PROSPECTS FOR COMPUTER-ASSISTED DIALECT ADAPTION

David J. Weber
UCLA and Summer Institute of Linguistics
William C. Mann
USC Information Sciences Institute

When a document is to appear in several dialects or closely related languages, there are many practical reasons for adapting it from one to another rather than preparing separate translations. However, manual adaption can be tedious, errorful, and requires a bidialectal adaptor (often unavailable) and/or qualified linguist (if available, very expensive). Computer-aided adaption might be an alternative, but is it feasible to write a computer program which contributes enough to be worth the bother and cost?

This paper describes an experiment in computationally assisting the adaption of text from one dialect of central Peruvian Quechua (a polysynthetic, agglutinative, American Indian language) to several others. The overall results are extremely encouraging: computer-aided dialect adaption is feasible and has important advantages over entirely manual methods.

Below we describe the dialect situation, the data and processes of the experimental program, and a field test of text produced by the program.

Six dialects differing phonologically, lexically, and grammatically were involved. The rich diversity of differences are dominated by a few kinds of systematic difference. The program treats these classes of difference separately rather than by a single method (such as string substitution); this requires a detailed analysis of the source dialect text.

Examples of the kinds of differences involved:

Phonological: the reflexes of four proto-Quechua (*Q) phonemes (*/ç̌/*/č/*/λ/*/ñ/) --in the government-mandated orthography-- are: Panao: tr ch ll ñ, Huallaga: ch ch ll ñ, Dos-de-Mayo: ch ts l n, Llata: ch s l n, Yanahuanca: tr ts l n, and Junin: tr ch l n.

Lexical: 'to get well/recover from an illness' is expressed with the root allchaka:- in Huallaga, aliya:- in Llata, and kachaka:- in Junin.

Grammatical: 1) Morphological: a suffix may be present in one dialect and absent in another; the forms or properties of corresponding suffixes may differ across dialects; there are different systems of indicating plurality within the verb: in some dialects there are 3-5 distinct pluralizers occurring in different "slots" and conditioned by what other suffixes occur in the word; in others there is only one pluralizer which has a fixed position; 2) Syntactic: the complements of phasal verbs ('begin', 'finish'...) are subordinated as adverbial clauses in some dialects but as infinitival (object) clauses in others.

For the source dialect (SD) is provided a root dictionary and a suffix dictionary; each entry contains the string of characters by which that morpheme is recognized, morphological category, morphophonemic properties,... and for roots, the *Q form. For each target dialect (TD) is provided a suffix dictionary, a list of the regular sound changes (RSC's) which when applied to a *Q root will yield the correct TD reflex, and a list of pairs of roots not cognate in the SD and TD.

A TD root dictionary is computationally derived by 1) applying the RSC's to the *Q form of each SD root and 2) substituting the TD root for non-cognate root pairs.

Text is adapted word by word, first analyzing SD words and then synthesizing TD words. (An early pencil-and-paper experiment suggested that for Quechua, word-by-word methods could effect approximately 95% of the required changes.) After orthographic adjustments, a simple, recursive, exhaustive search attempts to decompose each word into a root and zero or more suffixes by matching the word's characters to the strings of characters of dictionary entries subject to constraints of a built-in morphology. Tests are applied 1) during the search to test the suitability of a matching suffix as the immediate successor to what precedes, and 2) after all the word is matched to morphemes to test the overall suitability of that sequence of morphemes. These tests constrain possible decompositions to within managable limits. Word decomposition is complicated by various morphophonemic processes.

Pluralization differences are accommodated by 1) tagging as plural each decomposition which contains a plural morpheme, 2) deleting that plural morpheme, and 3) inserting the appropriate pluralizer for the TD word.

Synthesis of a TD word proceeds by 1) substituting the SD root by the corresponding TD root from the computationally derived TD root dictionary, 2) selecting for each morpheme the correct allomorph, 3) concatenation of the allomorphs, and 4) orthographic adjustment. Examples are shown in Figure 1.

Many words have multiple decompositions but this is tolerable because synthesis of alternative decompositions of one SD word usually yield identical TD words. Nonidentical alternatives for one SD word are left to the choice of the human editor/checker.

About 40 pages of text were adapted into each of 5 target dialects for use in the field test. Sampling indicates that the computer correctly changed about 760 morphemes per 1000 words of text; in the worst case native speakers

SD orthographic form:	(1) aywarkaykan	(9) aywarkaarinanpaq

The word analyzer produces, in succession:

length converted:	(2) aywarkaykan	(10) aywarka:rinanpaq
segmentation:	(3) aywa-rka-yka-n	(11) aywa-rka-:ri-na-n-paq
morphophonemic form:	(4) *aywa-RKA-YKA:-3	(12) *aywa-RKU-:RI-NA-3P-PAQ
plurality handled:	(5) (*aywa-YKA:-3)+PL	(13) (*aywa-RKU-NA-3P-PAQ)+PL

The word synthesizer produces, in succession:

re-pluralization:	(6) *aywa-YKA:-YA:-3	(14) *aywa-RKU-YA:-NA-3P-PAQ
allomorph selection:	(7) aywa-yka:-ya-n	(15) aywa-rku-ya:-na-n-paq
TD orthographic form:	(8) aywaykaayan	(16) aywarkuyaananpaq
	'they are going'	'in order that they go'

Figure 1

suggested about 190 additional changes per 1000 words. The computer converted text which otherwise would have been at best only marginally intelligible to a speaker of another dialect into --with a few exceptions-- fully comprehensible text. Thus the program brings a text being adapted close enough to the TD that it can be edited/corrected by a native speaker of the TD without much coaching or reference to the SD text.

Since inevitably there is a non-trivial residue of changes infeasible for the computer, its output requires subsequent manual correcting/editing. Therefore, rather than strive to make the program do everything imaginable, it is wise to do the overwhelming number of systemic, "low level" changing and not unduly complicate the program to accommodate too much. The test identified many relatively infrequent changes not handled by the present program. For most of them, computational adaption is not feasible. These are discussed in a version of this paper which has been published by Notes on Linguistics, SIL. It is available from The International Linguistics Center, 7500 Camp Wisdom Road, Dallas TX 75236 for $.75 , as Special Publication 1, Prospects for Computer-Assisted Dialect Adaption.

Conclusion: A computer can contribute significantly to adaption between dialects or closely related languages.