		se packetait hein tu sais, on- on disait que ça	packetait,		ça descendait de deça des fois. Puis le monde

Figure 1. Ottawa-Hull French concordance for 'pack' (Poplack 1989)

			CS FLAG →	pad 7	
013	623	(013) Okay. ... (inc). (A) You took my writing	pad,		eh? You took everything, eh? (6) (inc) le
063	1853	commencé ses périodes. Puis elle voulait avoir un	pad.		Nous-autres c' est un pad. (1) Ouais. (063) Elle
063	1853	elle voulait avoir un pad. Nous-autres c' est un	pad.		(1) Ouais. (063) Elle demande pour un pad là-bas
063	1854	est un pad. (1) Ouais. (063) Elle demande pour un	pad		là-bas, un pad là-bas c' est un affaire pour
063	1854	Ouais. (063) Elle demande pour un pad là bas, un	pad		là-bas c' est un affaire pour écrire dessus. (1)
068	1690	me promenais sur la grande-rue puis icitte avec un	pad		tu sais là. Ah sainte! C' était tannant. Quand j'
068	1694	tu es Eulalie aujourd'hui? Je leur montrais mon	pad,		tu sais? Bon bien ils me flippaient la page puis
			LOANWORD (UNMARKED) ↗	padé 1	
080	158	jusqu' aller en-arrière du cou icitte là, toute	padé		(inc) là. (1) Oui. (080) Ça d' épais, je vous
				PADI 1	← INTERVIEWER
099	456	à cinquante longueurs de n-- natation tandis que	PADI,		ça c' est un association internationale
				Padre 1	
056	869	(2) Ah ouais, ouais. (056) Dans le camp. Puis le	Padre,		c'est lui qui était comme interp-- interprète
				Padre-Foot 2	
056	858	là, de (A) German storm troopers? (F) Un nommé	Padre-Foot,		lui il a gagné la (A) Victoria Cross, (F) la
056	865	(1) Ah. (056) Ah oui, (A) fighting-Padre.	Padre-Foot.		Foot. (2) Puis vous l'avez rencontré là-bas
				pads 4	SPEAKER # →
008	813	pour jouer au hockey pour- on s' usait- pour des	pads.		(2) Ty-vrai? (1) Ah oui? (008) Ouais, on mettait
054	652	se mettait des- des livres de téléphone pour les	pads.		(1) Hein? (054) Des gros livres de téléphone
080	156	là, c' était toutes des- c' était toutes des	pads		ça d' épais, tu sais en ouate là ... (1) Oui
105	731	catalogues de chez Eaton' s puis on faisait des	pads		pour le goaler. (rire). (2) Ah mon-Dieu ça se
				Paf 2	
033	119	puis Holland. Puis ils ont fermé la porte.	Paf!		(2) Puis ça- a ty été là votre dernière job
091	1758	a frappé avec sa main, ça se peut puis ça a fait	paf!		Il m' a pas maganée puis il m' a pas sauté sur

Figure 2: Ottawa-Hull French concordance for 'pad' (Poplack 1989)

5.2. Secondary or reported data

Other types of data which are crucially important to the interpretation of bilingual speech production include information on speaker 1) characteristics, 2) attitudes and 3) perceptions.

5.2.1. Sociodemographic speaker characteristics. In the course of the 'sociolinguistic interview' described above, an attempt is made to obtain as much information as possible on the sociolinguistic background of each speaker. This typically includes a detailed account of the speaker's residential, educational, employment and linguistic history, as well as purely demographic information. On the basis of these and other data culled from the interviews, each speaker in the Ottawa-Hull sample was assigned a score on an English Proficiency Index (interpretable as a rough measure of level of bilingualism, since all of the informants have native abilities in French). The index is based on a combination of differentially weighted factors correlated with proficiency, including number of years of English-medium instruction, self-reports of English competence and propensity to use English according to situation, domain and interlocutor. All of this information is distilled into a 'sociolinguistic profile' for each speaker, which can be used as an independent variable in the explanation of his linguistic behavior.

5.2.2. Language attitudes. As part of our study of the New York Puerto Rican community, a detailed language attitude questionnaire (consisting of some 200 questions) based on standard social psychological methods was administered to each informant (Attinasi 1979). In reviewing the responses to these questions, some of which were self-contradictory, and others, ill-understood, it became apparent that by administering a questionnaire, the researcher not only predefines the possible attitudes that can be elicited (for closed questionnaires), but also the particular areas in which the respondent is permitted to express them (even in response to open-ended questions). Moreover, the very act of asking questions is likely to provoke some answer, regardless of whether the response reflects an idea that would even have occurred to the respondent if the interview had not taken place. In an attempt to

alleviate this problem in subsequent research, we exploited the fact that our French interviews were very long, and though generally not conducted in a question-answer format, tended to cover a number of topics related to the overall theme of francophone life in a bilingual setting.

From the conversations constituting the Ottawa-Hull corpus, we systematically extracted every overt remark that could be construed as reflecting an attitude about linguistic or ethnic matters, and proceeded, by content analysis, to exhaustively compare and group similar attitudes (Poplack & Miller 1985). We imposed no predetermined analytical or classificatory grid on them, but rather classed contrasting comments as a set of responses to some 'virtual' question. Over 100 such 'questions' emerged, many of them reminiscent of those familiar in traditional language attitude studies (e.g. Who speaks 'good' French? What do you think of two francophones who communicate in English?, etc.). Although not all informants provide a response to each, and some provide more than one, this method has the obvious advantages of not only revealing issues which are important to the informants, but of characterizing them in their own terms. Along with standard presentation of proportions of different answers to each question, we could also report what proportion of the respondents actually *brought up* the particular topic. This gives us access not only to opinions, but to the degree to which these opinions represent a real preoccupation of the bilingual informants in our sample. We were thus able to determine that though both minority and majority francophones manifest the same overt signs of linguistic insecurity (attitudes which are in fact pan-Canadian among the francophone populace), speakers residing in neighborhoods where French is the official and majority language reveal by their reported behavior and their preoccupations a covert linguistic *security* not shared by their minority counterparts, which is likely ascribable to the status of their language. Moreover, independent studies of the actual behavior of these groups show that these subtle attitudinal differences have identifiable linguistic correlates (Poplack 1988).

5.2.3. Speaker perceptions. Our linguistic analyses of the behavior of nonce borrowings and established loanwords have led us to consider

them as two (quantitatively different but qualitatively parallel) manifestations of the same phenomenon, as distinct from code-switching. But the psychological validity of this analytical decision for the bilingual speaker remained uncharted. We thus proposed to evaluate listeners' subjective reactions to different configurations of borrowed words (Poplack, Clément, Miller, Purcell & Trudel-Maggiore 1988). Adopting the matched guise procedure, we constructed a test tape consisting of sixteen stimuli, each containing a single English-origin form corresponding to one combination of the linguistic factors revealed to be significant in our earlier studies of loanword usage: 1) level of phonological integration (integrated or non-integrated), 2) level of morphological integration (integrated or non-integrated), and 3) levels of 'lexical' integration, here defined in terms of date of attestation of the word in French-language dictionaries and of its current diffusion across the community, as determined by the actual frequency of the word in the Ottawa-Hull French corpus. The instrument was administered to local native francophones, along with a questionnaire testing the identification, translatability and acceptability of borrowed words in different configurations of linguistic and social characteristics.

Subjective reactions to stigmatized linguistic variants are notoriously unreliable as predictors of actual usage. This problem is compounded in the case of incorporations from one language into another, as it may be impossible to determine whether eventual rejection is structural (i.e. refers to the manner in which the constituent is incorporated into the language), lexical (i.e. refers to the fact that the constituent does not form part of the lexicon of the judges' linguistic variety), or contextual (i.e. refers to the fact that the incorporation may be inappropriate to the type of interaction instantiated by the stimulus utterance). We therefore sought to reduce as far as possible the artificiality and contextual inappropriateness often associated by subjects with the simulation of stimuli by actors. To do this, we used as a source for our stimulus data actual utterances extracted from the Ottawa-Hull French corpus. Samples of the stimuli are provided in (2).

(2a) Stimulus 1: *boys* [b ɔ :Iz]
[-phonologically integrated] [-morphologically integrated] [attested before 1900] [widespread]

Pis l'homme qui sort avec les *boys* pis qui va à taverne pis qui rentre très tard, je trouve que tu retrouves ça ici. (026/882)

'And the man who goes out with the boys and who goes to the tavern and who comes home really late, I find that you find that here.'

(2b) Stimulus 3: *patroller* [patro:'le]
[+phonologically integrated] [+morphologically integrated] [unattested] [nonce]

Pis euh, fait que je peux pas voir pourquoi payer des gros salaires à ces policiers là, qui ont juste un mille carré à *patroller* là, tu sais? (019/1650)

'And uh, so I can't see why we should pay big salaries to those police officers, who have just one square mile to patrol, you know?'

The results of our study confirm and extend our earlier conclusions based on actual speaker behavior when using borrowed forms. A first important finding concerns the fact that subjects are often incapable of isolating an English-origin word in an otherwise French sentence if they have not been previously cued as to its existence, and this, regardless of the linguistic configuration of the word. Loanword identification appears to proceed as a lexical look-up operation. As might be expected, words categorized as forming an integral part of the French lexicon, i.e. those of long attestation and/or widespread diffusion, are identified as borrowed less frequently than unattested nonce borrowings. It is of interest, however, that the latter are still isolated less often than their widespread but unattested counterparts.

The linguistic configuration of the word assumes its role not for identification of the loanword, but for evaluation of the excerpt containing it. Speakers consistently rate borrowed forms more positively when they are integrated into French phonologically and morphologically, and this is true for each of the measures of acquiescence, affect, and surprisingly, normativeness. This pattern is as true of loanwords attested in French-language dictionaries since the turn of the century as of unattested nonce borrowings, lending further support to our decision to treat them together.

5.3. Data analysis

The discovery of linguistic patterns that hold for every speaker and every context is just as accessible to the intuitions of the variationist as to any other linguist. The difference arises when we deal with large quantities of natural speech data. There are correlations and variability from speaker to speaker and context to context that the variationist wants to account for that are less accessible to intuitions, and in fact, can only be clearly detected through quantitative analysis. These difficulties are exacerbated in the case of bilingual performance. For example, grammatical convergence which does *not* give rise to utterances which, when considered individually, are ungrammatical in the recipient language, but only to *preference* for an already existing structure with a counterpart in L_2, is a phenomenon which by nature eludes impressionistic observation. Similarly, there seems to be no self-evident way to intuit what it is that people are doing when they engage in intrasentential code-switching, by nature an aberration in terms of monolingual grammar. There are various strategies a speaker can adopt to minimize the clash between L_1 and L_2 phonologies, morphologies and syntax, and quantitative analysis can reveal which predominates in a given (social and linguistic) context.

Variationist linguistics (like other sciences of social behavior) cannot provide an immutable law for all eventualities. Linguists accustomed to observing natural interactions hear infelicitous or ungrammatical constructions produced by monolinguals on a regular basis. It is

thus not surprising that the same holds true for bilinguals. Quantitative analysis seeks to reveal the actual role (or the proportion) of initially questionable utterances within the larger system, i.e. whether they are idiosyncratic, or what some would call performance errors, or community norms. It can also shed light on the features of the environment which condition the choice of a particular structure.

5.3.1. The principle of accountable reporting. Two analytical principles underlying a quantitative variationist analysis are relevant to the study of language contact phenomena. The first is the principle of *accountable reporting* (Labov 1966). This requires not only that *all* the relevant examples of a phenomenon in some data set be incorporated into the analysis, but also, all of the contexts in which it *could have* appeared, but didn't. The sum total of occurrences and non-occurrences of variant realizations in a given context together constitute the *linguistic variable*, the key construct underlying variationist sociolinguistics. Thus, in studying variability in copula expression, for example, the variationist's data base will be constituted not only of all examples in which the copula was absent (3a), but also of those in which it surfaced ((3b) and (3c)):

(3a) If anybody (**ø**) in the way, well they'll mash him up. (4/275)
(3b) She**'s** older than this boy. (3/211)
(3c) His name **is** Son and his title **is** Nunez. (2/198)

The most immediate application of this principle to the bilingual context is in the determination of the *impact* of the various contact processes on the recipient language grammar. Language contact is (implicitly or explicitly) linked with linguistic change, but change is not brought about by a single deviant utterance. Processes like convergence and loanword incorporation are by nature quantitative. To assess the true role of a presumed change in the grammar of the *language*, it is necessary to count systematically the proportion of its occurrence, the contexts it has affected, and the speakers to whom it has spread.

The principle of accountable reporting poses special problems for bilingual data. In variable rule terminology, the examples in (3b)

and (3c) are known as 'non-applications' (of the copula deletion[7] 'rule'). But for at least some manifestations of language contact, no non-applications may be observed or inferred. In examining the claim (Klein 1980) that the Puerto Rican Spanish present reference system was converging with that of English, as evidenced by an increase in use of the progressive to refer to activity in progress at speech time, (an aspect also designated by the Spanish, but not English, simple present), it was a straightforward matter to extract from our bilingual corpus all morphologically simple and progressive present tense forms, and note for each, whether it referred to ongoing activity or to iterative/habitual actions or immutable truths. By comparing the proportions of different morphological forms used for each of these interpretations to each other and to both historical and synchronic monolingual Spanish data, it was possible to establish that no *increase* in the use of the progressive could be inferred, either over time or among those speakers with most bilingual ability in English. We thus concluded that if grammatical convergence were taking place in Puerto Rican Spanish, the present-reference system was not its locus (Pousada & Poplack 1982).

In terms of code-switching, however, the principle of accountability in its strict form is far more difficult to apply. This is because even if we could agree on where a true code-switch had in fact occurred, it is impossible to ascertain where one *could have* occurred but did not. This would require knowledge of the precise environments in which switching is permissible. Now since code-switching is first and foremost a discourse device, once the global situation is seen as appropriate, a code-switch is no more predictable at the *local* level than, say, a curse or a joke.

One way to resolve this is as follows: if we knew where code-switching was *prohibited*, as would be the case if there were purely syntactic restrictions on its occurrence, we could use this information to apply the principle of accountable reporting. In this connection, Sankoff & Poplack (1981) made use of the *equivalence constraint* on intrasentential code-switching (Poplack 1980, 1981) which states that codes may be switched intrasententially only when the word order of both languages is homologous on either side of the switch point. On this basis we could determine the syntactic boundaries at which a code-

switch was permissible (i.e. could have occurred) in addition to all those at which one actually *did* occur. We were thus able to estimate the *propensity* of switching at a given syntactic boundary. However, analysis of syntactic boundaries (even if limited to only permissible switch boundaries and even in a relatively short stretch of speech) is an extremely onerous task.

As far as borrowing is concerned, we have discovered no obvious way to determine the non-applications. Any content word in the language is fair game for borrowing (as to a far lesser extent, are function words). Only an infinitesimal number of them actually undergo this process, however, and still fewer proceed to achieve the status of established loanwords. We cannot predict which ones will be affected, since examination of the behavior of both nonce and established loanwords reveals that these do *not* tend to group naturally into specific semantic classes or to fulfill particular lexical 'needs' (Poplack, Sankoff & Miller 1988). Moreover, establishing the non-applications for loanwords would additionally require determination of the precise synonym(s) for every borrowed word. Even if this were feasible, there is no guarantee that any of them would appear in a given corpus, since in order for a lexical item to recur, a speaker must be talking about the thing to which it refers.

What we normally do in cases like these is extract the entire body of 'applications' (here, loanwords), and define a new 'dependent variable' within them. Poplack, Sankoff & Miller (1988) considered the entire corpus of 20,000 lone lexical items of English origin in French discourse. These potential candidates for loanword status were found to occur in four frequency categories in Ottawa-Hull French: *nonce* (used only once), *idiosyncratic* (used more than once but by a single speaker), *recurrent* (used more than 10 times) and *widespread* (used by more than 10 speakers), and we attempted to determine which were in fact true loanwords. This involved 1) locating a number of features associated with unambiguous loanwords (e.g. long-standing attestation, widespread dispersion, phonological, morphological and syntactic integration, recurrence, etc.) and 2) coding each token of each lexical type of English origin according to the extent to which it satisfied these criteria. We were thus able to draw a clear distinction between loan-

words and code-switches, in terms of their linguistic and social characteristics. As part of the same analysis we discovered that 'loanwords' and nonce borrowings could not be distinguished linguistically at any but the quantitative level, and only showed minor differences in terms of the speakers who used them. This confirmed our decision to treat them as manifestations of the same process.

5.3.2. Circumscribing the variable context. Perhaps the most controversial issue in the study of language contact phenomena is circumscription of the variable context. The first step a variationist will take in assessing contextual effects on the occurrence of one or another variant of a variable is to define the envelope of variation. If we want to determine the factors that promote, say, 'dropping the g' in forms like *workin'/working*, we must first locate the environments in which choice between the alternate realizations is even an option. In reviewing the potential candidates (i.e. forms containing the sequence *-ing*), we immediately discard tokens like *thing*, *ring*, *bring*, while retaining ones like *laughing*, *something*. Under main word stress, *-ing* is never reduced, though when unstressed, it often is. Inclusion of *thing* and *ring* in our data would not only have the effect of artificially lowering the overall deletion rate in the materials, since these would now include many contexts in which deletion never occurs, but more seriously, would blur the constraint hierarchy, or the pattern of *conditioning*, of the deletion process. How does this apply to the bilingual context?

Even if the analyst should be fortunate enough to dispose of a corpus containing many manifestations of language contact, s/he must still determine whether the other-language material constitutes a code-switch, or is a borrowing, or some other consequence of language contact. As we mentioned earlier, in empirical studies, it is often impossible, in a given sentence, to tell which of these processes has taken place. Though their results may be superficially similar, we submit that these processes are subject to different constraints and conditions, and that failure to separate them can only lead to confusing results.

5.3.2.1. Code-switching vs. borrowing. The problem of distinguishing code-switching and borrowing has prompted a number of studies on the characteristics of loanwords (e.g. Haugen 1950, Mackey 1970; Poplack & Sankoff 1984; Poplack, Sankoff & Miller 1988). It is generally reported that loanwords are phonologically, morphologically and syntactically integrated into the recipient language, and are recurrent and widespread. For nonce loans, however, the extralinguistic characteristics of recurrence in the speech of an individual and widespread distribution in the community do not hold. How can loanwords be distinguished from code-switches when this process is prevalent?

Close inspection of the results of the borrowing process (i.e. long-attested loanwords) reveals that they share a number of characteristics: they tend to be content words which take the same inflections and occupy the same syntactic slots as corresponding native recipient-language words. In the synchronic bilingual context, these facts can help distinguish loanwords from their original forms in the donor language, which of course take different inflections, if any, and may even occupy different slots. Specific tests for loanword status will vary from one language to another, depending on the particular morphological and syntactic features available.

Sankoff, Poplack and Vanniarajan (1990) studied combinations of Tamil, an OV language, and English, a VO language. Because of the differences in word order between the two languages, any switch involving an object NP will of necessity violate the word-order patterns of one or both languages. Yet it is precisely in object position where most of the tokens of English origin (generally consisting of single nouns) are found. Why should this language pair show so many apparently ungrammatical combinations, when the accumulating evidence suggests that languages are generally juxtaposed intrasententially in such a way as to result in *grammatical* sequences? There are at least two possible responses to this question. The first is that the structural makeup of the languages involved is disparate enough to permit few grammatical combinations. Should speakers of language pairs like Tamil/English wish to engage in code-switching, they would thus have no choice but to produce ungrammatical utterances. The second is that the 'offending' items are not in fact code-switches. This is where deter-

mination of the status of these elements becomes crucial. In these cases, we systematically compare their linguistic behavior with that of *unambiguous* code-switches and *unambiguous* loanwords. In the Tamil case, our analysis revealed that most of the single nouns in object position show the properties of borrowing and not of code-switching, i.e. they are accompanied by Tamil function words and carry Tamil case-marking. The fact that not all of the English-origin words are case-marked, however, again raises the question of whether the remainder are code-switches violating English word order. Quantitative analysis of *both* English-origin and *native Tamil* direct objects shows that, on the contrary, case-marking is variable on native Tamil as well as on borrowed English nouns. Moreover, comparison of marking rates shows that they are remarkably parallel. The borrowed forms contrast sharply with genuine code-switches from Tamil into English, which carry no Tamil case-marking, are accompanied by no Tamil function words, and begin and end only at syntactic boundaries which are equivalent in Tamil and English.

5.3.2.2. Nonce Loans versus Flagged Switches. In a study of bilingual behavior in English and Finnish, another postpositional language with case-marking, (Poplack, Wheeler and Westwood 1987), we again find that most of the English-origin material in Finnish discourse, consisting of single nouns and compounds, occurs in precisely those sites where true switches into English should be excluded.

As in the Tamil data, however, the majority of these nouns follow a Finnish function word and/or take the appropriate Finnish case-marker, indicating they are borrowings and not code-switches. Unlike the Tamil illustration, case-marking is obligatory in Finnish, but a good proportion of the English-origin nouns in the data are not case-marked.

Upon closer inspection, however, it became apparent that the presence of bare English-origin nouns in Finnish tends to be associated with an abnormal rate of certain discourse phenomena: in particular, pauses, ratification markers and flags, which in some conversations seem to be entirely confined to a switch-signaling function. Strikingly, the distribution of case-marking and discourse flagging of English-

origin single nouns tends toward complementary distribution. This confirms that most of these nouns (the case-marked ones) are nonce borrowings. The remainder are most logically treated as flagged, non-smooth single-word switches.[8]

5.3.2.3. Constituent Insertion. In a study of Moroccan Arabic/French bilinguals, Naït M'Barek & Sankoff (1988) found that by far the most frequent type of intrasentential language mixture is neither nonce borrowing, established borrowing, nor switching at equivalence sites, but rather insertion of a French NP, including at least determiner and noun, and optionally other elements, in a syntactic slot for an Arabic NP. For example, French DET + N is often inserted after an Arabic demonstrative or predeterminer *wahed*, contexts which take DET + N constructions in Arabic, but whose French counterparts would not permit the (second) determiner (see also Bentahila and Davies 1983). There are ten times as many NP insertions in all as there are switches at the equivalence site between Arabic DET and French noun.

That the process responsible for these data is NP insertion (rather than the equivalence switching predominant in the Puerto Rican data) is further confirmed by a greater statistical tendency for a second switch (back to Arabic) to occur after the French noun *only if this noun is in NP-final position.* If the NP continues, e.g. with an adjective or noun complement, then it is more likely to continue in French.

6.0. Discussion

The bilingual mechanisms discussed here are discretely different ways of solving the problem of combining material from two different languages. Each of them resembles the others in at least some aspect, and is distinctly different in another. Code-switching, constituent insertion and nonce borrowing are all (potentially) ways of alternating two languages *smoothly* within the sentence and in this, all contrast with flagged switching. Nonce borrowing differs from the other processes in that it involves syntactic, morphological and (variable) phonological integration into a recipient language of an element from a donor lan-

guage, whereas the other processes all maintain the monolingual grammaticality of the sentence fragment as determined by the rules of the respective language of its provenance. Indeed, nonce loans differ from established loanwords only *quantitatively* -- in frequency of use, degree of acceptance, level of phonological integration, etc. Constituent insertion differs from equivalence-based switching in that word-order constraints *across* switch boundaries need not be respected for those constituents eligible to be inserted. Switching at equivalence sites is the only mechanism which does not involve *insertion* of material from one language into a sentence of the other -- once a switch occurs, the rest of the sentence may continue in the new language (although further switches are also possible), whereas the other mechanisms generally require a return to the original language immediately after the nonce loan, inserted constituent, or flagged switch.

From a methodological point of view, it may be difficult to ascertain which mechanism has produced a given utterance. It seems clear that determining the status of the ambiguous item depends crucially on its linguistic and social context of occurrence. We have attempted to illustrate how quantitative variationist methodology, when applied systematically to representative corpora of bilingual discourse, can contribute to the resolution of these superficial ambiguities.

Notes

1 A preliminary version of this paper was prepared for a workshop on concepts, methodology, and data sponsored by the European Science Foundation Network on Code-switching and Language Contact in January, 1990. We thank the European Science Foundation for providing a forum for stimulating discussion of many of the issues presented here, and gratefully acknowledge the support of the Social Sciences and Humanities Research Council of Canada for much of the research on which this paper is based.

2 Throughout this paper we use *bilingual* to refer to *multilingual* as well.

3 Needless to say, some of these definitions, particularly those concerning the distinction between code-switching and borrowing, remain controversial. For detailed justification of those presented here we refer the reader to, e.g., Poplack et al. 1987, 1988; Naït M'Barek and Sankoff 1988, Sankoff et al. 1990.

4 This remains to be systematically studied.

5 This project has been generously supported from 1982 through the present by the Social Sciences and Humanities Research Council of Canada.

6 Codes refer to speaker number and line number of her/his utterance in the Ottawa-Hull French corpus.

7 Alternatively, the analyst may posit that (3a) is a non-application of the copula insertion rule.

8 Note that this type of flagging differs from the functional (or discourse) flagging reported among French/English bilinguals in Ottawa-Hull. In the Finnish/English materials flagging is associated with *production* difficulties, most likely attributable to the fact that the Finnish speakers in our sample did not belong to a community in which borrowing and code-switching are a discourse mode.

References

Attinasi, J. 1979. Language attitudes in a New York Puerto Rican Community. In R. Padilla (ed.), *Ethnoperspectives in bilingual education research: Bilingual education and public policy in the United States*. Ypsilanti, MI: Eastern Michigan University, 408-461.

Baetens Beardsmore, H. 1982. *Bilingualism: basic principles*. Avon, England: Multilingual Matters Ltd.

Baugh, John. 1979. Linguistic style shifting in Black English. Ph.D. Dissertation. University of Pennsylvania.

Bentahila, A. and Davies, E. 1983. The syntax of Arabic-French code-switching. *Lingua* 59:301-330.

Dorian, Nancy. 1981. *Language death.* Philadelphia: University of Pennsylvania Press.

Dorian, Nancy. 1989. *Investigating obsolescence: studies in language contraction and death.* (Studies in the social and cultural foundations of language 7) Cambridge: Cambridge University Press.

Gal, Susan. 1984. Phonological style in bilingualism. In Deborah Schiffrin (ed.), *Meaning, form and use in context*. GURT 84. Washington, D.C.: Georgetown University Press, 290-302.

Grosjean, F. 1990. The psycholinguistics of language contact and code-switching. *Papers for the workshop on concepts, methodology and data.* Strasbourg: European Science Foundation Network on code-switching and language contact, 105-116.

Gumperz, John. 1982. Conversational code-switching. In his *Discourse Strategies*. Cambridge: Cambridge University Press, 59-99.

Haugen, Einer. 1950. The analysis of linguistic borrowing. *Language* 26:210-231.

Hockey, S. & Marriott, I. 1980. *Oxford concordance program.* Version 1.0. Oxford, England: Oxford University Computing Service.

Klein, Flora. 1980. A quantitative study of syntactic and pragmatic indicators of change in the Spanish of bilinguals in the United States. In William Labov (ed.), *Locating language in time and space.* New York: Academic Press, 69-82.

Labov, William. 1966. *The social stratification of English in New York City.* Arlington, VA: Center for Applied Linguistics.

Labov, William. 1971. Some principles of linguistic methodology. *Language in society* 1:97-120.

Labov, William. 1984. Field methods of the project on linguistic change and variation. In John Baugh & Joel Sherzer (eds), *Language in use.* Englewood Cliffs, New Jersey: Prentice-Hall, 28-53.

Labov, William, Paul Cohen, C. Robins, & J. Lewis. 1968. *A study of the nonstandard English of Negro and Puerto Rican Speakers in New York City.* Philadelphia: U.S. Regional Survey.

Lavandera, Beatriz. 1978. The variable component in bilingual performance. Paper presented at GURT.

Mackey, W. F. 1970. Interference, integration and the synchronic fallacy. *Georgetown University round table on languages and linguistics 23.* Washington, D.C.: Georgetown University Press, 195-227.

Miller, C. & Shana Poplack. Forthcoming. Language contact and the stylistic repertoire.

Milroy, Lesley. 1980. *Language and social networks.* Baltimore: University Park Press.

Mougeon, R. & E. Beniak. 1991. *Linguistic consequences of language contact and restriction: The case of French in Ontario, Canada.* (Oxford studies in language contact). Oxford: Oxford University Press.

Naït M'Barek, M. & David Sankoff. 1988. Le discours mixte arabe/français: des emprunts ou des alternances de langue? *Revue canadienne de linguistique* 33:143-154.

Poplack, Shana. 1979. Function and process in a variable phonology. Ph.D. Dissertation. University of Pennsylvania.

Poplack, Shana. 1980. Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching. *Linguistics* 18(7/8):581-618.

Poplack, Shana. 1981. Syntactic structure and social function of code-switching. In R. Duran (ed.), *Latino discourse and communicative behavior.* New Jersey: Ablex, 169-184.

Poplack, Shana. 1985. Contrasting patterns of code-switching in two communities. In Henry Warkentyne (ed.), *Methods V: Papers from the fifth international conference on methods in dialectology.* Victoria: University of Victoria Department of Linguistics, 363-386.

Poplack, Shana. 1988. Language status and language accommodation along a linguistic border. In Peter Lowenberg (ed.), *Language spread and language policy: Issues, implications and case studies* (GURT 1987). Washington, D.C.: Georgetown University Press, 90-118.

Poplack, Shana. 1989. The care and handling of a megacorpus: the Ottawa-Hull French project. In Ralph Fasold & Deborah Schiffrin (eds), *Language change and variation* Amsterdam: Benjamins, 411-444.

Poplack, Shana, R. Clément, C. Miller, K. Purcell, & M. Trudel-Maggiore. 1988. Peut-on entendre l'intégration d'un emprunt? Paper presented at NWAVE-XVII. Université de Montréal.

Poplack, Shana & C. Miller. 1985. Political and interactional determinants of linguistic insecurity. Paper presented at NWAVE XIV. Georgetown University.

Poplack, Shana & David Sankoff. 1984. Borrowing: the synchrony of integration. *Linguistics* 22:99-135.

Poplack, Shana, David Sankoff, & C. Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26,1:47-104.

Poplack, Shana, S. Wheeler, & A. Westwood. 1987. Distinguishing language contact phenomena: evidence from Finnish-English bilingualism. In P. Lilius & M. Saari (eds), *The Nordic languages and modern linguistics 6*. Helsinki, 33-56.

Pousada, A. & Shana Poplack. 1982. No case for convergence: the Puerto Rican Spanish verb system in a language-contact situation. In Joshua Fishman & G. Keller (eds.), *Bilingual education for Hispanic students in the United States*. New York: Teachers College Press, 207-237.

Rand, D. & David Sankoff. 1988. *GoldVarb*. Logistic regression package for the Macintosh. Montréal: Université de Montréal.

Sankoff, David. 1979. *Varbrul 2S*. In Poplack, 1979. Appendix B.

Sankoff, David. 1982. Sociolinguistic method and linguistic theory. In L. Cohen et al. (eds), *Logic, methodology, philosophy of science VI*. Amsterdam: North Holland, 679-687.

Sankoff, David. 1988. Sociolinguistics and syntactic variation. In Frederick Newmeyer (ed.), *Linguistics: the Cambridge survey*. New York: Cambridge University Press, 140-161.

Sankoff, David. 1990. Dramatically contrasting language mixture strategies in two communities of fluent Arabic-French bilinguals. Paper presented at NWAVE XIX. University of Pennsylvania.

Sankoff, David, M. Naït M'Barek, & C. Montpetit. 1987. VSO/SVO bilingual syntax. Paper presented at NWAVE XVI. University of Texas at Austin.

Sankoff, David & Shana Poplack. 1981. A formal grammar for code-switching. *Papers in linguistics* 14,1:3-45.

Sankoff, David, Shana Poplack, & S. Vanniarajan. 1990. The case of the nonce loan in Tamil. *Language variation and change* 2,1:71-101.

Sankoff, David & Gillian Sankoff. 1973. Sample survey methods and computer-assisted analysis in the study of grammatical variables. In Regna Darnell (ed.), *Canadian languages in their social context.* Edmonton: Linguistic Research Inc., 7-64.

Sankoff, Gillian. 1974. A quantitative paradigm for the study of communicative competence. In Richard Bauman & Joel Sherzer (eds), *Explorations in the ethnography of speaking.* New York: Academic Press, 1-36.

Sankoff, Gillian & William Labov. 1985. Variation theory. Paper presented at NWAVE-XIV. Georgetown University.

di Sciullo, A., P. Muysken, & R. Singh. 1986. Government and code-mixing. *Journal of linguistics* 22,1:1-24.

Weinreich, Uriel. 1968. *Languages in contact.* The Hague: Mouton.

Wolfram, Walt & Ralph Fasold. 1974. *The study of social dialects in American English.* Englewood Cliffs NJ: Prentice-Hall.

Woolford, E. 1983. Bilingual code-switching and syntactic theory. *Linguistic inquiry* 14,3:519-536.

AMERICAN DIALECT RESEARCH

Offprint

EDITED BY

DENNIS R. PRESTON

This is an offprint from:

Dennis R. PRESTON
American Dialect Research
John Benjamins Publishing Co.
Amsterdam/Philadelphia
1993
ISBN 90 272 2132 4 (Eur.) / 1-55619-488-9 (US) (Hb)
ISBN 90 272 2133 2 (Eur.) / 1-55619-489-7 (US) (Pb)
© Copyright 1993 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm or any other means, without written permission from the publisher.