Deseret Language and Linguistic Society Symposium

Volume 27 | Issue 1 Article 9

3-23-2001

Full Issue

Deseret Language and Linguistics Society

Selected Proceedings of the

Deseret Language and Linguistics Society

2001 Symposium

Editor, Dallin D. Oaks

SELECTED PROCEEDINGS OF THE

DESERET LANGUAGE AND LINGUISTICS SOCIETY

2001 SYMPOSIUM

EDITOR, DALLIN D. OAKS

PROGRAM HELD 22–23 MARCH 2001
AT BRIGHAM YOUNG UNIVERSITY, PROVO, UTAH

Deseret Language and Linguistics Society

Selected Proceedings of the Deseret Language and Linguistics Society, 2001 Symposium is a peer-reviewed publication. The ideas presented are the authors' and do not represent any official position of the DLLS or Brigham Young University.

The DLLS Symposium is sponsored by BYU's College of Humanities.

We gratefully acknowledge Professor Melvin J. Thorne and his students at the BYU Humanities Publication Center for their editorial and production assistance in the preparation of these proceedings.

CONTENTS

Discourse Intonation and Speaking English as a Second Language

John P. Broderick

This article closely analyzes the deployment of discourse intonation in video and audio recordings of adult speakers of English as a Second Language (ESL). The speakers held university degrees and were enrolled in an intermediate/advanced-level conversation class that was part of an intensive program at an American university, preparing them to meet admission requirements for graduate study in the United States.

The data sample studied here was originally elicited by Carolyn M. El-Kadi as part of a study of classroom interaction (El-Kadi 1994 and 1996). I am most grateful to Dr. El-Kadi for her permission to analyze some of her data for a somewhat different purpose in this study. The focus here is on very detailed analysis of a small segment of data (approximately three minutes of video and audio recordings of a conversation between a teacher and three adult learners of English as a Second Language).

The analytical methodology used in this study is based on the work of American linguist Wallace Chafe and British linguists Michael Halliday, David Brazil, Malcolm Coulthard, and Catherine Johns. The unit of analysis is the intonation unit. Chafe's notion of consciousness is at the core of the analysis, as is his particular view of the intonation unit as the primary locus in language where the signaling of the status of information in consciousness is realized. Chafe posits three statuses that information can have in consciousness (active, semiactive, or inactive) and three parallel modes of verbalizing the three kinds of information (given, accessible, and new). Typically, a falling or rising nuclear tone verbalizes new information, and a fall-rise nuclear tone verbalizes accessible information. Given information is typically verbalized by phonologically nonprominent syllables in intonation units.

The plan of the article is as follows: (a) Review certain analytical concepts that are central to the research methodology used in this study. (b) Describe the design of the study. (c) Report the results of the analysis, discuss the results, and briefly relate them to classroom practice in teaching English as a second language.

Review of Analytical Concepts

Let us begin our review of analytical concepts with a brief discussion of Wallace Chafe's views concerning consciousness and the status of ideas in consciousness during conversational interaction.

Introduction to the Work of Wallace Chafe

For nearly thirty years, Wallace Chafe has been developing a comprehensive, coherent, and highly creative model of spoken discourse that has shed interesting new light on the relationship between cognitive experience and language. Throughout his career, he has based his research on careful analysis of naturally occurring language data. During the 1970s and '80s, Chafe published a series of articles addressing issues such as the relationship between discourse structure and human knowledge (1972) and between language and consciousness (1974). He also wrote about givenness, contrastiveness, definiteness, subjects, and topics in discourse (1976); about the relationship between knowledge, experience, and verbalization (1977a, 1977b, and 1979); and about cognitive constraints on the deployment of consciousness and on the flow of information (1980, 1987, and 1988). In 1994, he published a landmark book-length synthesis of these and other ideas entitled *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing.*

At the core of Chafe's work are (a) his particular notion of consciousness as the cognitive capacity in humans that makes coherent spoken discourse possible and (b) his particular view of the intonation unit as the primary locus in language where the operations of consciousness are realized. Even though he has addressed many discourse issues besides these two, the intonation unit and its relation to the flow of consciousness are central to his work.

There are many other linguistic researchers (cf. in particular Halliday 1967; Brazil, Coulthard, and Johns 1980; and Coulthard 1992) who have done insightful work on the same discourse intonation phenomena that interest Chafe, but no other researchers have so explicitly related their work to a theory of consciousness as has Chafe. In the discussion that follows, I will therefore focus on Chafe's ideas, but it should be noted that (especially in the analysis of my own research data) I have also incorporated certain ideas and analytical tools from these other researchers (cf. Broderick 1995 for a description and rationale).

Wallace Chafe's Definition of Consciousness

For Chafe, consciousness is above all a process, a "limited activation process . . . an active focusing on a small part of the conscious being's self-centered model of the surrounding world" (1994, 28). That is, at any given moment, only a small portion of the vast store of knowledge that a person possesses can have the special status that consciousness confers. Chafe compares consciousness to vision, stating that it has a focus that is embedded in a surrounding periphery. For example, if you are paying attention to the language of this article, I have just activated the idea of "paying attention" in your focal consciousness, and at this moment, i.e., before I now mention them again, the names Michael Halliday, David Brazil, Malcolm Coulthard, and Catherine Johns were in your peripheral consciousness. At the moment I just mentioned those names, they were reactivated in your focal consciousness. After the next few intonation units, those names will be back in peripheral consciousness, and if I do not mention them for a paragraph or two, they may fade from peripheral consciousness as well. And so it goes.

We have just seen an example of how items introduced by a speaker or writer in the language of a discourse will activate or reactivate ideas in consciousness. But the environment in which communication takes place also plays a role. Until I mention it now, the chair you are sitting in was in your peripheral consciousness simply by virtue of being perceptible. Now, of course, I have used language fully to activate it in your focal consciousness. But unless it is reactivated, it too, like the names Halliday, Brazil, Coulthard, and Johns will quickly be replaced by something else.

For these reasons, Chafe characterizes consciousness as *dynamic*: information constantly flows into and out of both focal (i.e., active) and peripheral (i.e., semiactive) consciousness (29–30). That consciousness has a focus and a periphery and that consciousness is dynamic are what Chafe calls *constant* properties of consciousness, as is the fact that consciousness has a point of view (in ordinary conversational language it is self-centered; in fiction, point of view can be manipulated in various ways). Another constant property of consciousness is that it needs to be oriented in space and time (30). (Chafe notes that a person, knocked out, upon regaining consciousness, asks, "Where am I?" "What time is it?")

Consciousness also has several variable properties (30–35): (a) Conscious experiences can arise from different sources (perceptible events, feelings, introspections). (b) Conscious experiences can be "immediate" (i.e., based on what one is perceiving, doing, feeling at the moment) or "displaced" (i.e., based on remembering or imagining) (cf. also Broderick 1999). (c) Conscious experiences can be factual or fictional. (d) Conscious experiences can be more, or less, interesting. (e) Conscious experiences can be verbal or nonverbal.

Though its essence is that of a dynamic process, Chafe also refers to consciousness as a place: "the crucial interface between the conscious being and his or her environment, the locus of remembering, imagining, evaluating, and speaking, and thus central to the functioning of the mind" (40).

The Intonation Unit

Before defining the intonation unit and its relationship to the flow of consciousness, Chafe briefly discusses "echoic" memory, the phenomenon, long noted by psychologists, whereby sound remains briefly available to consciousness after it is physically over. The intonation unit is, according to Chafe, "a unit of mental and linguistic processing . . . that seems to be of exactly the right size to be processed in its entirety with the help of echoic memory" (55).

In his 1987 article, "Cognitive Constraints on Information Flow," Chafe defined the intonation unit as "a sequence of words combined under a single, coherent intonation contour, usually preceded by a pause" (22). He went on in that article to add that the intonation unit is the vehicle of expression of temporarily activated information, that it typically contains about 5 or 6 words, and that intonation units typically begin about 2 seconds apart (22). In his 1994 book, he elaborates considerably. In discussing those elaborations, I will be referring to the intonation unit transcribed in 1a and 1b:

(1a) .. and so the **háll** is rèal ló=ng%.
(1b) .. and so the háll is rèal ló=ng.

Chafe uses the term "accent" to refer to syllable prominences that are realized as pitch deviations from a mid or neutral baseline, usually higher, but perhaps lower. He transcribes what he calls primary accent with an acute accent mark, which indicates that the pitch deviation is accompanied by extra loudness and/or length. He transcribes secondary accent with a grave accent mark, which indicates that the pitch deviation is not accompanied by extra loudness or length. Presumably, the type of pitch deviation, loudness, and length involved in "accent" are of a qualitatively different kind from similar phenomena associated with what is usually called "word stress"; however, Chafe does not explicitly say this.

My cited example 1a is an exact replication of an example of an intonation unit that Chafe discusses at length in his book (1994, 58–61). He says that this is a detailed "narrow" transcription (59). Throughout his book, intonation units are in fact represented in a less detailed "broad" transcription such as I have provided in 1b.

Let us now look at additional aspects of Chafe's definition of the intonation unit. First, intonation units are often, but not always, separated by pauses. Short pauses (of less than 0.2 seconds) are transcribed with two periods. Pauses of between 0.2 seconds and one second are transcribed with three periods. Pauses of longer than one second are transcribed with three periods followed by a number in parentheses indicating the exact length of the pause. (In my own data analysis, I time all pauses of more than 0.2 seconds.) Intonation units are not delineated by pauses alone, because they may occur without a preceding pause, and pauses may also occur within them.

Second, intonation units are in some way delineated by changes in fundamental frequency (the clearest manifestation of the "coherent intonation contour" referred to above). However, Chafe explicitly asserts that they need not be limited to one primary accent as is "arbitrarily required [of the tone unit] in the British tradition" (58). (In my own data analysis, I have in fact adopted the British convention of limiting each intonation unit to one primary accent. For my rationale, see Broderick 1995.)

Third, changes in duration can help delineate intonation units. The smaller type font transcribing syllables toward the beginning of 1a indicates rapid articulation (Chafe borrows the poetic term "anacrusis" as a label for this phenomenon). The equal sign after the vowel in the last syllable of the intonation unit in 1a indicates lengthening. He says this speeding up at the beginning and slowing down at the end of intonation units is common.

Fourth, he says changes in voice quality of various kinds can also accompany intonation unit boundaries. The percent sign at the end of 1a is used to transcribe what Chafe characterizes as "creaky voice (laryngealization or 'fry')" (60).

Fifth, intonation units end in an identifiable intonation contour. Chafe lists three possibilities: a falling contour beginning on the last primary accented syllable in the intonation unit, which is transcribed with a period, as in 1a and 1b above; a rising contour, transcribed with a question mark; and what he characterizes as "everything else" (i.e., contours indicating continuation), transcribed with a comma. If an intonation unit is cut off, or in some other way clearly missing a terminal contour, then no terminal punctuation is used in the transcription. (In my own data analysis, I distinguish between two contours indicating continuation: a comma to mark a fall-rise tone, which seems consistently to appear in intonation units verbalizing accessible information, and a double dash to mark a level tone, which seems consistently to appear in intonation units where the speaker is concerned more with inner thoughts rather than with assessing the status of information in the listener's consciousness and marking its verbalization accordingly. For my rationale, see Broderick 1995.)

Sixth, Chafe points out that intonation researchers have long noted a tendency for intonation units to group into what are called "declination units," sequences of several intonation units throughout which the dominant pitch level gradually falls (59). The points at which these declination units begin and end can also help to delineate intonation unit boundaries.

Given, Accessible, and New Information

Chafe distinguishes three types of intonation units: fragmentary, regulative, and substantive. Fragmentary units are precisely that: false starts or units cut off by another speaker. Regulatory units are of four types: (a) textual, e.g., "and then" and "well"; (b) interactional, e.g., "mhm" and "you know"; (c) cognitive, e.g., "let me see" and "oh"; and (d) validational, e.g., "maybe" and "I think." However, it is in *substantive* intonation units that the role of consciousness is most apparent in that the cognitive processes that mark givenness, newness, and accessibility have their domain (Chafe 1994, 63–64).

Ideas (events, states, or referents) may have three statuses in relation to consciousness: (a) "active," i.e., "lit up" in a person's focus of consciousness; (b) "semi-active," i.e., present in a person's peripheral consciousness (the person has background awareness of it, but it is not being actively focused on); and (c) "inactive," i.e., in long-term memory (but neither focally nor peripherally active) (Chafe 1987, 25).

Ideas that are newly activated in consciousness at a given point in a conversation are verbalized as "new." Ideas that are already active in consciousness at a given point in a conversation are verbalized as "given." Ideas that are reactivated from a previously semi-active state are verbalized as "accessible."

Chafe's 1987 article, "Cognitive Constraints on Information Flow," analyzes in great detail and from a number of discourse perspectives a brief narrative taken from a longer conversation. Chafe's transcription of it contains 40 numbered intonation units. In it, the speaker talks about a class he took in college, describing the professor's manner in vivid detail. After introducing the ideas of "a big undergraduate course that I had" and stating that "everybody loved the instructor," the speaker produced the intonation units which I have numbered 2 and 3:

(2) ... a=nd he was a ... real .. uh .. óld world ... Swíss= ... guy,
(3) .. this was uh .. a biólogy course,

In 2, the word *he* verbalizes given information, and the words *real old world Swiss* verbalize new information. In 3 the words *this* and *course* verbalize given information, and the word *biology* verbalizes new information. This is because, according to Chafe, "language gives more prominence to new ideas than to given ones, prominence being recognizable in terms of full nouns (more prominent) versus pronouns (less prominent), and strong accent (more prominent) versus weak accent (less prominent)" (1994, 71). These examples of given and new information and Chafe's characterization of how language typically verbalizes given and new information are entirely representative of a rich tradition of research on this aspect of discourse structure (cf. Chafe 1994, 161–85, for a review of work in that tradition). One of Chafe's special insights is, of course, that such prominences verbalize the status of information in "consciousness" as he has defined it.

His other innovation is the addition of a third information status, "accessible," to the traditional binary distinction between "given" and "new." I have already noted that he asserts that ideas that are "semi-active" in consciousness are verbalized as "accessible." But what exactly does that mean? Let us look at 4, 5, 6, and 7, which are intonation units that occurred later in the same narrative cited in 2 and 3:

(4) ... a=nd he= .. wou=ld .. immèdiately open his ... nótes up, [*his notes* = "accessible"]
(5) ... in the front of the róom, [*the room* = "accessible"]
(6) ... and évery ... évery lecture, [*every lecture* = "accessible"]
(7) .. stárted the same wáy.

Chafe identifies the following words in 4, 5, and 6 as verbalizing "accessible" information: *his notes* in 4, *the room* in 5, and *every lecture* in 6. Notice that in each case, the cited phrase contains a primary accent, a feature commonly associated with new information. What is it that sets these phrases off as "accessible" rather than new? According to Chafe, they are accessible because they verbalize concepts that belong to a set of expectations associated with a "schema," in this case the schema of a college course (1987, 29).

Another reason to analyze an expression as a verbalization of accessible information is that it reactivates an idea that was mentioned previously but not very recently in a conversation. Here is an example from the same narrative. The intonation unit numbered 2 in this article occurred very early in Chafe's cited narrative: it was the fourth intonation unit in the 40-intonation-unit segment analyzed in his article. The unit I here number 8 occurred very late in Chafe's analyzed segment: it was the thirty-fourth unit in that narrative:

(8) .. I I guess that's the .. old world stýle, [*old world style* = "accessible"]

The idea of "old world this or that" was not verbalized at all in the intervening 29 intonation units.

In my own data (not only that analyzed for this study but extensive samples analyzed for other studies), I have found a strikingly consistent correlation, on the one hand, between the fall-rise pitch contour and the verbalization of accessible information and, on the other hand, between the falling (or rising) pitch contour and the verbalization of new information. This is an important point, for the fall-rise contour provides an objective, formal marker of accessible information to supplement Chafe's more subjective indicators, i.e., membership in a conceptual schema or previous mention in the discourse. The following examples—4a, which contains part of 4, and 4b, an invented example—might help you to "hear" the distinction between the fall-rise contour that verbalizes accessible information and the falling contour that verbalizes new information:

(4a) [at the beginning of each class] he= .. wou=ld .. immèdiately open his ... notes up, (*notes up* verbalizes accessible information)

(4b) [Guess what John did during lunch yesterday?] He opened his notes up. (*notes up* verbalizes new information)

Let me briefly summarize our discussion so far of given, accessible, and new information in discourse. Chafe gives us clear formal criteria that will help to analyze "given" versus "new" information in conversational data: given information tends to be verbalized as pronouns or as weakly accented words; new information tends to be verbalized as full lexical items with strongly accented words. But all four of Chafe's examples that I have cited of "accessible" information—4, 5, 6, and 8—seem, using his criteria, formally indistinguishable from verbalizations of new information. Apparently, subjective semantic judgments about what might constitute a member of a conceptual schema, or about how long it has been since prior mention of an idea in the same discourse, seem to be the only basis for identifying "accessible" verbalizations. The distinction seems quite reasonable, conceptually, especially in light of the intuitive soundness of the distinction between focal and peripheral consciousness. It is therefore useful indeed to add the fall-rise intonation contour as a formal marker of information verbalized as accessible.

Research Design for This Study

The idea for this study had two sources: (a) Throughout the 1990s, while teaching a graduate course titled First and Second Language Acquisition, I developed an interest in interlanguage, the special and systematically structured variety of English that arises in second language learners, the study of which can reveal insights into the second language learning and teaching process. (For an overview of interlanguage research, cf. Gass and Selinker 1994, chapters 2, 6, and 7). (b) Also in the 1990s, I served as a dissertation advisor to Dr. Mary El-Kadi, and, while reviewing some of her data, I noted certain features of the discourse intonation of the international students that both distinguished them from their teacher and also indicated that interlanguage patterns might be at work. Dr. El-Kadi made video and audio recordings of 12 hours of an intermediate to advanced level conversation class that met one hour a day, five days a week for seven weeks. She recorded selected hours toward the beginning, some in the middle, and some toward the end of the seven-week period. Students in the class had scored approximately 500 on the Test of English as a Foreign Language (TOEFL). In her own research, she studied several selections of data toward the middle of the course. Her focus was on the analysis of interactional patterns such as turn taking and on the role of the teacher in both modeling and directing conversational interaction.

For this study, I first listened to extended samples of Dr. El-Kadi's data in order to formulate hypotheses. I then selected a short sample of the data that was three minutes and seven seconds in length and studied it in considerable detail. My research associate, Cristina Leira, spent approximately 20 hours producing a first draft of the transcription, focusing on segmenting it into intonation units. After that I spent more than 40 hours refining the transcription, timing the pauses, and analyzing the various discourse phenomena associated with each intonation unit.

The analysis reported in this paper is of a WAV sound file that was made from the video recording using a Sony IC recorder model ICD-R100. That WAV sound file was then analyzed using a sound analysis computer program made available through the home page of SIL International (formerly the Summer Institute of Linguistics). The home page address is http://www.sil.org. The title of the software is "Speech Analyzer: A Speech Analysis Tool, Version 1.06a" (© 1996–1998 by Summer Institute of Linguistics: Acoustic Speech Analysis Project; see JAARS_ICIS Waxhow, NC; e-mail: speech_project_jaars@sil.org). This computer program displays the basic sound wave in various degrees of detail (making it possible to measure the length of pauses quite accurately) and can also display intonation contours. A fully analyzed transcription of the data is available in the appendix of this article.

Results and Discussion

Throughout this section of this article, the reader should refer to the full transcript of the analyzed data that appears in the appendix. Even though there were three students in the class during the three-plus minutes which were analyzed, one of the students (designated "Y" in the transcript) took only two turns at talk (turn numbers 21 and 23 in the transcript), producing only three intonation units (21a, 23a, and 23b), two of which were completely (23a) or partially (21a) unintelligible. The two students whom I focus on in the analysis are designated "K" and "G" in the transcript. "K" is a native speaker of Japanese, and "G" is a native speaker of Spanish. The teacher is designated "R" in the transcript. (These are the first letters of their first names.) An additional focus is on the substantive intonation units produced by those two students rather than on the regulatory or fragmentary units—this is because the mechanisms signaling verbalization as given, accessible, or new are operative only in the substantive units.

Table 1 lists the number of turns at talk in the analyzed segment of data and also the total number of intonation units, the numbers of each subtype of intonation unit (substantive, regulatory, and fragmentary), and the number of each type produced by the teacher and by each of the three students.

In the segment of analyzed data there were 56 turns at talk. The teacher, R, took 26 turns (about half the total); student K took 16 turns; student G took 12 turns; student Y took 2 turns.

There were a total of 123 intonation units, of which 63 (again, about half) were produced by the teacher. Student K produced 35 intonation units; student G produced 22; student Y produced 3.

Table 1: Turns at Talk and Intonation Units

	Total	Teacher (R)	Student (K)	Student (G)	Student (Y)
Turns at Talk	56	26	16	12	2
Total Intonation Units	123	63	35	22	3
Fragmentary Intonation Units	12	3	6	2	1
Regulatory Intonation Units	25	10	8	7	0
Substantive Intonation Units	86	**50**	**21**	**13**	2

Of the total of 123 intonation units, 12 were fragmentary units, 3 produced by the teacher, 6 by student K, 2 by student G, and 1 by student Y.

Of the total of 123 intonation units, 25 were regulatory units, 10 produced by the teacher, 8 by student K, 7 by student G, and none by student Y.

Of the total of 123 intonation units, 86 were substantive units, 50 produced by the teacher, 21 by student K, 13 by student G, and 2 by student Y. The focus of my analysis was on how student K and student G deployed discourse tones to signal the status of information in these substantive intonation units and on how their deployment of discourse tones differed from that of their teacher. Table 2 again lists the number of substantive intonation units produced by the teacher (R) and by student K and student G. It also reports on how many of those units verbalized new information and how many verbalized accessible information.

Of the 50 substantive intonation units produced by the teacher, 38 verbalized new information and 12 verbalized accessible information.

Of the 21 substantive intonation units produced by student K, 20 verbalized new information and only 1 verbalized accessible information.

Of the 13 substantive intonation units produced by student G, 12 verbalized new information and only 1 verbalized accessible information.

Table 3 focuses on the intonation units produced by the teacher, by student K and by student G that verbalized new information and reports on the tones used to signal that information status.

Notice that the teacher (R) always used either falling tone (in statements and *wh* questions) or rising tone (in yes/no questions) to signal the verbalization of new information, and he never used level tone to do so (as is appropriate, since level tone, as used by native speakers, indicates that the speaker is not in fact engaged in monitoring the status of information in the listener's consciousness and thus is not at that moment actively participating in the process of conversational interaction). But note that student K never used falling or rising tone to signal new information (as would

Table 2: Substantive Intonation Units Verbalizing New and Accessible Information

	Teacher (R)	Student (K)	Student (G)
Total Substantive Intonation Units	50	21	13
Substantive Units Verbalizing New Information	**38**	**20**	**12**
Substantive Units Verbalizing Accessible Information	12	1	1

Table 3: Tones Used to Verbalize New Information in Substantive Intonation Units

	Teacher (R)	Student (K)	Student (G)
Total Substantive Units Verbalizing New Information	38	20	12
Subtotal with Falling or Rising Tone	**38**	**0**	**4**
Subtotal with Level Tone	**0**	**20**	**8**

have been appropriate) but instead used level tone (inappropriately). Student G used falling or rising tone appropriately 4 of 12 times and inappropriately 8 of 12 times.

What is most interesting about the results of this study relates not to the percentages of intonation units of the various types and subtypes but to this manner in which the discourse tones are realized. Fully competent speakers of English signal the verbalization of active ideas as new information in discourse by using a falling tone on the tonic syllable of the intonation unit in statements and *wh* questions, and rising tone on the tonic syllable in yes/no questions. The speech of student K (the native speaker of Japanese) was most remarkable in this regard. In all 20 intonation units that he produced which verbalized new information, he used a level tone instead of a falling tone. Clearly, his interlanguage system does not yet use a falling tone as a means of marking new information. The speech of student G (the native speaker of Spanish) shows a similar tendency, but with exceptions. Of 12 intonation units that she produced which verbalized new information, she used a level tone instead of a falling tone in 8 of them. However, she did use the falling tone in the other 4 intonation units. Her interlanguage system thus contains the falling tone as a means of marking new information; however she uses it only a third of the time.

As noted earlier, each of the students being analyzed here produced only one intonation unit with the fall-rise tone that marks semiactive ideas verbalized as accessible information. We have already seen that this is a much lower percentage than that of the teacher relative to the number of intonation units verbalizing new information. Of additional interest is the manner in which student K realizes the fall-rise tone (not as a fall-rise, but as a level tone on a higher pitch—cf. intonation unit 42g in the transcript).

42 K a 2:21.3 ...(1.3) But sòmetimes
my mòther sáid--
b 2:24.2 ...(0.5) It's not fo=r
c 2:25.5 .. It's nót
d 2:26.3 .. *It*'s not .. góod for *me*--
e 2:28.1 Só--
f 2:28.4 .. *I* háve to--
g 2:29.4 ...(0.3) go exchànge the
clóthes, ((Fall-rise realized as a
level tone on a higher pitch.))

In the one instance where student G uses the fall-rise tone to mark accessible information (cf. 36d), she realizes it in the manner of a native speaker.

36 G a 2:06.0 ...(1.2) **Nó**. ((Sung on
three notes: level, very
high, level.))
b 2:07.8 ((Unintelligible.))
c 2:08.8 ...(1.0) *He*'s old enòugh
that--
d 2:11.0 ...(0.7) to know whàt .. *he*
wants to wéar,

These findings based on very careful analysis of a relatively small segment of data confirm impressions based on less detailed analysis of larger portions of the data elicited for Dr. El-Kadi's earlier study.

What conclusions relative to the process of second language acquisition might we draw from the findings of this study? First, in regard to student K, though he scored 500 on the TOEFL, he still has some way to go in mastering the refinements of the English discourse intonation system. He seems to know how to segment his speech into intonation units (though with difficulty—cf. his higher number of fragmentary units in Table 1), but he seems not to have mastered the actual phonetic realizations of the relevant discourse tones, as indicated by the complete absence of falling tones in his speech. The way in which he uses a raised level tone to realize the fall-rise tone in the one intonation unit verbalizing accessible information indicates that he is at least at the beginning stages of acquiring the system. One might even go so far as to say that he has acquired the tones at the "emic" (as in "phonemic") level, but has not yet acquired the tones at the "etic" (as in "phonetic") level.

Student G (the native speaker of Spanish), on the other hand, is well on the way to mastering the "etic" realizations of the system of discourse tones and may already fully have mastered them at the "emic" level.

Given the design of this study, it is not possible to assert with unqualified confidence whether the difference between the interlanguage systems of student K and student G in regard to the realization of discourse tones is due to their levels of competence as individual language learners or whether it may be explained by differences between their native languages (Japanese and Spanish, respectively) and English in the use of discourse tones; i.e., tones in Spanish, but not Japanese, may function more similarly to English. This question deserves attention in future research.

Although the findings of this study may need additional verification in order to make strong and conclusive inferences in regard to classroom practice, it is nonetheless reasonable to propose the use of data samples, such as the one analyzed in this study, in developing classroom exercises to assist students in mastering the English system of discourse intonation. Specifically, I propose the construction and use of exercises that focus on comparing the speech of a teacher (who is a native speaker of English) with that of students in data samples similar to the one analyzed in this article. The teacher's opening monologue in intonation units of turn 1 (a through k) is interesting in that it models all three of the most common discourse tones: the falling tone, the fall-rise tone, and the rising tone.

1 R a 0:02.6 Wéll.

b 0:02.7 *I*'ve got twò things plánned for *you*. [falling tone; new information]

c 0:04.8 .. this mórning, [fall-rise tone; accessible information]

d 0:05.4 ...(0.3) Úm--

e 0:06.2 ...(0.5) While *we* are wàiting for the óthers, [fall-rise tone; accessible information]

f 0:08.0 in càse *they* **dó** come, [fall-rise tone; accessible information]

g 0:09.0 tell *me* whàt *you*'re going to do this wée=kend. [falling tone; new information]

h 0:10.5 ...(3.0) *It*'s alrèady stárting now. [falling tone; new information]

i 0:15.0 ...(1.0) Wátcha gonna do, [fall-rise tone; accessible information]

j 0:16.8 ...(0.7) Do *you* have any pláns? [rising tone; new information]

k 0:18.3 .. Kóji? [rising tone; new information]

Or one could point out to students how the teacher in the data sample in this article models the correct tone when he repeats the student's previous intonation unit in 13a and 17a. (Note that even though it is unlikely that the teacher had explicit knowledge of a system like the one used here for describing discourse tones, he seems, in 13a and 17a, intuitively to have repeated the student's previous intonation unit specifically to model the correct tone.)

12 K a 0:44.0 Three
b 0:44.5 Thrèe casséttes--
c 0:45.5 ...(2.8) *Thàt'*s enóugh--
13 R a 0:48.8 ...(0.3) *Thàt'*s [enóugh.]

16 K a 0:53.5 I
b 0:53.8 *I* búy--
17 R a 0:54.3 *You* [búy *it*.]
18 K a 0:54.8 [*I* dó--]

More specific proposals concerning classroom practice must await additional research into the facts of discourse intonation in the interlanguage of learners of English as a second language, but clearly the principal finding of this study—that adult learners of English as a second language tend to use a level tone to mark new information instead of a falling tone in statements and *wh* questions and a rising tone in yes/no questions—can be used to give students practice in this, perhaps the most important, element of the English discourse intonation system.

REFERENCES

Brazil, David. 1985. *The communicative value of intonation in English.* Discourse Monograph No. 8. University of Birmingham: English Language Research.

———. Malcolm Coulthard, and Catherine Johns. 1980. *Discourse intonation and language teaching.* London: Longman.

Broderick, John P. 1995. Given, accessible, and new information: A comparison of Wallace Chafe's approach to analyzing discourse intonation with that of Brazil, Coulthard, and Johns. Paper presented to the Fortieth Annual Conference of the International Linguistic Association, Washington, D.C., March 11, 1995.

———. 1999. Wallace Chafe's light subject constraint in conversational discourse in the immediate mode of consciousness. *Word: Journal of the International Linguistic Association* 50 (2) (August): 143–54.

Chafe, Wallace. 1972. Discourse structure and human knowledge. In *Language comprehension and the acquisition of knowledge,* ed. Roy O. Freedle and John B. Carroll, 41–69. Washington, D.C.: V. H. Winston & Sons.

———. 1974. Language and consciousness. *Language* 50:111–33.

———. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In *Subject and topic,* ed. Charles N. Li, 25–55. New York: Academic Press.

———. 1977a. Creativity in verbalization and its implications for the nature of stored knowledge. In *Discourse production and comprehension,* ed. Roy O. Freedle, 41–55. Norwood, N.J.: Ablex.

———. 1977b. The recall and verbalization of past experience. In *Current issues in linguistic theory,* ed. Roger W. Cole, 215–46. Bloomington: Indiana University Press.

———. 1979. The flow of thought and the flow of language. In *Discourse and syntax,* ed. Talmy Givón, 159–81. New York: Academic Press.

———. 1980. The deployment of consciousness in the production of a narrative. In *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production,* ed. Wallace L. Chafe, 9–50. Norwood, N.J.: Ablex.

———. 1987. Cognitive constraints on information flow. In *Coherence and grounding in discourse,* ed. Russell S. Tomlin, 21–51. Amsterdam and Philadelphia: John Benjamins.

———. 1988. Linking intonation units in spoken English. In *Clause combining in grammar and discourse,* ed. John Haiman and Sandra A. Thompson, 1–27. Amsterdam and Philadelphia: John Benjamins.

———. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* Chicago: University of Chicago Press.

Coulthard, Malcolm. 1992. The significance of intonation in discourse. In *Advances in spoken discourse analysis,* ed. Malcolm Coulthard, 35–49. London and New York: Routledge.

El-Kadi, Carolyn M. 1994. *Linguistic theory applied to teaching practice: Looking through linguists' eyes at an urban ESL classroom.* Norfolk, VA: Old Dominion University. Doctoral Dissertation.

———. 1996. Discourse analysis of classroom interaction and the training of classroom teachers. In *Georgetown University roundtable on languages and linguistics 1995,* ed. James E. Alatis et al., 198–212. Washington, D.C.: Georgetown University Press.

Gass, Susan M., and Larry Selinker. 1994. *Second language acquisition: An introductory course.* Hillsdale, N. J.: Lawrence Erlbaum Associates.

Halliday, M. A. K. 1967. *Intonation and grammar in British English.* The Hague: Mouton.

JAARS-ICIS, Waxhaw, NC. "Speech analyzer: A speech analysis tool, version 1.06a." © Summer Institute of Linguistics, 1996–1998. (E-mail: speech_project_jaars@sil.org)

APPENDIX: TRANSCRIPTION OF DATA

Fragmentary Intonation Unit
(No terminal punctuation)
Regulatory Intonation Unit
Substantive Intonation Unit
Text Verbalizing Given Information
Text Verbalizing Accessible Information
Text Verbalizing New Information
0:00.0 = Minutes : Seconds . Tenths of Seconds
.. Pause of 0.2 Seconds or Less
...(0.0) Timed Pause of More than 0.2 Seconds
á, é, í, ó, ú, ý Primary Phrasal Accent (Tonic Syllable)
à, è, ì, ò, ù, ỳ Secondary Phrasal Accent
Bóldface: Contrastive Accent (on Tonic Syllable)

= Lengthening of Preceding Segment
, Fall-rise Tone on Preceding Tonic Syllable
? Rising Tone on Preceding Tonic Syllable
. Falling Tone on Preceding Tonic Syllable
! Rise-fall Tone on Preceding Tonic Syllable
-- Neutral Tone on Preceding Tonic Syllable
[ ] or [[ ]] Simultaneous Articulation
((Comment or Clarification))
R: English Teacher
K: Student (Native Speaker of Japanese)
G: Student (Native Speaker of Spanish)
Y: Student (Native Speaker of Japanese)

1	R	a	0:02.6	Wéll.
		b	0:02.7	*I*'ve got twò things plánned for *you*.
		c	0:04.8	.. this mórning,
		d	0:05.4	...(0.3) Úm--
		e	0:06.2	...(0.5) While *we* are wàiting for the óthers,
		f	0:08.0	in càse *they* **dó** come,
		g	0:09.0	tell *me* whàt *you*'re going to do this wée=kend.
		h	0:10.5	...(3.0) *It*'s alrèady stárting now.
		i	0:15.0	...(1.0) Wátcha gonna do,
		j	0:16.8	...(0.7) Do *you* have any pláns?
		k	0:18.3	.. Kóji?
2	K	a	0:18.7	.. *It* will be ráining--
		b	0:20.0	.. Sáturday and
		c	0:21.2	...(0.4) Sàturday and Súnday--
		d	0:22.6	I will
		e	0:23.7	Maybe *I* will stay hóme--
		f	0:25.0	...(0.4) in the dórm--
		g	0:25.6	...(0.4) and watch TV ((teevée))--
		h	0:27.0	or rènt a móvie--
3	R	a	0:29.0	...(1.0) Do *you* have a VCR? ((véeceearr))
4	K	a	0:30.9	Yés.
5	R	a	0:31.2	...(0.7) And you get

		b	0:32.5	Where do *you* go to rent the tápes,
6	K	a	0:34.0	...(0.5) Blockbuster vídeo--
7	R	a	0:35.6	Rìght on twénty--
		b	0:37.1	...(0.3) fí=rst--
		c	0:37.7	.. strée=t?
8	K	a	0:38.1	...(0.03) Yés.
9	R	a	0:39.0	...(0.4) How mány do *you usually watch* in a wéek.
10	K	a	0:41.2	...(0.4) Maybe
11	R	a	0:42.2	In a règular week
12	K	a	0:44.0	Three
		b	0:44.5	Thrèe cassèttes--
		c	0:45.5	...(2.8) *Thàt*'s enóugh--
13	R	a	0:48.8	...(0.3) *Thàt*'s [enóugh.]
14	K	a	0:49.3	[Ríght,]
15	R	a	0:50.3	...(0.3) And wou-
		b	0:50.8	.. do *you* cook pòpcorn or ánything?
		c	0:52.9	when yuh
16	K	a	0:53.5	I
		b	0:53.8	*I* búy--
17	R	a	0:54.3	*You* [búy *it*.]
18	K	a	0:54.8	[*I* dó--]
19	R	a	0:55.5	Do *you* have any suggéstions for *him*? ((Addressed to G and Y.))
		b	0:58.0	*It*'s a ràiny wéekend.
		c	0:59.3	*Three movies* take sìx hó=urs.
		d	1:01.5	What élse can *he* do.
20	G	a	1:03.2	...(3.0) Cléan--
		b	1:07.0	[((Laughing))]
21	Y	a	1:07.0	[...(1.5)] ((unintelligible)) cléan--
22	G	a	1:10.0	[[((Laughing))]]
23	Y	a	1:10.0	[[((Unintelligible))]]
		b	1:14.0	.. cléa=n yóu=r róo=m--
24	R	a	1:16.0	...(0.5) Do *you* wánna do *that*? ((K shakes head from side to side.))
		b	1:18.0	...(0.5) Nó?
25	K	a	1:20.0	...(0.5) *I* don't care--
26	R	a	1:21.8	*You don't càre* if it is dírty or clèan.
		b	1:23.8	..Okáy,
		c	1:24.3	...(1.0) Wéll,

d 1:25.2 I
e 1:25.3 ..*Wè*'ve gotta think of some óther actìvity.
f 1:28.1 ...(0.5) Maybe *they* have some suggéstion,
g 1:30.5 .. *You* can hèar what ***théy'*****re** going to dò,
h 1:32.0 ...(0.4) Ché=ck òut with *them.*
i 1:33.4 ..And sée *what they're gonna do.* ((Very softly spoken; see video.))
27 K a 1.34.7 ..Whát *you* gonna do-- ((To Gloria, almost inaudible; see video.))
28 G a 1:35.2 ...(1.0) *I heard they'*re cleaning hóuse. ((Followed by laughter.))
b 1:38.8 ...(0.4) *Sàturday* .. eh .. is a góod day for cleaning.
c 1:42.2 And *I* may go to the mall--
d 1:44.5 My oldest néphew--
e 1:46.0 is going to Orlàndo Flórida--
f 1:46.7 with the bánd--
g 1:47.5 ...(0.4) And *he* wants néw .. clòthes.
29 R a 1:49.5 ...(0.3) *He* wànts new clóthes?
30 G a 1:50.5 Yés.
b 1:51.0 ...(1.0) ((Unintelligible))
31 R a 1:52.7 ... (1.6) For ***hím***.
32 G a 1:54.7 ...(0.3) Yés.
33 R a 1:55.5 ..Whý do *you* have to go= for *hím.*
34 G a 1:57.1 Oh because maybe *I* buy sómething for *me*--
((Giggles through 35a-e.))
35 R a 1:59.8 Úh,
b 2:00.8 ...(0.8) Convénient.
c 2:02.2 Húh?
d 2:02.6 Do *you* advíse *him* on--
e 2:04.3 ..what clóthes he should wear?
36 G a 2:06.0 ...(1.2) **Nó**. ((Sung on three notes: level, very high, level.))
b 2:07.8 ((Unintelligible.))
c 2:08.8 ...(1.0) *He*'s old enòugh that--
d 2:11.0 ...(0.7) to know whàt .. *he* wants to wéar,
37 R a 2:13.3 ...(0.9) How old ís *he.*
38 G a 2:16.0 Fóurteen--
39 R a 2:16.5 ...(1.7) Did *you* decide on ***yóur*** *clothes* at *fòurteen*? ((Addressing K.))
40 K a 2:20.1 .. Yés--
41 R a 2:20.7 *You* díd!
42 K a 2:21.3 ...(1.3) But sòmetimes my mòther sáid--

b 2:24.2 ...(0.5) It's not fo=r
c 2:25.5 .. It's nót
d 2:26.3 .. *It's* not .. góod for *me*--
e 2:28.1 Só--
f 2:28.4 .. *I* háve to--
g 2:29.4 ...(0.3) go exchànge the clóthes, ((Fall-rise realized as a level tone on a higher pitch.))
h 2:30.8 .. with hér--
43 R a 2:32.0 *Exchange it* for sòmething élse.
b 2:33.5 .. Húh?
c 2:33.8 ...(0.3) *Something* that **shé** liked.
44 K a 2:36.0 ...(1.8) It was
b 2:38.0 ...(0.5) She gave
c 2.39.0 .. gave *me* móney--
d 2:39.9 .. So--
e 2:40.2 ...(0.8) I have to
f 2:41.7 ...(2.2) *I* háve to= --
g 2:45.5 ...(1.7) depènd on *hér*--
45 R a 2:48.0 ...(0.5) Uhúh.
b 2:48.9 ...(1.0) So *you hàd to* respèct [**hér** wishes.]
46 K a 2:50.9 [Yés--]
47 R a 2:51.5 **Hér** taste.
48 K a 2:52.0 .. Uhúh--
49 R a 2:52.5 Èven though *you* dìdn't líke *them.*
((R then switches eye contact to G.))
b 2:53.8 ...(1.4) At fourtéen.
50 G a 2:56.7 ...(1.7) Yés,
b 2:58.2 and *he* is **nót** at home.
51 R a 2:59.5 ...(0.5) Nó.
52 G a 3:00.3 Nó--
53 R a 3:00.7 ...(0.4) *He* won't excháng e *it?*
54 G a 3:02.3 ...(1.1) Úh-Uh.
55 R a 3:03.6 ...(1.0) *He* knòws what *he* wánts,
b 3:05.3 and *he* géts *it.*
56 G a 3:05.8 And
b 3:06.6 .. Yé=s. ((Ends at 3:07.2.))

Toward a Psychological Analysis of the Sentence from the Work of Lashley, Chomsky, Wundt, Polanyi, and Skousen's AML

Bruce L. Brown

Lashley (1951) and Chomsky (1957) clearly demonstrated the inadequacy of "left-right" associationistic models in accounting for language and other kinds of holistically patterned behavior. Both argued persuasively that the kind of holistic dependencies among elements that characterize language syntax cannot be explained through behavioristic S-R connections. Chomsky (1957, 18–25) began his attack on behavioristic theories of language by demonstrating the inadequacy of Markov processes (a precise embodiment of S-R chaining theory) in accounting for patterned sequences of behavior. In particular, he showed that the kinds of holistic dependencies among elements that characterize syntactic structures in language could not be accounted for with left-right associationistic models, but rather, would require a top-down hierarchical approach.

In his influential paper on the problem of serial order in behavior, Lashley (1951) made a similar case for the necessity of hierarchical explanation, but from a neurological point of view. The lines of his argument were quite different from Chomsky's. Chomsky's argument was essentially formal and based upon artificial models of logical mechanisms. Lashley's argument was neurological, but also conversational and straightforward. He first reviewed a variety of anecdotal observations concerning language and then asked what kind of neurological organization would be necessary to account for them.

He pointed out that a given set of phonemes in spoken words (or of letters in typed words) can occur in a number of combinations, such as the reverse combinations *right* and *tire* (p. 115). Lashley then made the very obvious point that "the order must therefore be imposed upon the motor elements by some organization other than direct associative connections between them" (1951, 115). He further argued that words stand in relation to sentences as letters do in relation to words, and that words also have no intrinsic temporal valence as implied by the associative chaining models. Drawing upon an analysis of the language translation process, he argued that this syntactic order is also not to be attributed to the thought process—the same thought can be expressed with quite different temporal structures in different languages. Translators translate holistic thoughts, not word by word. As he summarized: "the mechanism which determines the serial activation of the motor units is relatively independent, both of the motor units and (also) of the thought structure." Lashley (p. 115) argued that language is not the only example of this kind of syntactically structured behavior, that a multitude of skilled behaviors in man and other animals display this kind of implicit hierarchical structure and cannot be explained in terms of associative connections among the elements.

Wundt (1912, Chap. 7, "Die Satzfügung") reasoned from a very different perspective. His primary task was to explain the formation of sentences. He reasoned that any explanation of the sentence that focuses only upon its surface structure would obviously be inadequate. He characterized the sentence as "both a simultaneous and a sequential structure" (see p. 21 of Blumenthal). It is simultaneous because at any given moment it is present in consciousness as a totality even as the individual words are spoken. We focus upon the whole of what we are saying even as the words flow forth in a habitual way that is not introspectible to us. As he said:

> The sentence, however, is not an image running with precision through consciousness where each single word or single sound appears only momentarily while the preceding and following elements are lost from consciousness. Rather, it stands as a whole at the cognitive level while it is being spoken. If this should ever not be the case, we would irrevocably lose the thread of speech. (quoted in Blumenthal 1970, 21)

Like Chomsky, Wundt held that any explanation of the sentence that focuses only upon its surface structure would be obviously inadequate. But unlike Chomsky's position, both Wundt's account and that of Lashley left open the question of whether the psychology of the sentence requires one to posit the literal existence of syntactical rules in the human psyche.

Clearly a strong case can be made for an explanation of patterned serial behavior that does not attribute it to associative connections among the elements. However, we cannot consider that demonstration to be equivalent to making the case for rule-based explanations. Some have taken it this way. In particular, the Chomskian approach put phrase structure rewrite rules and transformational rules in center stage and imbued them with ontological status, thus opening the way for a new era of mentalism in the behavioral sciences. The new artificial intelligence (AI) brand of cognitive psychology further built upon this unbridled mechanistic mentalism, much to the detriment of a truly cognitive approach to explanation. The excesses of the AI movement were at least as outrageous as those of the behaviorists a decade or two earlier. The behaviorists insisted on mechanistic explanations, but also on the law of parsimony. Neo-cognitivists seem to be willing to sacrifice parsimony as long as a computer metaphor is satisfied, to guarantee mechanistic explanation.

But parsimony still makes sense. There is no reason to create complex, burdensome explanations if simpler ones will suffice. Polanyi's characterization of the nature of skills led the way for us here. He began his discussion of the psychology of skills (1962) with the trenchant statement:

> I shall take as my clue for this investigation the well-known fact that *the aim of a skilful performance is achieved by the observance of a set of rules which are not known as such to the person following them.* (49)

He then went on to offer explanations of the physical principles underlying swimming and riding a bicycle, but with the caveat that one certainly would not have to understand those explanations to perform either of these skills. Either of these skills is acquired tacitly through trial and error, or through apprenticeship, but without explicit awareness of the principles involved. This approach to explaining the acquisition of skills (including linguistic skill) is consonant with influential theories of perception, such as the "transactionalism" of Ames (1946) and Kilpatrick (1961) and J. J. Gibson's theory (1966) of "direct perception."

A full explanation of the principles involved in any of these skills is probably beyond our present scientific capability. However, Polanyi (1962) offered the following as a first approximation of the physical principles involved in riding a bicycle:

> Again, from my interrogations of physicists, engineers and bicycle manufacturers, I have come to the conclusion that the principle by which the cyclist keeps his balance is not generally known. The rule observed by the cyclist is this. When he starts falling to the right he turns the handle-bars to the right, so that the course of the bicycle is deflected along a curve towards the right. This results in a centrifugal force pushing the cyclist to the left and offsets the gravitational force dragging him down to the right. This manoeuvre presently throws the cyclist out of balance to the left, which he counteracts by turning the handlebars to the left; and so he continues to keep himself in balance by winding along a series of appropriate curvatures. A simple analysis shows that for a given angle of unbalance the curvature of each winding is inversely proportional to the square of the speed at which the cyclist is proceeding. But does this tell us exactly how to ride a bicycle? No. You obviously cannot adjust the curvature of your bicycle's path in proportion to the ratio of your unbalance over the square of your speed; and if you could you would fall off the machine, for there are a number of other factors to be taken into account in practice which are left out in the formulation of this rule. Rules of art can be useful, but they do not determine the practice of an art; they are maxims, which can serve as a guide to an art only if they can be integrated into the practical knowledge of the art. They cannot replace this knowledge. (49–50)

Obviously, being able to explain bicycle riding at this high level of abstraction and physical decomposition is not a prerequisite to performing the skill. There are many six- and seven-year-old children who have mastered the skill of riding a bicycle, but the explanation given above would probably mean very little to any of them. Nor would it make sense to hypothesize the existence of "rules" of this kind in their heads, a kind of inborn, unconscious, unintrospectible BRAD ("bicycle riding acquisition device"). There are many ways to explain what one is doing, some explicitly physical (such as the foregoing), some metaphorical, some complex, and some simple, but probably none of these levels of explanation fully captures or exhausts what is actually going on.

I remember hearing Chomsky say in a talk at McGill University in 1967 that the logical capability implicit in the linguistic performance of a typical three-year old is more complex than the principles of calculus. At the time I found that statement preposterous. With thirty-five more years of experience the statement now seems obvious and correct. The two sentences "They are easy to please" and "They are eager to please" at first seem alike in structure, and their surface structure is similar. However, an impersonal transformation shows that they are very different in deep structure: "It is easy to please them," but not "It is eager to please them." Polanyi would explain this in terms of the contrast between *tacit* knowledge and *explicit* knowledge. We have a tacit apprehension of linguistic principles of great depth and subtlety, but we do not have explicit knowledge of the principles involved. Chomsky's subtle and complex linguistic rules could be viewed in this framework as being an explicit spelling out of the logic underlying what every person can do linguistically without taking thought, without being able to introspect. T. G. R. Bower (1977) and his colleagues have shown that Piaget's (1954) developmental stages for children are much too conservative. Infants and young children have a tacit mastery of various cognitive tasks long before they can give proper explicit accounts, and Piaget made the mistake of basing his stages on what children would say, what they could explain.

One of the major approaches to language and cognition to come forth in the past thirty years is the work of the Parallel Distributed Processing (PDP) Group (Rumelhart, McClelland, and the PDP Research Group 1986), so-called "neural nets" or "connectionist models." The connectionist models capitalize on this "levels of explanation" approach, with the proposal that fairly simple associationistic mechanisms can be modeled on a computer to create close approximations to behavior that appears to be rule-governed. Chandler (1995) summarizes their major achievement: "They have shown that rule-like regularities can emerge from the massed interaction of relatively simple processes operating on homogeneous networks of information even though those networks contain and refer to no explicit representations of those rules" (234–35). The strategy is an ingenious one, and it has won for D. O. Hebb's neurological behavioristic associationism (on which PDP is based) a new hearing within contemporary cognitive psychology.

Skousen's (1989) analogical modeling of language (AML) also accounts for seemingly rule-governed behavior without recourse to explicitly represented rules. The approach is based upon a very simple principle of "natural statistics": to *minimize the number of disagreements* (Skousen 1992). In the same way that the complexities of hypothesized internalized linguistic rules can be avoided with this approach, the complexities of statistical decision theories can also be avoided. That is, there is no need to posit that the learner acquires some kind of "probabilistic rule" for dealing with linguistic categorization. Rather, his performance can be accounted for by the simple proposition that he samples from his own stored linguistic experiences using this one basic principle. Close approximations to actual performance can be achieved by adjusting the level of "imperfect memory." It is intriguing how such a simple hypothesized process can create complex behavior that could be explained at the highest level in terms of a complex and subtle rule system of the kind Chomsky has described.

Both the connectionist models (PDP) and AML are what Skousen (1995, 227) referred to as "procedural" as contrasted with rule approaches, which are "declarative." As procedural models, both AML and PDP avoid the major conceptual problems encountered in rule-based models. Skousen (1995) identified at least three such problems: rule-governed approaches cannot deal with "leakage" across category boundries; they are not robust in dealing with missing information or ill-formed context in the way that actual speakers are; and they are pushed to revert to a competence/performance distinction to account in an ad hoc way for failures of the model to deal with real, dynamic aspects of language.

AML has a number of features to recommend it over other available procedural language models. One is its explicit incorporation of episodic memory into the learning process. Another is its potential to account for more general perceptual processes beyond language. Both Skousen (1995) and Chandler (1995) have pointed to a number of failings of the connectionist models that AML seems to overcome. For one, connectionist models, once trained, are deterministic and cannot handle probability matching. Furthermore, connectionist network training can often require an inordinately long time even for simple behaviors, can get stuck in local minima, and even when trained cannot adjust to learn new input but rather collapses into predicting nonsense (the so-called "catastrophe problem"). AML, on the other hand, is particularly good at probability matching in a way that corresponds to actual human language learners. Also, no training is necessary, there are no local minima, and it adjusts well to new input, even contradictory input.

There are particular problems yet to be solved in the application of AML. One of the biggest problems is computational. With commonly used computational methods, each variable that is added essentially doubles the processing time as well as the memory requirements of the computer. Also, the notable successes of AML have been in the more well-defined areas of phonetics/phonology, orthography, and morphology. Application to more abstract and difficult areas of semantics and syntax has yet to be demonstrated. However, initial work with syntax looks promising. Lonsdale (2001), for example, has found some success in translating from French to English using *analogical cloning*, following the method of Jones (1996).

The *probability matching* aspect of analogical modeling is particularly interesting to psychologists in that it foreshadows the possibility of higher level theoretical integration with other established principles of human and animal behavior. A case in point is the well known *matching law* of Richard Herrnstein (1961) whereby probabilities of response are found to match probabilities of reinforcement. There are probably many linguistic examples of probability matching of this kind. Tucker and his colleagues (1968), as one example, have documented a linguistic probability matching in native French speakers with respect to the categorization of grammatical gender of "artificial" French words. They found a close match between the gender selection probabilities for various invented words and the gender probabilities for words with the same endings in *Petit Larousse.* Skousen (1995) recognized this capacity of AML to deal with the ubiquitous phenomenon of probability matching as one of the many advantages of AML over neural networks. AML can be seen as a sophisticated extension of associationistic principles, one that makes them capable of accounting for seemingly rule-governed behavior.

Given the arguments for the superiority of the AML approach to the modeling of human linguistic behavior, it could be argued that this paper has come full circle back to the associationistic approach criticized by Lashley and Chomsky. However, this is not just a case of "rocks break scissors, scissors cut paper, and paper covers rocks." A better metaphor would be an upward spiral, where the associationism implied in analogical modeling represents a much higher level of sophistication than the simple left-right associationistic chain theory that still falls vulnerable to the Lashley/Chomsky critique. Nor does it mean that with the continued ascension of analogical modeling we would expect to witness the demise of rule-governed approaches. In the concluding paragraphs of his fundamental work on analogical modeling, Skousen discussed the place of rule approaches:

> Despite the many arguments, both empirical and conceptual, in favor of an analogical approach to the description of language, there is a place for rule approaches too. An optimal rule description serves as a kind of metalanguage that efficiently describes past behavior and allows us to talk about that behavior. Whenever we attempt to summarize behavior or to discover relationships in data, our viewpoint is structuralist. But if we wish to predict language behavior rather than just describe it, we must abandon rule approaches. Rule descriptions have great difficulty in explaining actual language usage. (1989, 139)

Skousen went on to compare language rules with Boyle's Law and

Charles's Law as general physical laws that are only approximations to the real behavior of gasses. They are fairly accurate in accounting for gas molecules acting in the aggregate under most conditions, yet they have no real existence except in the minds of scientists. He made this comparison with linguistic rules:

> In no literal sense can it be said that individual gas molecules follow these laws. In a similar way, linguistic rules are meta-descriptive devices that exist only in the minds of linguists. Speakers do not appear to use rules in perceiving and producing language. Moreover, linguistic rules can only explain language behavior for ideal situations. As in physics, an atomistic approach seems to be a more promising method for predicting language behavior. (1989, 140)

This is reminiscent of Polanyi's characterization of a skillful performance as being achieved by the observance of a set of rules which are not known as such to the person following them. Linguistic behavior can be described in a general way by rules, but an analogical modeling approach is probably much closer to the actual psychological processes involved and accounts better for actual linguistic behavior (performance). Skousen's illuminating comments on the place of rules and analog constitute a fitting conclusion to his first published book on analogical modeling. They are also, perhaps, a promising prelude to the construction of a serious account of the psychology of the sentence, that mysterious process by which our holistic thoughts are automatically converted into a string of words.

References

Ames, Adelbert, Jr. 1946. Binocular vision as affected by relations between uniocular stimulus-patterns in commonplace environments. *American Journal of Psychology* 59: 333–57.

Blumenthal, Arthur L. 1970. *Language and psychology: Historical aspects of psycholinguistics*. New York: John Wiley and Sons.

Bower, T. G. R. 1977. *A primer of infant development*. San Francisco: W. H. Freeman.

Chandler, Steve. 1995. Non-declarative linguistics: Some neuropsychological perspectives. *Rivista di Linguistica* 7, no. 2: 233–47.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Gibson, James J. 1966. *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Herrnstein, Richard J. 1961. Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior* 4: 267–72.

Jones, Daniel. 1996. *Analogical natural language processing*. London: UCL Press.

Kilpatrick, F. P. 1961. *Explorations in transactional psychology*. New York: New York University Press.

Lashley, Karl S. 1951. The problem of serial order in behavior. In *Cerebral mechanisms in behavior*, ed. L. A. Jeffress. New York: John Wiley and Sons. 112–36.

Lonsdale, Deryle W. 2001. Recent advances in analogical cloning. Presentation at the Deseret Language and Linguistics Society 2001 Symposium. Brigham Young University.

Piaget, Jean. 1954. *Origins of intelligence*. New York: Basic Books (original French edition, 1936).

Polanyi, Michael. 1962. *Personal knowledge: Towards a post-critical philosophy*. New York: Harper and Row.

Rumelhart, David E., James L. McClelland, and the PDP Research Group. 1986. *Foundations*. Vol. 1 of *Parallel distributed processing: Explorations in the micro-structure of cognition*. Cambridge, MA: MIT Press.

Skousen, Royal. 1989. *Analogical modeling of language*. Dordrecht: Kluwer Academic Publishers.

Skousen, Royal. 1992. *Analogy and structure*. Dordrecht: Kluwer Academic Publishers.

Skousen, Royal. 1995. Analogy: A non-rule alternative to neural networks. *Rivista di Linguistica* 7, no. 2: 213–31.

Tucker, G. Richard, Wallace E. Lambert, Andre Rigault, and Norman Segalowitz. 1968. A psychological investigation of French speakers' skill with grammatical gender. *Journal of Verbal Learning and Verbal Behavior* 7: 312–16.

Wundt, Wilhelm M. 1912. *Die sprache*. Book 2, vol. 1 of the *Volkerpsychologie* series. Leipzig: Engelmann.

Evidence for Borrowing as the Reason for Exceptions to the Spanish Sound Change *f* to *h*

David Riding

Even though the Spanish sound change from *f* to *h* in word initial, prevocalic position has been declared "an irrefutable law of Spanish" (García de Diego, as cited by Levy 1973, 205), there are still many Spanish words that begin with *f*, such as *familia, fastidioso, fortuna, femenino,* etc. Scholars have generally explained these exceptions as borrowings from other languages, mainly Latin (Corominas & Pascual, in *Diccionario Crítico Etimológico Castellano e Hispánico* 1980, afterwards referred to as *DCECH;* Blake 1988). But some linguists question the validity of the exception argument, due to the lack of solid evidence. Blake warns about the possibility of circular reasoning, "namely, that all learnèd words preserve [f-], while all cases of *f*-preservation confirm the presence of a learnèd word" (1988, 53). However, two types of evidence show the validity of the conventional suggestion, that exceptions to this sound change are due to borrowing. First, words preserving the *f* are usually cultural vocabulary, which is more susceptible to borrowing. Second, words not having undergone the change from *f* to *h* have generally not undergone any of the other Spanish sound changes, meaning that these words appeared in the language after the changes occurred.

f TO *h*

The change from *f* to *h* in word initial position is one of the main phonetic elements that distinguish Spanish from other Romantic languages (Iribarren-Argaiz 1998). For example, Spanish *hacer* and Portuguese *fazer* both mean "to do," while Spanish *horno* and French *fourneau* both mean "oven." Word initial *f* did not change to *h* in every situation, only when followed by a single vowel, so words beginning with *fl-, fr-, fie-,* or *fue-* did not change.[1] At first the *h* was like the *h* in English, but later even the *h* dropped out to a silent letter, preserved solely in the orthography (Menéndez Pidal 1956, 199). Although evidence of the sound change can be found in literature throughout the first half of the second millennium (*DCECH*), the pivotal century was the 1300s, when the new feature became an accepted part of spoken Castilian Spanish, which would grow to be the dominant Spanish dialect (Blake 1988). There are two prevailing theories as to why *f* went to *h*. The first is that Castilian changed under the influence of a pre-Romantic language, most likely Basque, which has no labio-dental phoneme /f/ (Penny 1991, 79–80). The second theory, stated by Spaulding, is language internal; thus: "It *may* stand for the evolution of a bilabial *f*, which without difficulty becomes aspirate *h* by opening the lips, and which *may* have existed among the Romans"(1962, 90, *emphasis by Spaulding*). This phenomenon still occurs in many modern dialects, as native speakers are prone to pronounce words such as *fuerte* as *juerte* (Espinosa, as cited by Spaulding 1962, 91).

BORROWED WORDS PRESERVING *f*

Penny (1991) and Patterson (1982) show three ways Spanish has received vocabulary from Latin. First, from popularisms: words descended directly from Latin, used continually throughout the centuries. Second, from learnèd, or borrowed, words: whenever the Spaniards needed a word for a new concept (generally a nonmaterial aspect of life) they would find it in the Latin literature and copy it almost exactly. If these words, such as *feliz, fugaz, fábula,* and *formar,* were copied after the sound change from *f* to *h,* they retain the *f.* Any difference between the Latin word and its borrowed Spanish reflex will only be in the ending, which is sometimes modified to fit Spanish morphology. Third, Spanish has received vocabulary from semilearnèd words, which were learned from oral Vulgar Latin. They are words heard in such places as church or law courts, words such as *fallecer, fe, feria,* and *falso.*

Cultural vocabulary is much more likely to be borrowed than basic vocabulary. Cultural vocabulary includes culture-specific and conceptual words, while basic vocabulary "includes items such as pronouns, numerals, body parts, geographical features, basic actions, and basic states" (Crowley 1997, 171–72). They are words like *head, man, woman, flour, string,* etc. About 30% of the Latin words in *Vocabularium seu lexicon ecclesiasticum, Latino-Hispanicum* (Fernández de Santaella 1744) have common reflexes in modern Spanish and are of the type that would be susceptible to the *f* to *h* change. Of these, about half showed a transformation of *f* to *h* and half did not.[2] The items that show the change, meaning they descend directly from Latin and are not borrowed, can all be classified as basic vocabulary. Although not an exhaustive list, these are some good examples (in Spanish, not Latin): *higo* (fig), *hincar* (to nail, fasten), *hilo* (string), *hoja* (leaf), *hormiga* (ant), *horno* (oven), *humo* (smoke), *honda* (sling), *haba* (a kind of bean), *hacer* (to do, to make), *hablar* (to speak), *harina* (flour), *hacha* (ax), *herramienta* (tool), *hervir* (to boil), *hebilla* (buckle), *heder* (to stink), and *hosca* (dark) (Fernández de Santaella 1744). The words that not only preserve the *f* but are almost identical to the Latin do not adhere to Crowley's definition of basic vocabulary, meaning they are more cultural, or conceptual: *fabricar* (to manufacture, to make), *fácil* (easy), *fábula* (fable), *falso* (false, in a lying manner), *fama* (fame), *familia* (family), *fecundo* (fertile), *fértil* (fertile), *feliz* (happy), *feria* (holiday), *feroz* (ferocious), *fiar* (to trust), *fe* (faith), *figura* (figure), *fúnebre* (pertaining to a funeral or death), *fortuna* (fortune), and *firmeza* (strength, steadfastness) (Fernández de Santaella 1744).

Of this second list, Patterson says that *falso, fe, feria,* and *fiar* are all genetic words, not borrowed (1982, 21). Since they preserve *f* but are not borrowed, they are semilearnèd words, taken from verbal Vulgar Latin. According to Penny, these words should have religious or legal significance, which they do. *Fe* and *fiar* both come from the same root, *fidēs,* meaning "faith." In accordance with the definition of semilearnèd words, they have gone through some sound changes but not all (they preserve the *f*) (Penny 1991, 32). Although *feria* shows no sound change from the Latin, it does have a definite religious history. Its original definition was "holiday," many of which were religious. Then the name *feria* was applied to every day of the week, first to replace the names of pagan gods (such as *el Sol* and *la Luna y Martes*) and also to remind ecclesiastical workers that every day should be used for religious devotion. Portuguese preserves the word *feria* in the names of the days of the week (Fernández de Santaella 1744). Although it is a limited subset, this data corresponds with Penny's description of the three types of Latinisms in Spanish: popularisms, which have gone through the sound changes, are basic vocabulary; semilearnèd words, which show some but not all sound changes, are generally religious vocabulary; and learnèd words, which show little sound change, are less-concrete vocabulary.

WORDS WITH *f* LACK OTHER SOUND CHANGES

In addition to this semantic evidence, there is also phonetic evidence that many words were borrowed. There are many pairs of words in Spanish that are similar in definition and phonetic form and can both be traced to the same Latin root, such as *hablar* and *fabular*, both from the Latin *fābulare*. The difference between the words in each pair is that one preserves the *f* while the other has changed from *f* to *h* and undergone the other relevant sound changes that occurred in Spanish. Step-by-step analysis of the changes in the *h*-words demonstrates how the *f*-preserving words lack these changes, having a form almost identical to the Latin. If these words had preserved the *f* for a reason other than borrowing, they would not so systematically lack the other sound changes as well. This lack of change shows that they appeared in Spanish after the changes took place. I will present six such pairs, giving present and past definitions, etymologies, and derivations. The definitions help show that the word in each pair that has changed, for example *hablar* in the case of *hablar* and *fabular*, is a more basic vocabulary item than the word that was borrowed at a later time. Citations on derivations refer to sound change rules; the derivations are my own. Each change is in the correct chronological order as far as my sources specify, but some are uncertain.

fastidio, hastío

The Spanish words *fastidio* and *hastío* both trace their origins back to the Latin word *fastīdium* (*DCECH*), meaning "weariness, monotony, an aversion, loathing" (Diamond 1961). Both Spanish words have meanings close to their common root (Peers et al. 1960). In the early seventeenth century,[3] *fastidio* carried this same meaning, but *hastío* had an additional, more specific meaning in everyday Spanish—having little desire to eat and the abhorrence of food because of an upset stomach (de Covarrubias 1943). This would seem similar to the way we use the word *fastidious* in English, which is the characteristic of being hard to please because of previous experience in the area (*Oxford English Dictionary* 2001). The first instance of *hastío* in literature was in 1495, while *fastidio* appeared in 1251 (Corominas 1967). The word *hastioso*, an adjectival form, was later replaced by *fastidioso* (*DCECH*).

Fastīdium went through many changes to become *hastío*[4]:

fastīdĭum ——>	*fastīdĭu*—word-final *m* is eliminated in Latin by first century BC (Penny 1991, 74)
fastīdĭu ——>	*fastidio*—vowel leveling from classical Latin to Vulgar Latin (syllable final *u* becomes *o*) (Resnick 1981, 47)
fastidio ——>	*fastiyo*—*dy* (a stop followed by a palatal glide) becomes just the palatal glide
fastiyo ——>	*fastío*—*y* in contact with *e* or *i* is deleted (Resnick 1981, 67)
fastío ——>	*hastío*—*f* is converted to *h*

hijo, filial

Hijo and *filial* are not exact cognates, but they do have the same root, so they can be used to show how *hijo* has changed over time. *Hijo*, meaning "son" or "child" descends from *filĭu*, the Latin word for the same concept, while *filial* (same as the English word *filial*) is an adaptation of *filialis*, meaning "befitting a son" or "filial" (Diamond 1961, 58; *DCECH,* 359; Peers et al. 1960, 424, 475). The English definition of *filial* from the *OED* is "of or pertaining to a son or daughter."[5] *Filial* was in use by then, meaning "that which pertains to the son, like filial love" (de Covarrubias 1943, 595). We can see the sound change *f* to *h* taking place as we find examples of *fijo* (1100 AD) and *hijo* (1062) occurring around the same time (Corominas 1967).

Hijo came to its present form in this manner:

filĭu ——> *fiʎo*—*l* palatalized before a glide (the *ĭ*) (and the same *u* to *o* change)
fiʎo ——> *fiyo*—*ʎ* becomes the semivowel *j*
fiyo ——> *hiyo*—this positioning is uncertain
hiyo ——> *hiʒo*—in the twelfth century, *y* goes to *ʒ*
hiʒo ——> *hiʃo*—*zh* sound goes to a *sh*
hiʃo ——> *hixo*—in the sixteenth century, the *sh* becomes the velar fricative
hixo——> *hijo*—no sound change, just a change in orthography (Resnick 1981, 39)

humear, fumigar

The Latin word *fūmĭgāre* (to smoke) has given Spanish the words *humear* and *fumigar* (Diamond 1961; *DCECH*). *Humear* has descended from the original Latin and means basically the same thing: "to smoke, emit smoke, fumes, or vapors." *Fumigar* was borrowed later from Latin and is similar to the English *fumigate*: "to fumigate, smoke, fume, purify, or mediate by vapors" (Peers et al. 1960, 483, 437). De Covarrubias does not mention either of these words but does define *humo,* "smoke" (noun), the same way we do today. He also notes that *perfumar* and *perfume* come from the same root (1943).[6] *Humear* first appeared in literature in the mid-thirteenth century (*humo* appeared in 1088). *Fumigar* came much later, in 1817, which explains why de Covarrubias makes no mention of it (*DCECH*).

Here is the transformation from *fūmĭgāre* to *humear*:

fūmĭgāre ——> *fūmĭγāre*—between vowels, *g* (like other voiced stops) is weakened to γ, a voiced velar fricative
fūmĭγāre ——> *fūmĭāre*—γ drops out completely (Harris-Northall 1990, 7–10).
fūmĭāre ——> *fumeare*—*ĭ* > *e* in unstressed position, other vowels level
fumeare ——> *humeare*—*f* to *h*
humeare ——> *humear*—final *e* is dropped (Penny 1991, 96–97).

hembra, feminino

This is another example where the two words do not come from exactly the same word but from the same root. *Hembra* is from Latin *fēmĭna* (a female, a woman), and *femenino* is from the adjective *fēmĭnīnus* (feminine) (Diamond 1961). Current Spanish defines *hembra* as a "female animal or plant" and only vulgarly as a woman (Peers et al. 1960). In the seventeenth century, the word was commonly applied to any type of female, human, animal, or vegetable (de Covarrubias 1943). *Femenino* simply means "feminine" in the present day (Peers et al. 1960), but de Covarrubias has no earlier definition of it. *Hembra,* which has been used throughout the centuries since Latin, was first recorded in its present form in the late twelfth century. We do not see *femenino* until 1438.

fēmĭna ——> *femna*—vowels in syllables next to stressed vowels (the *e* is stressed in this case) that are adjacent to *r* or *l,* or sometimes *s* or *n,* tend to drop out; this does not happen to vowels that begin or end words

femna ——> *femra*—dissimilation of *m* and *n*

femra ——> *fembra*—epenthesis (adding a consonant) to break up a difficult cluster (Resnick 1981, 71–72)

fembra ——> *hembra*—*f* to *h*

If *femenino* had descended from Latin and not been borrowed later it would not have retained the *i* between the *m* and the *n*, which itself causes a lot more changes. The changes that have occurred in *femenino* can be explained thus:

fēmĭnīnus ——> *femeninus*—*ĭ* goes to *e* in Vulgar Latin

femeninus ——> *femenino*—borrowings tend to modify endings to fit the language (Penny 1991, 96, 210).

huir, fugitivo

Huir, "to flee, to escape" descends from Latin *fŭgĕre,* meaning "to flee, to flee away from" (Peers et al. 1960, 483; *DCECH,* 422; Diamond 1961, 61). *Fugitivo,* "fugitive, runaway," and other words such as *fugaz* (fleeting, volatile) and *refugio* (refuge) all come from the same root as *fŭgĕre*: *fugio* (Peers et al. 1960, 437; Llauró Padrosa 1957, 263). *Fugio* is also the root of the musical term *fugue,* which in Italian means "flight," alluding to the feeling of the music (*OED* 2001). The specific word borrowed into Spanish for *fugitivo* is *fugitīvus* (Llauró Padrosa 1957, 263). *Huir* has been elevated a bit since the time of de Covarrubias. Roughly translated, he says *huir* "commonly denotes cowardliness in the military, unless one is trying to trick the enemy or the enemy is so powerful he has no chance. Even in these situations many will stay, preferring to die in battle than to flee" (1943, 704). Now "fleeing" has a more general sense; it may be done for both good and cowardly reasons. In the seventeenth century, *fugitivo* had a narrower definition, as it generally referred to an escaped slave (who could be found because he was walking in irons). *Fugaz,* which used to just mean something with the condition to flee, like a rabbit, came to mean "fleeting" when it was applied poetically to things like time and ages (de Covarrubias 1943). *Huir* was recorded with an initial *f* as early as 1054 and with an initial *h* in 1490. *Fugaz* was recorded in literature in 1580 (Corominas 1967).

How to get *huir* from *fŭgĕre*:

fŭgĕre ——> *fugīre*—documented Vulgar Latin (*DCECH,* 614)

fugīre ——> *fuyire*—*g* before *e* or *i* goes to *y*

fuyire ——> *fuire*—*y* before *e* or *i* drops out (Resnick 1981, 67)

fuire ——> *fuir*—final *e* drops off (Penny 1991, 97)

fuir ——> *huir*—*f* to *h*

The presence of the *g* in words like *fugitivo* and *refugio* shows they were borrowed at a later time.

habla, fábula

Both of these words are directly related to the Latin *fābula,* "conversation; story without a guaranteed history." They are related to *hablar* and *fabular,* which descend from *fābulare,* which means "to talk, to converse" (*DCECH,* 296; Diamond 1961, 56). A *fábula* is a "fable, legend, fiction, storytale," which is one part of the Latin *fābula,* while *habla* is "speech, language, idiom, dialect," the other part of the definition. In addition to being consistent with what we think of as a fable, in the seventeenth century, *fábula* also meant "the rumor and talk of the town," or the gossip it seems. It was "something without foundation" (de Covarrubias 1943, 579). This gives new insight on what Paul means when he says, "And they shall turn away their ears from the truth, and shall be turned unto fables" (2 Tim. 4:4). *Hablar,* which used to be a transitive word, was first used in literature in 1492. We can first see *fábula* in a book from 1438 AD (*DCECH*).

fábula ——> *fabla*—*u* is deleted for the same reason as the *i* in *femina*: it is near a stressed syllable and next to an *l*
fabla ———> *habla*—*f* to *h* (Penny 1991, 96)

The presence of the *u* in *fábula* shows us that it is a learnèd word.

CONCLUSION

Although there is some disagreement among scholars as to the source of words that preserve an *f* at the beginning despite the rule that *f* went to *h,* evidence is available that these words are a result of borrowing. Words that preserve the *f* are generally cultural words, which are more likely to be borrowed, and those which have undergone the sound change are mostly basic vocabulary. Also, words beginning with *f* have not undergone any of the other sound changes in Spanish, showing they were introduced into the language after these changes had occurred. Further research could apply these same methods to other Spanish sound changes to see whether their exceptions are also a result of borrowing.

NOTES

1. *f* could not have gone to *h* in the times of Vulgar Latin because that is when *e* > *ie* and *o* > *ue*. This change to diphthongs had to happen before *f* went to *h* or else words with diphthongs today would have experienced the sound change as well (Menéndez Pidal 1956, 200).

2. The *Dictionary of Liturgical Latin* (Diamond 1961) shows an even greater percentage of words with initial *f,* since many more words had been borrowed by the time of its printing. The Fernández de Santaella dictionary is used to give definitions closer to what would have been in use when the sound change was completed.

3. All seventeenth century definitions are from De Covarrubias, *Tesoro de la Lengua Castellana o Española,* which was originally written in 1611.

4. The transcriptions of the sound changes are done mostly according to Spanish orthography. For example, *y* is used for a palatal approximant, which is *j* in IPA (the International Phonetic Alphabet). IPA is used when there is no letter available, namely, *ʎ* for a palatal lateral approximant, *ʒ* for a *zh* sound, and *ʃ* for an *sh* sound.

5. Although his older definition of *hijo* is the same as it is currently, de Covarrubias gives us some interesting insight into his view of what a son really is, calling a child "what binds the love of its parents, as both agree to love it" (1943, 689).

6. The *OED* explains the development of this concept in English: originally *perfume* meant pleasant odorous fumes given off by burning something such as incense. The word eventually came to mean any odor emitted in particle form by a sweet-smelling substance.

REFERENCES

Blake, Robert J. (1988). Sound change and linguistic residue: The case of [f-] > [h-] > [Ø]. In *Georgetown University round table on languages and linguistics.* 53–62. Washington, D.C.: Georgetown University Press.

The Church of Jesus Christ of Latter-Day Saints. (1990). *Holy Bible.* Salt Lake City, UT.

Corominas, Joan. (1967). *Breve diccionario etimológico de la lengua Castellana—Segunda edición.* Madrid: Editorial Gredos.

Corominas, Joan, and José A. Pascual. (1980). *Diccionario crítico etimológico Castellano e Hispánico* (Vols. I & III). Madrid: Editorial Gredos.

Crowley, Terry. (1997). *An introduction to historical linguistics*. 3rd edition. Auckland, New Zealand: Oxford University Press.

De Covarrubias, Sebastián. (1943). *Tesoro de la lengua Castellana o Española: Según la impresión de 1611, con las adiciones de Benito Remigio Noydens publicadas en la de 1674*. Edición preparada por Martín de Riquer. Barcelona: S. A. Horta.

Diamond, Wilfrid. (1961). *Dictionary of liturgical Latin*. Milwaukee: Bruce Publishing Company.

Fernández de Santaella, Rodrigo. (1744). *Vocabularium seu lexicon ecclesiasticum, Latino-Hispanicum: Ex Sacris Bibliis, conciliis, pontificum, ac theologorum decretis, divorum vitis: Dictionariis, aliisque probatissimis scriptoribus, concinnatum*. Ed. Joanne De Lama Cubero.

Harris-Northall, Raymond. (1990). *Weakening processes in the history of Spanish consonants*. London and New York: Routledge.

Iribarren-Argaiz, Mary C. (1998). The influence of Basque on phonological changes in Castilian Spanish: A case of languages in contact. *Romance Quarterly*, 45 (1): 3–34.

Levy, John F. (1973). Tendential transfer of Old Spanish *HEDO* < **FOEDU** to the family of *HEDER* < **FOETĒRE.** *Romance Philology*, 27 (2): 204–10.

Llauró Padrosa, Juan. (1957). *Diccionario manual morfológico Latino-Español*. 2nd edition. Barcelona: Librería Elite.

Menéndez Pidal, Ramón. (1956). *Orígenes del Español: Estado lingüístico de la Península Ibérica hasta el siglo XI*. 4th edition. Madrid: Espasa-Calpe, S. A.

Oxford University Press. (2001). *Oxford English dictionary on-line*. Available: http://dictionary.oed.com/

Patterson, William T. (1982). *The genealogical structure of Spanish*. Washington, D. C.: University Press of America.

Peers, Edgar A., José V. Barragán, Francesco A. Vinyals, and Jorge A. Mora. (1960). *Cassell's Spanish dictionary*. New York: Funk and Wagnalls.

Penny, Ralph J. (1991). *A history of the Spanish language*. Cambridge: Cambridge University Press.

Resnick, Melvyn C. (1981). *Introducción a la historia de la lengua Española*. Washington, D.C.: Georgetown University Press.

Spaulding, Robert K. (1962). *How Spanish grew*. Berkeley: University of California Press.

Travlang. (1995–2000). *Travlang's English-French on-line dictionary [on-line]*. Available: http://dictionaries.travlang.com/EnglishFrench/

Improving Speech Recognition for a Communicative CALL Task

Michael Emonts, Deryle Lonsdale, C. Ray Graham, Michael Rushforth, PSST! Research Group, Brigham Young University

When speech application developers produce communicative activities for computer-assisted language learning (CALL), they often experience difficulties in getting the off-the-shelf programs to recognize learner responses with a high degree of accuracy. This paper describes the measures we have taken to improve recognition rate with the OGI toolkit and chronicles the improvement based on each change made. Changes include (1) limiting the universe of possible (probable) responses by making the task more explicit; (2) applying linguistic knowledge to the altering of word pronunciations; (3) converting .wav files to other file types; (4) altering the recognizer used for speech recognition.

Need for Interaction in Language Learning

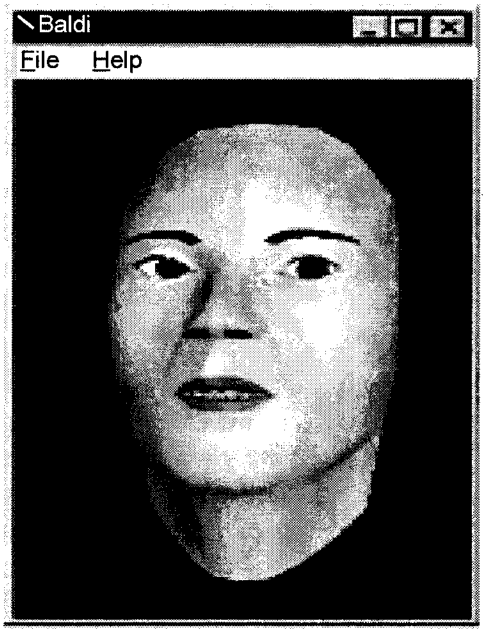
Language acquisition research has strongly suggested that there are two major learner activities that contribute to second language acquisition: receiving comprehended input and producing comprehensible output (for summaries of this research see Ellis 1994, 273–84; Lightbown and Spada 1999, 39; and Gass and Selinker 1994, 200–201, 276–78). For many years now, computers with interactive multimedia capabilities have provided an excellent source of comprehensible input for language learners. And interactive exercises provided through the computer interface have provided a mechanism through which comprehension could be verified, thus assuring that the material was not only comprehensible but also comprehended by the particular learner. The computer interface has also provided a number of activities through which learners can practice speaking skills in a rather mechanical way.

CALL, however, has lacked the capacity to provide opportunities for the learner to practice what language acquisition specialists have called comprehensible or forced output—that is, oral output that is comprehended and responded to appropriately. This inability of computers to respond appropriately to oral output from learners has been a major drawback to CALL (Egan 1999, 280–81; Harless, Zier, and Duncan 1999, 313–14).

Why Use the OGI Toolkit

The CSLU/OGI toolkit[1] best suited our needs in designing interactive language learning tasks. The OGI toolkit provides a user-friendly graphical interface with high-level functionality that enables users to easily design dialog scenarios. Along with the toolkit comes Baldi, an agent featuring modern facial animation technologies that are articulatorily correct (see Figure 1). The use of Baldi provides learners an opportunity to interact with a visible agent, thus making the overall interaction more realistic and meaningful. The fact that it is freely distributable and has a well-documented website further adds to the benefits of using the OGI toolkit.

Figure 1. Baldi



Improving Speech Recognition Accuracy

The "directions" activity was chosen as the task in which to improve the system. In this activity, a map is presented showing the location of the student's house and the current location of the student's friend who is trying to visit (see Figure 2). The learner's goal is to provide step-by-step directions that will lead the friend to the student's house. When a correct direction is given, the friend then follows that direction, thus coming closer to the student's house (see Figure 3).

In order to improve the recognition accuracy for this activity, a program needed to be created that would enable us to quantify the accuracy of each speaker's utterances and to eliminate the need for extensive testing. To do so, we collected samples of correct directions from various adult speakers and stored them in .wav format. These files were then used as examples of input to the system that should be recognized, thus speeding up the testing process. Recognition accuracy was then quantified by simply determining what percentage of utterances were recognized correctly, as specified by a phrase-structure grammar. Although the corpus collected was relatively small, it provided a benchmark upon which we could progress. The program that was developed for testing accuracy allowed us to isolate factors that lead to a poor recognition rate.

Figure 2. Directions need to be given to lead the student's friend to the correct house

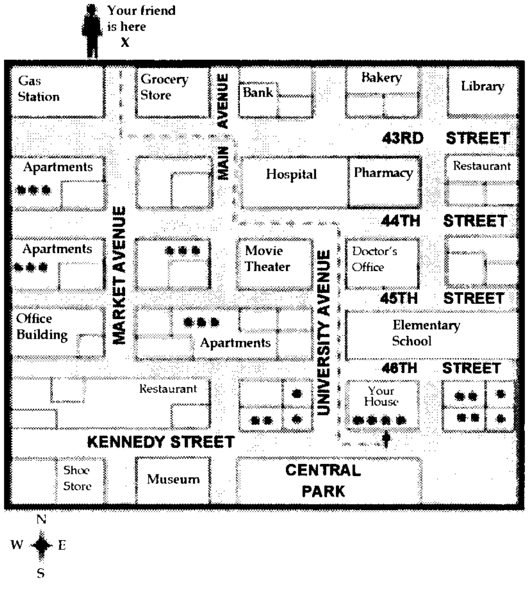


Figure 3. As correct directions are given, progress is shown

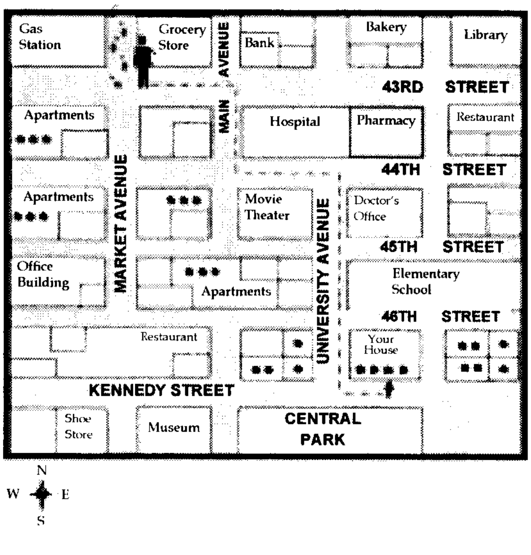


Table 1. Effect on accuracy by various recognition system input parameters

Test	Number Correct	Total Number of Files	Percent Correct
Grammar			
Controlled grammar	24	24	100%
One vs. two	13	24	54%
Go vs. walk	6	24	25%
North, south, east, west	19	24	79%
North, south, east, west	33	100	33%
Optional [kh] in "walk"	38	100	38%
Optional [kh] in "block"	34	100	34%
Recognizers(N,S,E,W)			
Adult_8kHz_0	33	100	33%
Adult_8kHz_0	39	100	39%
Child_16kHz_0	77	100	77%
Child_16kHz_1	92	100	92%
Adult_16kHz_0	97	100	97%
Grammar			
One vs. two	94	100	94%
One through five	86	100	86%
File types			
LINEAR	86	100	86%
RIFF	86	100	86%
NIST	92	100	92%
Grammar			
Walk vs. go	93	100	93%
Flexible syntax	91	100	91%
Optional [kh] in "walk"	93	100	93%

The first step was to simplify the grammar to the extent that it consisted only of the correct direction (i.e., "go south one block"). It is no surprise that 100% accuracy was obtained with a deterministic grammar. From this point, however, any change made to the grammar would lower accuracy percentage and allow us to observe the sources of recognition error. Table 1 shows the changes made to the controlled grammar and their influence on recognition rate. Introducing the number "two" to the grammar, for instance, made the system have to determine whether the input utterance was "go south one block" or "go south two blocks." Unfortunately, this small change in grammar reduced the response accuracy (i.e., the system's ability to correctly understand the speaker's utterance) to only 54%. Introducing "walk" to the controlled grammar reduced the accuracy to only 25%. Allowing the grammar to determine which direction was spoken (i.e., north, south, east, or west) was not as troublesome, yielding 79% accuracy.

At this point, it became clear that a larger number of test files was required to ensure reliability of testing results. The number of files tested was increased to 100, and surprisingly, the above grammar distinguishing the four directions only yielded 33% accuracy.

Because each word in the grammar was recognized using the default pronunciations that came with the toolkit, changes made to the pronunciations appeared to marginally improve results. For example, the representation of "walk" in Worldbet, the transcription scheme used in the toolkit, is {w > kc kh}. When a speaker says, "walk two blocks south," however, the "k" is rarely aspirated as indicated by the "kh" in the toolkit's pronunciation (cf. Ladefoged 1993, 49–55). Making the aspiration on the "k" optional (i.e., {w > kc [kh]}) improved recognition accuracy from 33% to 38%. The inclusion of optional aspiration in the "k" of "block," however, damaged recognition rate, lowering accuracy from 38% to 34%. This is clearly due to the fact that stops at the ends of utterances are typically released.

Using the basic grammar that included only the four directions (33% accuracy), changes were made by replacing the default recognizer (i.e., Adult_8kHz_0, an 8kHz sample adult model) with other recognizers provided in the toolkit. As Table 1 illustrates, huge differences in recognition rate were exhibited, depending on the recognizer being used. Recognizers that utilized a 16kHz sampling rate exhibited the best recognition accuracy, with the adult version leading the way with a 97% recognition rate.

Although the accuracy was reasonably high, it must be remembered that the grammar was still very simple. The recognizer at this point only has to choose among four possibilities (i.e., "go north one block," "go east one block," "go south one block," and "go west one block"). Further complicating the grammar by adding the flexibility of numbers one through five lowers the accuracy to 86% but greatly increases the grammar's complexity.

The choice of file formats for the input .wav file also affects recognition. Converting the file type from the default LINEAR format to NIST, for example, raises the overall recognition rate from 86% to 92%.

Adding a flexible syntax rule to the grammar greatly increased the complexity of the grammar yet only marginally affected recognition accuracy. The flexible syntax rule allows the user to speak in either of two correct syntaxes (i.e., "go south one block" and "go one block south"). The inclusion of this syntactic rule only weakened the recognition rate to 91%.

Figure 4. Final grammar that distinguishes eighty possible directions

```
set G1 {
        {"instruction"
        "$dir = south | east | west | north;
        $dist = one | two | three | four | five;
        $displace = go | walk;
        $x = [*sil%% | *any%%] $displace
[*sil%%] $dist [*sil%%] $block [*sil%%] $dir [*sil%%
| *any%%];
        $y = [*sil%% | *any%%] $displace
[*sil%%] $dir [*sil%%] $dist [*sil%%] $block [*sil%%
| *any%%];
        $path = ($x | $y);"}
}
```

After the optional aspirated "k" was added to the pronunciation model for "walk," we attained a final accuracy rate of 93% for a reasonably complex grammar. Following all of these alterations, the grammar (see Figure 4) can distinguish among the four directions, five possible numbers of blocks, two ways of traveling (walk and go), and two syntactic methods (described above). The recognizer is now able to use this grammar to determine whether the utterance is one of the few correct directions among the eighty different possibilities.

Possible Areas of Future Research

Although 93% accuracy is a vast improvement, an error rate of 7% is very serious in language pedagogy. Having correct responses rejected 7% of the time may be a serious detriment to learning and provide much discouragement (Mostow and Aist 1999, 415–16). In order to further increase the recognition rate of this activity, we make the following suggestions.

First, it was found that many errors were caused by a poor recording environment. Problematic .wav files were often saturated with loud background noise or even truncated so that they didn't contain the entire utterance. A better recording environment or better recording equipment (or both) would probably increase accuracy. Second, the recognition rate can be enhanced further by finding the optimal settings of various internal parameters, such as penalties and rejection medians. Third, developing a new recognizer specifically created for the "directions" activity would likely reduce the error rate immensely. To do so, the neural network would need to be retrained on a large corpus of applicable utterances. Furthermore, since the activity is geared toward learners of English, developing a recognizer specifically trained on foreign-accented English would also prove beneficial.

Conclusion

To summarize, we have successfully improved the recognition rate of our "directions" activity to an accuracy level of 93%. We have found that by altering the task design, simplifying the grammar, changing the speech recognizer, converting the format of the .wav files, and altering the word pronunciations, recognition rate may be increased substantially.

Notes

1. See the website for the Center for Spoken Language Understanding at the Oregon Graduate Institute (www.cslu.ogi.edu/toolkit).

References

Egan, Kathleen B. 1999. Speaking: A critical skill and a challenge. *CALICO Journal* 16(3): 277–93.

Ellis, Rod. 1994. *The study of second language acquisition.* Oxford: Oxford University Press.

Gass, Susan M., and Larry Selinker. 1994. *Second language acquisition: An introductory course.* Hillsdale, NJ: Erlbaum.

Harless, William G., Marcia A. Zier, and Robert C. Duncan. 1999. Virtual dialogues with native speakers: The evaluation of an interactive multimedia method. *CALICO Journal* 16(3): 313–37.

Ladefoged, Peter. 1993. *A course in phonetics,* 3rd ed. Fort Worth: Harcourt Brace Jovanovich College Publishers.

Lightbown, Patsy M., and Nina Spada. 1999. *How languages are learned.* Oxford: Oxford University Press.

Mostow, Jack, and Gregory Aist. 1999. Giving help and praise in a reading tutor with imperfect listening—Because automated speech recognition means never being able to say you're certain. *CALICO Journal* 16(3): 407–24.

An NLP System for Extracting and Representing Knowledge from Abbreviated Text

Deryle Lonsdale, Merrill Hutchison, Tim Richards, William Taysom

This paper presents a new natural language processing (NLP) system called LG-Soar. The system is based at its most fundamental level on the Soar cognitive modeling intelligent agent architecture (Newell 1990). The system represents the integration of three major processing components: (1) regular-expression-based text preprocessing; (2) the Link Grammar parser; and (3) the Soar intelligent agent architecture. The result is a robust, versatile text processing engine useful for difficult-to-handle input. Unlike a related Soar-based NLP system, NL-Soar (Lewis 1993), this new system is not specifically designed for cognitive modeling of natural language use.

The project addresses several interesting challenges from an NLP perspective. The overall goal was to mine content from problematic text. Most existing systems perform well only on well-structured, completely grammatical text. Another goal was to address complicated linguistic issues in the development of a usable system. We also sought to output the information into a variety of usable formats. Finally, the project was meant to test the feasibility of integrating this particular set of components within a unified agent architecture.

The system was designed to handle the parsing of genealogical information from a file containing profiles of several thousand colonial American individuals. This well-known resource, built from Savage's dictionary (1860–1862), was previously scanned via OCR and placed in raw-text form on the internet.

The LG-Soar system operates as follows:

1. A genealogical entry is read in from a preprocessed input file.
2. Each entry is split into individual sentences.
3. Each sentence is parsed with the Link Grammar parser.
4. The discourse representation module creates semantic/discourse representations (based on parse contents) for all sentences in the entry.
5. Output is generated according to various formats.

These steps are discussed in further detail in the rest of the paper.

Preprocessing

The preprocessing stage of LG-Soar uses a collection of subroutines and regular expressions in the Perl scripting language to create machine-readable entries for the parser. Duties of the preprocessor include creating and numbering entries for the individuals found in the input text and reformatting information about those individuals into tokenized, plain-text sentences that can be parsed by LG-Soar.

The input file text is difficult to deal with in its original form. A number of abbreviations used in the text were meant to represent several different words and require analysis of the context to correctly substitute the right word for each abbreviated form. A further complication is manifest in tokenizing individual sentences. Care must be taken not to truncate the sentence because of an abbreviation. Other problems include place names or words that, after appearing once, are abbreviated later on in the text; an incomplete list of substitutions for abbreviations; and occasional corpus errors. The text itself, although difficult to process, is structured well enough to allow a basic level of automated information extraction. Text similar to the genealogical information used in this project is found on the internet, and the tools used in this project could be adapted for processing this type of semistructured data.

The preprocessor creates an entry that consists of a surname, a given name, and information about the individual. Surnames always appear in full capitals and head the paragraph of information about a family. Individuals belonging to the family of that surname also always appear in full capitals. Information about an individual is parsed and appended to the entry until a new individual is encountered.

Perl was chosen to implement the preprocessor because of its built-in functions for pattern matching and text manipulation. The first step in building the preprocessor was to analyze the input text. A keyword-in-context (KWIC) browser was used to determine the context for a particular interpretation of an abbreviation. Each context is represented as a quoted string. The short string is a type of simple regular expression. Underscores stand for the abbreviation. Complex Perl regular expressions are represented as Perl scalar variables. Parentheses indicate optionality. These contexts are combined into a single line with the word that will replace the abbreviation and are later extrapolated into one large regular expression as the preprocessor is started.

Prior to running the preparser, lists of abbreviations and their interpretations, complex regular expressions, and common words must be created. Many of the abbreviations and their substitutions are listed in the information included with the original file (which is subsequently removed from the input text). Unlisted abbreviations are determined by analyzing the corpus. The abbreviations are used as the key that indexes a concatenated string of possible interpretations. Complex regular expressions match a complex type of data, such as a date or occupation, and are built by hand using the KWIC browser. These serve to increase the readability of the context strings because the complexity can be hidden in simple variables that nest the larger, more difficult expressions. The Unix command "grep" is used to create a file of words that appear capitalized. This file is used to create another file that includes any uncapitalized word that matches a word from the capitalized word file.

Sentence boundaries are determined with a simple heuristic: Look at the words preceding and following a period. The sentence ends if the word before the period isn't an abbreviation or the word following the abbreviation is a common word. This heuristic is successful since sentences ending in an abbreviation are uncommon. Abbreviations are processed as the sentence is being concatenated and tokenized. The preparser looks up the abbreviations in the index to find the string of concatenated interpretations. The string is split into individual interpretations that serve as a key to locate the extrapolated context regular expression. The underscores are replaced with the actual abbreviation in the original expression and matched against the line of input. If it matches, the word is replaced. A sample entry and the result of its preprocessing are shown in Figure 1.

Future work on the preprocessor may include techniques to further simplify later stages of processing within the LG-Soar system by splitting long sentences joined by conjunctions into small sentences and using pattern matching to explicitly replace the subject in sentences that make use of anaphora.

THE SOAR FRAMEWORK

Soar is a computer system that models human cognition (Newell 1990). It is architecturally predisposed to goal-directed problem solving and thus is ideally suited to complex tasks. Implemented in an agent-based framework, it is ideal for web search and similar applications. Its overall design derives from the fact that it was meant to instantiate a unified theory of cognition. More details can be found in the relevant literature. Soar has already been used very successfully in a diverse array of applications.

One of the motivations for using Soar in this project is that NL-Soar has been used successfully for representing and tracking referents in discourse. Because the goal of this work was not to model human cognition directly, it was deemed more appropriate to develop LG-Soar, a new system tailored to information-extraction and data-integration tasks.

LG-Soar processing requires fairly clean (if not completely grammatical) textual input. For the purposes of this paper, it can be assumed that the input to the system is preprocessed text as described previously. The output from the parser part of the system is some representation of structure that will allow for the next stage of processing. Note that the usual NLP parser output representations, tree structures, are not always conducive to further processing; they are often cumbersome and inadequate.

Figure 1. Sample raw text from Savage and its preprocessed counterpart[1]

EATON, THOMAS, Haverhill, m. at Andover 6 Jan. 1659, Unice Singletary of Salisbury; freem. 1666, was k. by the Ind. 15 Mar. 1698.

⇩

Thomas Eaton, married at Andover 6 January 1659,
Unice Singletary of Salisbury.*****
Was freeman 1666.******
Was killed by the Indians 15 March 1698. *****

Note the expansion of abbreviations, determination of sentence boundaries, and canonicalization of the name.

There are several reasons why it was decided to use the Link Grammar parser for this application. First, the parser is freely available for research purposes. Second, it is robust and can handle a much larger range of grammatical, semigrammatical, and ungrammatical structures than traditional parsers can gracefully without failing. Third, the system builds explicit relations suitable for the next stage of processing. The system also runs quite fast, compared to traditional parsers; this is a consideration when handling large volumes of data. It is also written in the C programming language, which facilitates integration with the Soar system. Finally, the LG approach yields a linguistic description that is more appropriate for the task than traditional phrase-structure grammars can provide.

The LG-Soar system was constructed by integrating two systems: Soar and the Link Grammar parser. This was possible since Soar and LG both use C at their lowest levels. In addition, Soar supports Tcl, the Toolkit Command Language, which is used to integrate various computer architectures. Tcl thus acts as "glue" between the Link Grammar engine and the Soar engine. A nontrivial amount of C and Tcl code was therefore written to tie the two systems together. The result is LG-Soar, a system that includes Tcl commands for calling the Link Grammar functions and passing information into the basic Soar processor.

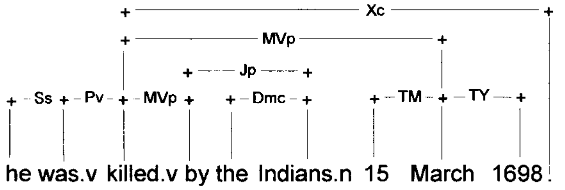
THE LINK GRAMMAR COMPONENT

The Link Grammar parser is a system designed to permit flexible and robust parsing of natural-language text. It produces a shallow parse, meaning that it does not aim at a complete, theory-dependent, linguistic parse of the sentence with all of its morphological, syntactic, and semantic complexity. Rather, it seeks to describe the major components of a sentence in as simple terms as possible. The basic unit of structure is the link.

Each sentence consists of links, and each link connects two words. Links are of various types and correspond loosely to functional relationships, like associating a subject with its predicate, a verb or preposition with its object(s), an auxiliary verb with its main verb, and so on. A link label specifies the type of relationship between the words at each end of the link. Potential links are specified by highly technical rules. In addition, it is possible to assign a score to overall linkages and also to penalize individual links.

Figure 2 shows a sample link parse for a sentence from an entry in the Savage text.

Figure 2. Sample link parse



For example, here the subject and verb are linked via an Ss (singular subject) link, a determiner and its head noun are associated via a Dmc (determiner) link, and the month and associated year are associated via a TY (time/year) link.

A couple of sample LG rule entries appear in Figure 3.

It is clearly beyond the scope of this paper to describe how the parser works and how the grammar knowledge is developed; however, a few points can be made concerning what types of linguistic knowledge had to be added to handle the Savage text entries. For example, the basic parser only recognizes one month/day order (May 24), whereas Savage uses formats like "24 May." Similarly, it only recognized years after 1900; this had to be extended back several centuries. It was also necessary to allow years to postmodify verbs, even without prepositions (e.g., "died April 1655"). Savage also rather idiosyncratically inserted a comma between arguments in verb frames (e.g., "He married 6 July 1694, Ann Lynde."); constructions like this had to be allowed for. The basic system also recognized dates as direct objects and as comma-introduced appositives, as in constructions like "He died of smallpox, 24 October 1678." By penalizing such links, the problem was corrected. Savage also used telegraphic-style prose, such as allowing singular nouns without determiners: "He was son of Thomas." Rules were added to the grammar to permit these kinds of constructions. Finally, several domain-specific words (e.g., "freeman") had to be added to the system's general-purpose lexicon.

Figure 3. Sample LG rule entries

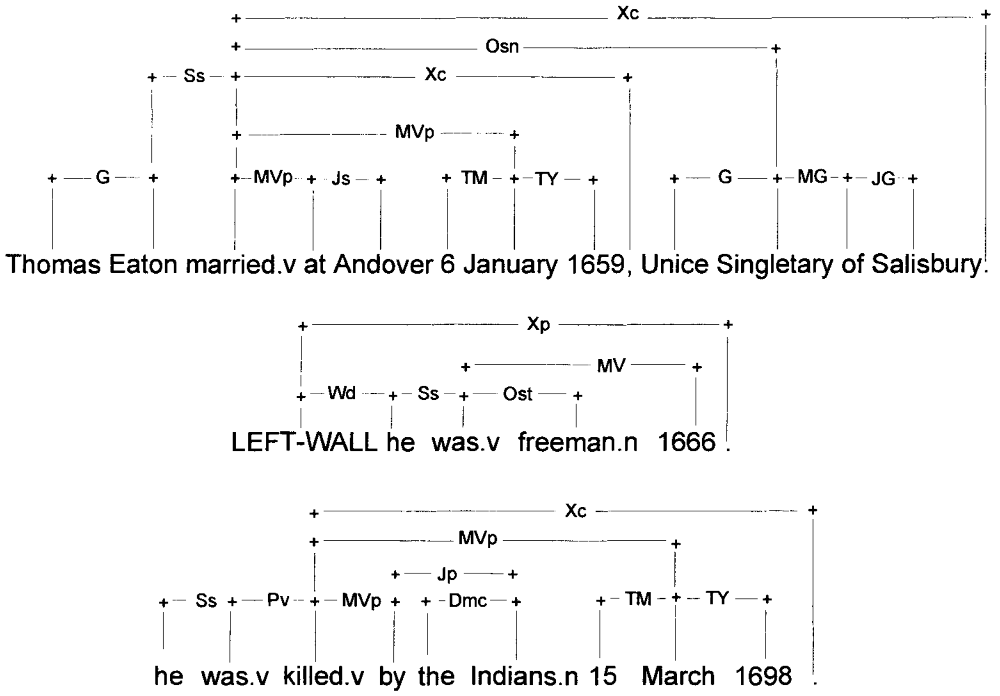
```
words/words.y: % year numbers
NN+ or NIa- or AN+ or MV- or ((Xd- & TY- & Xc+) or TY-)
or ({EN- or NIc-} & (ND+ or OD- or ({{@L+} & DD-} &
([[Dmcn+]] or ((<noun-sub-xnoappositive> or TA-) & (JT- or IN-
or <noun-main-xnoyear>))))));

<vc-fill>: ((K+ & {[[@MV+]]} & O*n+) or ({O+ or B-} & {K+}) or
[[@MV+ & {Xc+} & O*n+]]) & {Xc+} & {@MV+};
```

The result is an extremely robust parser that has been enhanced with the linguistic knowledge necessary to handle genealogical and biographical text. Whereas the text would cause severe difficulty for conventional parsers, LG-Soar was able to deal with it very satisfactorily. Figure 4 presents a couple of examples.

Semantic Processing

After the text has been preprocessed and the syntactic links generated, semantic processing takes over. This involves

Figure 4. More sample link parses


Xc
Osn
Ss
Xc
MVp
G
MVp
Js
TM
TY
G
MG
JG
Thomas Eaton married.v at Andover 6 January 1659, Unice Singletary of Salisbury.
Xp
MV
Wd
Ss
Ost
LEFT-WALL he was.v freeman.n 1666 .
Xc
MVp
Jp
Ss
Pv
MVp
Dmc
TM
TY
he was.v killed.v by the Indians.n 15 March 1698 .

the translation of the Link Grammar parse into a representation of discourse objects and their anaphoric relationships. A subset of the linguistic approach called Discourse Representation Theory (DRT) was chosen as the basis for the representation (Kamp & Reyle 1993). A series of intermediate representations of semantic content is followed until the appropriate output format is generated by the system. This section discusses semantic processing and its associated representations.

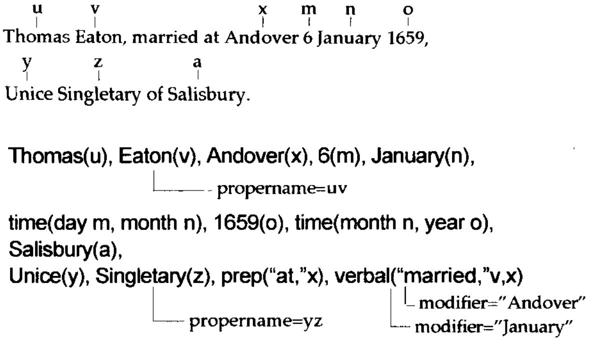
LG-Soar implements a series of three translations between intermediate semantic representations:

1. converting a syntactic link parse to a protoDRS
2. converting a protoDRS to a DRS
3. converting a DRS to user-directed output formats

The representation of a particular set of semantic entities and relationships is called a Discourse Representation Structure (DRS). A DRS is designed as a simple and easily visualizable means of specifying the content of discourse in the context of its predication, in a manner akin to first-order logic. The approach also places a great deal of emphasis on determining pronoun reference. Any DRS has two kinds of elements: discourse referents and conditions. Discourse referents function basically like variables in logic, and conditions function as predicates over the discourse referents. For example, in the sentence "He was killed by the Indians 15 March 1698," "he" might be assigned the discourse referent *u*, and "the Indians" assigned *v*. Then conditions placed on *u* and *v* might involve the fact that *u* is represented in the existing framework by a masculine third person singular pronoun, that *v* refers to some Indians, and finally that *u* "was killed by" *v* (or, if the passive voice is disregarded, that *v* killed *u*). At present the system only implements the most basic features of DRT; however, additional features and constructions can easily be added to the existing framework.

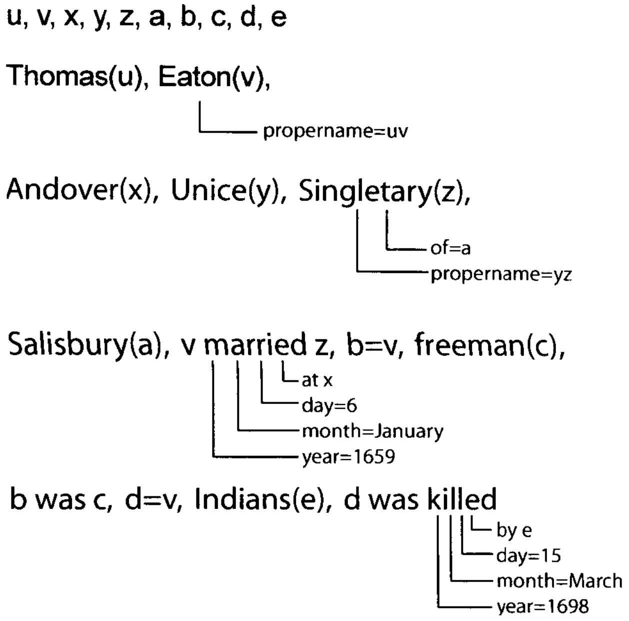
Although the DRS is the principal semantic representation built by LG-Soar, its creation is preceded by that of a protoDRS. The protoDRS derives its information from the Link Grammar parse links. It has discourse referents as arguments in conditions and includes pointers to words in the parsed sentence as arguments. For example, the verbal condition in the sentence "John worked in a factory" will have as arguments the discourse referent associated with John, the word "worked" identified as the verb, and the word "in" identified as the introduction to a modifying phrase. The LG-parse-to-protoDRS translation depends entirely on the structure of the Link Grammar parse. First, each link triggers the construction of discourse referents and conditions. After these have been initialized, relationships between them are established. For example, the "S" link connects the main verb of a sentence to its grammatical subject. The link triggers the construction of a discourse referent for the subject and a verbal condition for the relevant verb. Inferred relationships establish that the discourse referent is the subject of the verb phrase. An example of a protoDRS for a sentence is given in Figure 5.

Figure 5. Sample protoDRS


u v x m n o
Thomas Eaton, married at Andover 6 January 1659,
y z a
Unice Singletary of Salisbury.
Thomas(u), Eaton(v), Andover(x), 6(m), January(n),
propername=uv
time(day m, month n), 1659(o), time(month n, year o),
Salisbury(a),
Unice(y), Singletary(z), prep("at,"x), verbal("married,"v,x)
modifier="Andover"
propername=yz
modifier="January"

Once a protoDRS has been built, a DRS can then be created from it. For the most part, the protoDRS-to-DRS transition involves transferring conditions from the protoDRS while removing word pointers, and also formulating relationships with complex conditions in the DRS. During this stage of processing, anaphoric relations are also determined. The rules for the construction of the DRS from the protoDRS make use of some knowledge beyond that which is expressed in the Link Grammar parse. Possessive pronouns are a good example. The Link Grammar parser simply treats possessive pronouns as determiners linked to the noun which they modify. In the protoDRS the relationship is represented as a condition that marks the pronoun as a determiner and the noun as its argument. During the protoDRS-to-DRS transition, the determiner is checked against a list of possessive pronouns. If the determiner is a possessive pronoun a "pos-s" condition is added to the protoDRS. This new condition marks the possessive pronoun as the possessor and the noun of the determiner condition as the possessed. The presence of the "pos-s" condition then triggers its transfer to the DRS. A sample DRS is given in Figure 6.

Figure 6. Final DRS for sample sentence


u, v, x, y, z, a, b, c, d, e
Thomas(u), Eaton(v),
propername=uv
Andover(x), Unice(y), Singletary(z),
of=a
propername=yz
Salisbury(a), v married z, b=v, freeman(c),
at x
day=6
month=January
year=1659
b was c, d=v, Indians(e), d was killed
by e
day=15
month=March
year=1698

The architecture of LG-Soar allows for the arbitrary extension and branching of this three-step series of translations. Robustness is achieved by ignoring parts of a sentence from which no acceptable parse can be determined. A similar criterion is used during translation phases. At each level or representation, only certain "triggering" configurations allow for the generation of structure at the subsequent level. This allows certain links to be ignored during the first phase. During the second phase, semantically uninterpretable conditions, perhaps arising from a massive failure in the Link Grammar parse, do not prevent the salvaging of some relationships. At each phase, only relevant information is transferred. In future work, additional knowledge sources can be used at some levels in order to make distinctions that were not explicitly represented at the previous level.

By the time the DRS construction is complete, the syntax of the source sentence has been suppressed, and all content is described as discourse referents and conditions. Having constructed the DRS, specific user-oriented output representations can be generated.

OUTPUT FORMATS

The system is capable of outputting the extracted information in a variety of formats. For example, predicate-argument relationships such as those depicted in Figure 6 can be output directly. DRT has defined a data structure called discourse representation structures; the data can be output in DRS format as well. A tool called CLIG (computational linguistics interactive grapher) has a Tcl/C implementation; it was integrated into LG-Soar successfully to output DRSs from the extracted information. Potentially most useful, though, is the GEDCOM (genealogical data communication) format, which is the de facto standard for exchanging genealogical data. The LG-Soar system is capable of outputting the extracted information in GEDCOM format, which can be used by a large variety of personal history products. Only a few highly specific and pertinent DRS conditions trigger the GEDCOM data structures. For example, a verbal condition with the verb "died" or "killed" indicates someone's death. The advantage of constructing the GEDCOM data structure from the DRS (as opposed to, say, the Link Grammar parse) is that the DRS as a semantic representation denotes many possible syntactic constructions identically. So rules for constructing the GEDCOM data at the DRS level can easily cover more possible sentences than rules at a previous level.

FUTURE WORK AND APPLICATIONS

This work has focused on processing one type of text: Savage's monumental work. However, the goal was to develop a much more widely applicable system. For example, only English text was addressed in this paper, yet many languages follow the same conventions observed in Savage's text, particularly for biographical and genealogical information. Because Link Grammar parser versions have also been developed for other languages (e.g., German and French), it should be possible to integrate them into LG-Soar. The processing of semistructured text was the focus of this paper; however, handling completely unstructured (i.e., free) text should also be possible within our approach. In addition, completely structured text (e.g., from a spreadsheet) should likewise be possible. Additional knowledge sources could be added, such as lexical semantic resources like the WordNet lexical database (Fellbaum 1998). WordNet has been integrated with other Soar projects, and having this resource in the system will allow some automatic inferencing that is now being hand-coded in the discourse section (e.g., the fact that if someone is killed by the Indians on a particular date, that date is his death date). Finally, another exciting aspect of the LG-Soar system follows from the fact that Soar is a machine fully capable of autonomous learning. Though machine learning was turned off in the development of the system as described in this paper, it is perfectly reasonable to assume that many aspects of the task as described can be learned by the system. This should allow it to deal with unseen difficulties and to further optimize processing.

NOTES

[1] This example reflects previous factoring of the data from its original presentation layout.

REFERENCES

Fellbaum, Christiane, ed. 1998. *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Kamp, Hans, and Uwe Reyle. 1993. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory.* Dordrecht: Kluwer Academic Publishers.

Lewis, Richard L. 1993. *An architecturally-based theory of human sentence comprehension.* Technical Report CMU-CS-93-226, Ph.D. dissertation, Carnegie Mellon.

Newell, Allen. 1990. *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Savage, James. 1860–1862. *A genealogical dictionary of the first settlers of New England: Showing three generations of those who came before May, 1692, on the basis of Farmer's register.* Boston: Little, Brown and Co. 4 vols.

Sleator, Daniel, and Davy Temperley. 1993. Parsing English with a link grammar. In *Third International Workshop on Parsing Technologies.*

Lexicostatistics Applied to the Historical Development of Three Languages of the Philippines

Hans Nelson

For years, historical linguists have attempted to classify language family relationships using a variety of methods. One such method is lexicostatistics. "Lexicostatistics . . . is a technique that allows us to determine the degree of relationship between two languages by comparing the vocabularies of the languages and determining the degree of similarity between them" (Crowley 1998, 171). Lexicostatistics is used to determine "(1) time depth (glottochronology), (2) subgrouping, and (3) genetic relationship" (Anttila 1972, 397). The vocabulary used for such comparisons is taken from the Swadesh list of basic vocabulary. This list of 100 English words, developed by Morris Swadesh, is an attempt to formulate a list of lexical items that resist cultural influences and therefore are not easily affected by neighboring languages. Thus, when comparing two languages it is assumed that the words on the Swadesh list have not changed significantly from their original form (Campbell 2000, 177).

Campbell and others challenge this aspect of lexicostatistic methodology, saying that there is no culture-free basic vocabulary in a language. But even though there may not be an impenetrable list of culture-free vocabulary, there are lexical items in a vocabulary that are less likely to have been borrowed from other languages (Anttila 1972, 397). The lexicostatistic approach can, in fact, be useful for subgrouping large language families that are in close geographic proximity with relatively limited lexical data available for analysis. Thus the lexicostatistic approach is useful for an analysis of the languages among the islanders of the Philippines, an area only about the size of New Mexico, with roughly eighty-five to one hundred known languages. The approach could be effective for at least roughly categorizing and subgrouping these variants.

This paper will use lexicostatistics to look at the historical relations among Tagalog, Ilokano, and Bikolano, three languages of the Philippines, with regard to the wave theory. More specifically, this paper explores whether Tagalog and Ilokano are the most closely related of the three languages being compared. This analysis uses Swadesh's list of 100 basic words. The rate of retention, the rate of cross-linguistic loss, and the dates of divergence among the three languages are also calculated. Additionally, the effectiveness of glottochronology is examined as it pertains to effectively finding relationships within the Austronesian family, more specifically the major dialects in use on the Philippine island of Luzon.

Background

Before proceeding, it will be useful to consider a few notions in greater detail. We will first consider the contrast in the kind of questions in which lexicostatistics and glottochronology are most usefully applied. Glottochronology uses word comparisons between languages for calculating dates of divergence. In doing so, it uses Swadesh's list of vocabulary. As previously noted, the Swadesh word list is assumed to be a culture-free, or at least culturally resistant, word list that can help determine dates of divergence more accurately. While the vocabulary is resistant to change, it does slowly change, and it does so at a constant rate that gives some idea of how long particular languages have been diverging from each other. This is analogous to Carbon C dating techniques. Because ^{14}C is radioactive, it has a constant rate of decay, or half-life. The ratio of ^{14}C to ^{12}C in a particular object can determine its age. Likewise, glottochronological methods are typically used to show and explore genetic linguistic relationships among languages, meaning that once their relation is determined, a system is set up in relation to time as opposed to geography.

Glottochronologists try to avoid including loanwords in their work. Loanwords, especially from distant, unrelated regions, contaminate this constant rate of retention. "It is very important that you exclude copied (or borrowed) vocabulary . . . as these can make two languages appear to be more closely related to each other than they really are" (Crowley 1998, 175).

In contrast to glottochronology, lexicostatistics is, as Anttila states, "a wider field of statistics in the service of historical vocabulary studies" (1972, 396). We have noted that languages can be described not only in terms of their genetic relation to one another, but also in terms of their geographic relation, or what Comrie calls *areal* relations (1989, 11). The idea of comparing languages by geographic region rather than time depth is termed "wave theory." As noted by one scholar: "When no particular linguistic innovation can be given chronological priority, subgrouping results in a brush-like tree without depth (one node)" (Anttila 1972, 304). And it is in the geographic dimension where lexicostatistic work can be useful. In both lexicostatistics and glottochronological studies, borrowing from languages that are geographically distant (and thus generally linguistically distant or unrelated) diminishes accuracy in determining the relationship between two geographically adjacent languages. However borrowing from neighboring languages is expected and not harmful to determining their degree of relation.

Lexicostatistics proves to be an effective initial grouping strategy for large language families, though when applied glottochronologically, its calculations for dates of divergences admittedly seem arbitrary, without historical evidence to support the figures. Lexicostatistics also proves less a delineator of genetically dated family relations than it does of geographic family relations. Its main function is to determine the degree to which items in more than one lexicon are related. This method will determine the closeness of the three languages we are considering.

It will be useful to further discuss the languages under consideration. Tagalog and Ilokano were among the three languages selected for comparison due to obvious geographic proximity. They coexist with native speakers of each language separated by only tens of miles. After spending time in the Luzon region, and as a fluent speaker of Tagalog, I became vividly aware of the many different languages in such a small area. As one travels the countryside it becomes evident to the astute listener that these languages may well share a branch of the same linguistic tree. Given the geographic proximity of Tagalog and Ilokano, I became curious as to their origins and their relationship with one another. Let it be stated that the three languages chosen, as well as many or even most other languages spoken in the Philippines, are not merely dialects of each other, but truly languages in their own right. At times the languages share similarities in grammar, syntax, and lexicon, yet they remain fully separated languages and should be treated as such. Originally I had considered using Kapampangan as the third comparison language in this study, but because of limited data resources and time constraints, it was not feasible to do so. However an acceptable replacement, Bikolano, was found. While its geographic area of usage lies slightly south of the area originally envisioned, it is a suitable replacement because it is still geographically adjacent to Tagalog.

As background for what follows, it is necessary to explain how Spanish, Indonesian, and Sanskrit have influenced the three languages under consideration. In 1521, Ferdinand Magellan claimed the land of the Philippines for Spain, whose imperial rule lasted until the United States of America gained possession of the islands after victories in the Spanish-American War of 1898–1901. During the almost four hundred years prior to this, the Spanish had dominated Philippino culture and influenced the language. As to the origins of the Indonesian language influence in the Philippines, Francisco states that Indian influence in the Philippines could be dated to between the tenth and twelfth centuries A.D. on the basis of the linguistic evidences shown in the earliest Old Malay and Old Javanese inscriptions which have been discovered there (1965, xiii). Sanskrit influence entered the Islands by way of the Tamil of the South Indian (Dravidian) culture, which had been in contact with the islands hundreds of years before Indian influences reached the area (Makarenko 1992, 65). Thus, Tagalog, the national language of the Philippines, has been influenced significantly through contact with Sanskrit, Indonesian, and Spanish. Because of these influences, many of Tagalog's original lexical properties have been replaced by loanwords. Not only has Tagalog been influenced, but most other languages in the Philippines have also experienced somewhat parallel modifications. In order for the vocabulary of Tagalog or other Philippine languages to be compared in a historical context with other languages in close geographical proximity, the vocabularies of the languages must be considered in their preinfluenced state to the degree possible.

METHODOLOGY

The first step in this study was to remove any borrowed words from languages that are geographically distant. As mentioned earlier, words borrowed from distant languages tend to skew results in determining the degree of relation between two languages. However, if one is seeking to show a geographic relation between languages, then some inclusion of loanwords from geographically neighboring languages is helpful. The next step was to perform a lexicostatistical comparison of the languages. Once that was completed, then the dates of divergence were determined through glottochronological comparisons.

Any distant lexical borrowing involving the Swadesh list of basic vocabulary presents a problem in calculating rates of language divergence and comparing the degree of similarity between the languages under consideration. A study was undertaken by Zorc using the Swadesh-200 word list (a variation based on the original 100-word list) in which he determined the proto-Tagalic stress of vocabulary, yet he failed to remove distantly geographic borrowed words from his list (1972, 13). Therefore his findings in regard to these borrowed words are not credible. In my study, those words that are known to be borrowed from Spanish, Sanskrit, and Bahasa Indonesian were removed.

After identifying a word borrowed from one of these languages, my initial response was an attempt to replace the word with an older form or synonym in the particular language (see Appendix 2). If no word replacement could be found, then the word was simply dropped from the list. If a particular borrowed word appeared to be more influenced by Philippine languages than by some other language, the word was left in the revised list without modification. Through this process, the original list of 100 words (as shown in Appendix 1) was whittled down to 89 words (see Appendix 3). In Appendix 2, the chart is color coded as follows: white letters on black for Indonesian, black letters on light grey for Sanskrit, and white letters on medium grey for Spanish. Bracketed words are those original borrowed words in the languages from which they were taken. For example, in Appendix 2 the Tagalog, Ilokano, and Bikolano word for *seed* (No. 24) is printed on light grey background; this shows that in all three cases it is borrowed from the Sanskrit language. Next to the Tagalog word *binhi* (No. 24) the Sanskrit word *hiji* from which the Tagalog word is borrowed appears in brackets. Since the Ilokano and the Bikolano words are also printed on light grey and no bracketed word appears next to them, they are also borrowed from the same Sanskrit word shown in the Tagalog column.

In the subsequent discussion of research procedure and findings, the following abbreviations will be utilized. *Tagalog, Ilokano, Bikolano, Spanish, Indonesian,* and *Sanskrit* have been abbreviated to *Tag., Ilk., Bik., Spa., Indo.,* and *Sak.* respectively. Of the 100 words forming the original Tag. word list (Appendix 2), there were 4 loanwords from Spa., 13 loanwords from Indo., and 7 loanwords from Sak (counting only one of the two words listed for No. 71). Thus, 24 percent of Tag.'s core vocabulary (the 100 original words) is composed of loanwords. In Ilk. there were 3 loanwords from Spa., 12 from Indo., and 7 from Sak. Therefore, 22 percent of Ilk.'s basic vocabulary is composed of loanwords. In Bik., there existed only 1 loanword from Spa., 12 from Indo., and 4 from Sak., totaling 17 percent of the core vocabulary of the language as borrowed. From this perspective, there appears to be little doubt as to the relative influence from Spa., Indo., and Sak. on the core vocabulary of these three Philippine languages. Regardless of total core influence, Indonesian clearly dominates in influence.

In the standard Swadesh word list, words No. 1, *I,* and No. 2, *you,* were found to closely resemble the Indo. *aku* and *kau* in all three languages (see Appendices 2 and 3). Yet they resemble and appear to be influenced more by the closer western neighbors of Austronesian than by the more distant Malay East. Thus, they were retained in the list and not treated as loanwords.

With regards to No. 3, *we,* only the inclusive form for each language remained in the list because of the obvious borrowing of the exclusive form in all three languages. The (inclusive) Bik. word *kita* resembles the Indo. word *kita* yet is retained for comparison of the entire entry row. All three languages must be compared, not just two of the languages. The Bik. word No. 4, *ini,* seems to be borrowed from the Indo. *ini.* However it was not removed from the list as a borrowed form because it filled the Bik. No. 4 slot necessary for there to be a three language comparison of the word. If No. 4, *ini,* were removed from the Bik. column, it would mean that both words from Tag. and Ilk. would have no third word from Bik. for comparison. If this were done, the entire entry row would need to be removed. Appendix 5 explains how the other borrowed words listed in Appendix 2 were evaluated, resulting in the revised Swadesh list provided in Appendix 3.

Appendix 3 shows the revised Swadesh list of words that reflects all word removals and replacements for all three languages based on the evaluations in Appendix 5. Words which are shaded light grey and given an asterisk are those words that are still in the list despite replacement or modification in some fashion from the original Swadesh list in Appendix 1. Words or whole lines which are shaded dark grey are those words or list numbers that have been completely removed from the list of words considered for comparison.

GLOTTOCHRONOLOGICAL COMPARISON

In order to enable a glottochronological study, I made a comparison of the remaining 89 basic vocabulary word-list items to discover any cognate forms (see Appendix 3). The cognate sets are marked *A*, *B*, or *C*, depending upon the number of different forms of the word in need of representation (see Appendix 4), (Crowley 1998, 176). All Tag. word forms are marked with the letter *A* to show a single word form. If a second form is found in Ilk. unlike the first *A* form, then the letter mark *B* is assigned to denote a new form and so forth with the letter *C* showing an even different form than that of *A* or *B*. If all forms of a particular word in all three languages are cognates, they were all marked with the letter *A*. This process of marking the cognate sets continues for each Swadesh list number until all 89 words are reviewed for their relationship to each other.

LIMITED LEXICAL DATA

Biko. presented a new problem in the application of glottochronology. Because of its limited lexicon, some words could not be obtained for the basic vocabulary list, so a justifiable means had to be developed for giving these a fair trial *in absentia*. After much consideration and deliberation, a method was decided upon. There was a total of 14.6 percent of the lexical data missing. These missing words are represented on a background of light grey (see Appendix 4). A system was devised whereby the missing word would receive a letter based upon a percentage corresponding to the ratio order of the others. For example, the number of times in a given combination that the first two words were *A* and the last was *A* was 59.5 percent of the time, and the number of times in a given combination that the first two words were *A* and the last was *B* was 40.5 percent of the total. So, it then follows that if there were 4 of the 13 missing items in which the first two words were marked *A*, then approximately 2 of those statistically should be *A*, and 2 should be given the letter *B*. By this means the 13 missing items were reconciled. Table 1 shows the number breakdown (see also Appendix 4).

COGNATE PERCENTAGES

With that dilemma resolved, the cognate percentage figures were available for calculation. The cognate percentage calculations shown in Table 2 clearly demonstrate that 46 of the 89, that is, about 52 percent, of the core vocabulary words in Tagalog and Ilokano are cognates. Tagalog and Bikolano also share about 52 percent of their basic vocabulary, and Ilokano and Bikolano share

Table 1. Combinations

SL#	Tagalog	Ilokano	Bikolano	%Occurrence	#Missing	Distribution
1	A	A	A	59.5%	4 AA	2A
2	A	A	B	40.5%		2B
3	A	B	A	44.1%	9 AB	4A
4	A	B	B	0.0%		OB
5	A	B	C	55.9%		5C
				Total Missing	**13**	**13**

Table 2. Cognate percentages

Tagalog			Tagalog		
46 of 89	Ilokano		51.69%	Ilokano	
46 of 89	27 of 89	Bikolano	51.69%	30.34%	Bikolano

about 30 percent of their vocabulary as cognates.

The means for determining cognates, as suggested by Crowley (1998, 178), was primarily based upon the inspection method wherein "intelligent guesswork" was applied in determining whether or not the two forms were cognates. Also, some systematic sound correspondences were used among the languages to determine true cognates. For instance, Blust points out the relationship of the *r* and *g* in the Southern Luzon languages of the Philippines (1991, 73–129). Other sound changes worth mentioning were *d* and *r* along with the standard *l* and *r* changes. The phonological aspect of the comparison undertaken in this study will not be discussed in any further depth, for it was not the primary focus in determining cognate relationships.

Dates of Divergence

As a next step, the glottochronology formula was invoked to calculate the dates of divergence among the languages. The figure of 86 percent was used as the established *r*, or constant change factor, in the mathematical formula to work out the time depth, or the period of separation, of two languages; $t = \log C/2\log r$; as given by Campbell (2000, 179). The following findings emerged: (1) Tag. and Ilk. diverged about 2,200 years ago. (2) Tag. and Bik. diverged about 2,200 years ago. (3) Ilk. and Bik. diverged about 4,000 years ago. These findings strongly imply that Tag. and Ilk. are languages of a common family, as are Tag. and Bik. It seems that Bik. diverged from Tag. at about the same time Ilk. diverged from Tag. Both are related more to Tag. than to each other.

Conclusion

Prior to the research, it had been theorized by the investigator that Tagalog would have a greater cognate relationship with Ilokano than it would with Bikolano. This assumption did not hold true. Tagalog seems to have an equal relationship with both of the languages. When the percentages are examined, it appears that the ratio of common core vocabulary is the same for Bik. and Tag. as it is for Ilk. and Tag. The similarity of the relationships between Ilk. and Tag. and Bik. and Tag. could result from the specific list of comparison words chosen or from the partial subjectivity in the determination of cognates. The results are surprising, considering that Bikolano's area of usage is farther south and therefore it is less of a geographic neighbor to Tagalog than is Ilokano.

Concluding anything about subgrouping causality at this point would be presumptuous and not consistent with the limited evidence. However, the notion that Tagalog is a parent to the two other languages appears justified. The fact that it shares 50 percent of its cognates with both of these languages is an interesting finding and a stimulus for further research. Further studies are envisioned which will attempt comparisons of additional languages that the island of Luzon has to offer. These studies will build on the data gathered during the research just completed. A limited review of the literature reveals little existing scholarly work of this type using languages of the Philippines. As to the effectiveness of using glottochronological methodology within the Philippine language cluster, definitive conclusions are premature. Insufficient studies of a similar type exist with which these results can be compared.

This research seems to demonstrate that glottochronology could be effective in showing core cognate percentages of languages and in turn giving historical linguists a starting point for determining language groupings among families. Relative to Campbell's concerns about the Swadesh list of basic vocabulary, it must be admitted that the 100-word list is not a totally culture-free, unchanging list. However, it can be effectively argued that by removing the borrowed words from the original list, one can form an essentially culture-free word list sufficient to undertake exploratory research. The most important contribution of this study may be the platform and stimulus it provides for additional research applications and modifications of the methodology in the future.

REFERENCES

Anttila, Raimo. 1972. *An introduction to historical and comparative linguistics.* New York: Macmillan.

Blust, Robert. 1991. The greater central Philippines hypothesis. *Oceanic Linguistics* 30 (2): 73–129.

Campbell, Lyle. 2000. *Historical linguistics: An introduction.* Cambridge: MIT Press.

Comrie, Bernard. 1989. *Language universals and linguistic typology.* 2nd ed. Chicago: The University of Chicago Press.

Constantino, Ernesto. 1971. *Ilokano dictionary.* Honolulu, Hawaii: University of Hawaii Press.

Crowley, Terry. 1998. *An introduction to historical linguistics.* 2nd ed. Auckland, New Zealand: Oxford University Press.

English, Leo James. 1997. *English-Tagalog dictionary.* 23rd ed. Metro Manila, Philippines: National Book Store, Inc.

English, Leo James. 1996. *Tagalog-English dictionary.* 12th ed. Metro Manila, Philippines: National Book Store, Inc.

Francisco, Juan R. 1965. *Indian influences in the Philippines.* Manila, Philippines: Benipayo Printing Co., Inc.

Grace, George W. 1964. The linguistic evidence. *Current Anthropology* 5 (5): 361–68.

Llamzon, Teodoro A. 1978. *Handbook of Philippine language groups.* Quezon City, Philippines: The Ateneo de Manila University Press.

Makarenko, V. A. 1992. South Indian influence on Philippine languages. *Philippine Journal of Linguistics* 23 (1–2): 65–77.

Rubino, Carl R. Galvez. 1998. *Ilocano.* New York: Hippocrene Books.

Silverio, Julio F. 1980. *New English-Pilipino-Bicolano dictionary.* Caloocan City, Philippines: National Book Store, Inc.

Wolff, John U. 1971. *Beginning Indonesian: Part one.* Ithaca, New York: Cornell University.

Zorc, R. David. 1972. Current and proto Tagalic stress. *Philippine Journal of Linguistics* 3 (1): 43–57.

Appendix I

Swadesh 100-Word List of Basic Vocabulary

SL#	English	Tagalog	Ilokano	Bikolano
1	I	ako	siak	ako
2	you (sig.)	ka	sika	ika
3	we (incl.,excl)	tayo, kami	datayo, dakami	kita, kata
4	this	ito, iri	daytoy	ini
5	that	iyan, iyon	dayta	idto
6	what	ano	ania	ano
7	who	sino	sino	isay
8	no	hindi, wala	saan, awan	dai, buku
9	all	lahat	amin	gabos
10	many	marami	naruay	orog
11	one	isa	maysa	saro
12	two	dalawa	dua	duwa
13	big	malaki, dakila	dakkel	dakula
14	long	mahaba	atiddog	maigot
15	small	maliit	bassit	saditsadit
16	woman	babae	babai	babaye
17	man	lalaki	lalaki	lalaki
18	person	tao	tao	tawo
19	fish	isda	sida	sira
20	bird	ibon	billit	gamgam
21	dog	aso	aso	ido
22	louse	lisa	lis-a	lisa
23	tree	punong kahoy	kayo	poon
24	seed	binhi	bukel	butud
25	leaf	dahon	bulong	saka
26	root	ugat	ramut	pungo
27	bark	banakal	ukis ti kayo	upak
28	skin	balat	kudil	
29	flesh	balat	lasag	
30	blood	dugo	dara	dago
31	bone	buto	tulang	butud
32	egg	itlog	itlog	sogok
33	grease	grasa	manteka	taba
34	horn	sungay	sara	
35	tail	buntot	ipus	ikog
36	feather	balahibo	dutdot	
37	hair	buhok	buok	buhok
38	head	ulo	ulo	ulo
39	ear	tainga	lapayag	talinga
40	eye	mata	mata	mata
41	nose	ilong	agong	dongo
42	mouth	bibig	ngiwat	ngoso
43	tooth	ngipin	ngipen	ngipon
44	tongue	dila	dila	dila
45	claw	kuko	kuko	kawit
46	foot	paa	saka	bitis
47	knee	tuhod	tumeng	tohod
48	hand	kamay	ima	kamot
49	belly	tiyan	tian	tikab

50	neck	leeg	tengnged	liog
51	breast	suso	suso	daghan
52	heart	puso	puso	
53	liver	atay	dalem	
54	drink	uminom	uminum	
55	eat	kumain	mangan	kumaon
56	bite	kumagat	kagat	kagat
57	see	makita	kita	hilingon
58	hear	makinig	mangngeg	
59	know	alam	am-ammo	maraman
60	sleep	tumulog	turog	turog
61	die	mamatay	matay	
62	kill	patayin	patayen	
63	swim	lumangoy	aglangoy	
64	fly	lumipad	agtayab	lumpat
65	walk	lumakad	magna	lakaw
66	come	lumipat	umay	dolokon
67	lie	humiga	agidda	kabuwaan
68	sit	umupo	agtugaw	
69	stand	tumayo	agtakder	
70	give	ibigay	ited	bugay
71	say	magsalita, magwika	sarita	tataramon
72	sun	araw	init	aldaw
73	moon	buwan	bulan	bulan
74	star	bituin	bitwen	bituon
75	water	tubig	danum	
76	rain	ulan	tudo	uran
77	stone	bato	bato	bako
78	sand	buhangin	darat	baybay
79	earth	mundo	daga	kinaban
80	cloud	ulap	ulep	ambon
81	smoke	usok, aso	asuk	aso
82	fire	apoy	apuy	kalayo
83	ash	abo	dapo	abo
84	burn	sunog	uramen	pasuon
85	path	daan	dana	agihan
86	mountain	bundok	bantay	bukid
87	red	pula	nalabaga	pula
88	green	berde	berde	
89	yellow	amarilyo	amarilio	amarilio
90	white	maputi	puraw	
91	black	maitim	nangisit	itom
92	night	gabi	rabii	banggi
93	hot	mainit	napudot	
94	cold	malamig	lamek	takig
95	full	puno	napno	pano
96	good	mabuti	naimbag	magayon
97	new	bago	baro	bago
98	round	mabilog	natimbukel	bilog
99	dry	tuyo	namaga	alang-alang
100	name	pangalan	nagan	ngaran

Appendix 2

List of Borrowed Words - Color Coded

SL#	English	Tagalog	Ilokano	Bikolano
1	I	ako [aku]	siak	ako
2	you (sig.)	ka [kau]	sika	ika
3	we (incl.,excl)	tayo, kami [kami]	datayo, dakami	kita, kata [kita]
4	this	ito, iri [ini]	daytoy	ini [ini]
5	that	iyan, iyon	dayta	idto
6	what	ano	ania	ano
7	who	sino	sino	isay
8	no	hindi, wala	saan, awan	dai, buku
9	all	lahat	amin	gabos
10	many	marami	naruay	orog
11	one	isa	maysa	saro
12	two	dalawa	dua [dua]	duwa
13	big	malaki, dakila	dakkel	dakula
14	long	mahaba	atiddog	maigot
15	small	maliit	bassit	saditsadit [sedikit]
16	woman	babae	babai	babaye
17	man	lalaki [lalaki]	lalaki	lalaki
18	person	tao	tao	tawo
19	fish	isda	sida	sira
20	bird	ibon	billit	gamgam
21	dog	aso	aso	ido
22	louse	lisa [liksa]	lis-a	lisa
23	tree	punong kahoy	kayo	poon [pohon]
24	seed	binhi [hiji]	bukel	butud
25	leaf	dahon	bulong	saka [sakha]
26	root	ugat	ramut	pungo
27	bark	banakal	ukis ti kayo	upak
28	skin	balat	kudil	
29	flesh	balat	lasag	
30	blood	dugo	dara	dago
31	bone	buto	tulang	butud
32	egg	itlog	itlog	sogok
33	grease	grasa [grasa]	manteka [manteca]	taba
34	horn	sungay	sara [sara]	
35	tail	buntot	ipus	ikog
36	feather	balahibo	dutdot	
37	hair	buhok	buok	buhok
38	head	ulo	ulo	ulo
39	ear	tainga	lapayag	talinga
40	eye	mata [mata]	mata	mata
41	nose	ilong	agong	dongo
42	mouth	bibig	ngiwat	ngoso
43	tooth	ngipin	ngipen	ngipon
44	tongue	dila [lidha]	dila	dila
45	claw	kuko	kuko	kawit
46	foot	paa [pada]	saka [sakha]	bitis
47	knee	tuhod	tumeng	tohod
48	hand	kamay	ima [lima]	kamot
49	belly	tiyan	tian	tikab

50	neck	leeg	tengnged	liog
51	breast	suso	suso	daghan
52	heart	puso	puso	
53	liver	atay [hati]	dalem	
54	drink	uminom [minum]	uminum	
55	eat	kumain	mangan	kumaon
56	bite	kumagat	kagat	kagat
57	see	makita	kita	hilingon
58	hear	makinig	mangngeg	
59	know	alam	am-ammo	maraman
60	sleep	tumulog	turog	turog
61	die	mamatay [mati]	matay	
62	kill	patayin	patayen	
63	swim	lumangoy	aglangoy	
64	fly	lumipad	agtayab	lumpat
65	walk	lumakad	magna	lakaw
66	come	lumipat	umay	dolokon
67	lie	humiga	agidda	kabuwaan
68	sit	umupo	agtugaw	
69	stand	tumayo	agtakder	
70	give	ibigay	ited	bugay
71	say	magsalita, magwika [carita] [vaka]	sarita	tataramon
72	sun	araw	init	aldaw
73	moon	buwan	bulan	bulan
74	star	bituin	bitwen	bituon
75	water	tubig	danum	
76	rain	ulan	tudo	uran
77	stone	bato [batu]	bato	bako
78	sand	buhangin	darat	baybay
79	earth	mundo [mundo]	daga	kinaban
80	cloud	ulap	ulep	ambon
81	smoke	usok, aso	asuk	aso
82	fire	apoy [api]	apuy	kalayo
83	ash	abo [abu]	dapo	abo
84	burn	sunog	uramen	pasuon
85	path	daan	dana	agihan
86	mountain	bundok	bantay	bukid
87	red	pula	nalabaga	pula
88	green	berde [verde]	berde	
89	yellow	amarilyo [amarillo]	amarilio	amarilio
90	white	maputi [pudi]	puraw	
91	black	maitim	nangisit	itom
92	night	gabi	rabii	banggi
93	hot	mainit	napudot	
94	cold	malamig	lamek	takig
95	full	puno [penuh]	napno	pano
96	good	mabuti [bhuti]	naimbag	magayon
97	new	bago	baro	bago
98	round	mabilog	natimbukel	bilog
99	dry	tuyo	namaga	alang-alang
100	name	pangalan	nagan	ngaran

Spanish
Sanskrit
Indonesian
Original Item

Appendix 3

Revised Swadesh List (reflects removals, replacements, and substitutions):

SL#	English	Tagalog	Ilokano	Bikolano
1	I	ako	siak	ako
2	you (sig.)	ka	sika	ika
3	we (incl.,excl)	tayo*	datayo*	kita*
4	this	ito*	daytoy	ini
5	that	iyan*	dayta	idto
6	what	ano	ania	ano
7	who	sino	sino	isay
8	no	hindi*	awan*	dai*
9	all	lahat	amin	gabos
10	many	marami	naruay	orog
11	one	isa	maysa	saro
12	two	dalawa	aduwa*	duwa
13	big	dakila*	dakkel	dakula
14	long	mahaba	atiddog	maigot
15	small	maliit	bassit	saday*
16	woman	babae	babai	babaye
17	man	lalaki	lalaki	lalaki
18	person	tao	tao	tawo
19	fish	isda	sida	sira
20	bird	ibon	billit	gamgam
21	dog	aso	aso	ido
22	louse	kuto*	kuto*	kuto*
23	tree	punong kahoy	kayo	poon
24	seed	buto*	bukel	butud
25	leaf	dahon	bulong	saka
26	root	ugat	ramut	pungo
27	bark	banakal	ukis ti kayo	upak
28	skin	balat	kudil	
29	flesh	balat	lasag	
30	blood	dugo	dara	dago
31	bone	buto	tulang	butud
32	egg	itlog	itlog	sogok
33	grease	taba*	taba*	taba
34	horn	sungay	sara	
35	tail	buntot	ipus	ikog
36	feather	balahibo	dutdot	
37	hair	buhok	buok	buhok
38	head	ulo	ulo	ulo
39	ear	tainga	lapayag	talinga
40	eye	mata	mata	mata
41	nose	ilong	agong	dongo
42	mouth	bibig	ngiwat	ngoso
43	tooth	ngipin	ngipen	ngipon
44	tongue	dila	dila	dila
45	claw	kuko	kuko	kawit
46	foot	paa	saka	bitis
47	knee	tuhod	tumeng	tohod
48	hand	kamay	ima	kamot
49	belly	tiyan	tian	tikab

50	neck	leeg	tengnged	liog
51	breast	suso	suso	daghan
52	heart	puso	puso	
53	liver	atay	dalem	
54	drink	uminom	uminum	
55	eat	kumain	mangan	kumaon
56	bite	kumagat	kagat	kagat
57	see	makita	kita	hilingon
58	hear	makinig	mangngeg	
59	know	alam	am-ammo	maraman
60	sleep	tumulog	turog	turog
61	die	yumao*	matay	
62	kill	patayin	patayen	
63	swim	lumangoy	aglangoy	
64	fly	lumipad	agtayab	lumpat
65	walk	lumakad	magna	lakaw
66	come	lumipat	umay	dolokon
67	lie	humiga	agidda	kabuwaan
68	sit	umupo	agtugaw	
69	stand	tumayo	agtakder	
70	give	ibigay	ited	bugay
71	say	magsabi*	kuna*	tataramon
72	sun	araw	init	aldaw
73	moon	buwan	bulan	bulan
74	star	bituin	bitwen	bituon
75	water	tubig	danum	
76	rain	ulan	tudo	uran
77	stone	bato	bato	bako
78	sand	buhangin	darat	baybay
79	earth	daigdig*	daga	kinaban
80	cloud	ulap	ulep	ambon
81	smoke	usok, aso	asuk	aso
82	fire	apoy	apuy	kalayo
83	ash	abo	dapo	abo
84	burn	sunog	uramen	pasuon
85	path	daan	dana	agihan
86	mountain	bundok	bantay	bukid
87	red	pula	nalabaga	pula
88	green	lunti*	nalangto*	
89	yellow	dilaw*	kiaw*	darag*
90	white	maputi	puraw	
91	black	maitim	nangisit	itom
92	night	gabi	rabii	banggi
93	hot	mainit	napudot	
94	cold	malamig	lamek	takig
95	full	puno	napno	pano
96	good	maganda*	naimbag	magayon
97	new	bago	baro	bago
98	round	mabilog	natimbukel	bilog
99	dry	tuyo	namaga	alang-alang
100	name	pangalan	nagan	ngaran

Altered Item*

Removed Item

Appendix 4

Cognate List:

SL#	English	Tagalog	Ilokano	Bikolano
1	I	A	A	A
2	you (sig.)	A	A	A
3	we (incl.)	A	A	B
4	this	A	B	C
5	that	A	B	C
6	what	A	A	A
7	who	A	A	B
8	no	A	B	C
9	all	A	B	C
10	many	A	A	B
11	one	A	A	B
12	two	A	A	A
13	big	A	A	A
14	long	A	B	C
15	small	A	B	C
16	woman	A	A	A
18	person	A	A	A
19	fish	A	A	B
20	bird	A	B	C
21	dog	A	A	B
22	louse	A	A	A
24	seed	A	B	A
25	leaf	A	B	A
26	root	A	B	C
27	bark	A	B	C
28	skin	A	B	A
29	flesh	A	B	A
30	blood	A	A	A
31	bone	A	B	A
32	egg	A	A	B
33	grease	A	A	A
35	tail	A	B	C
36	feather	A	B	A
37	hair	A	A	A
38	head	A	A	A
39	ear	A	B	A
40	eye	A	A	A
41	nose	A	A	A
42	mouth	A	B	C
43	tooth	A	A	A
45	claw	A	A	B
47	knee	A	B	A
48	hand	A	B	A
49	belly	A	A	B
50	neck	A	B	A
51	breast	A	A	B
52	heart	A	A	A
55	eat	A	B	A
56	bite	A	A	A

57	see	A	A	B
58	hear	A	A	A
59	know	A	A	A
60	sleep	A	A	A
62	kill	A	A	B
63	swim	A	A	B
64	fly	A	B	A
65	walk	A	B	A
66	come	A	B	C
67	lie	A	A	B
68	sit	A	B	C
69	stand	A	B	C
70	give	A	B	A
71	say	A	B	C
72	sun	A	B	A
73	moon	A	A	A
74	star	A	A	A
75	water	A	B	C
76	rain	A	B	A
77	stone	A	A	A
78	sand	A	B	C
79	earth	A	B	C
80	cloud	A	A	B
81	smoke	A	A	A
82	fire	A	A	B
84	burn	A	B	C
85	path	A	A	A
86	mountain	A	B	C
87	red	A	B	A
88	green	A	B	C
89	yellow	A	A	B
91	black	A	B	A
92	night	A	A	B
93	hot	A	B	C
94	cold	A	A	B
96	good	A	B	C
97	new	A	A	A
98	round	A	B	A
99	dry	A	B	C
100	name	A	A	A

Total 89

Calculated Letter

Appendix 5

With list word No. 4, the Indo. word *ini* was removed in Tag. because a more suitable replacement was found using the word *ito*. List word No. 8, *no*, meaning negative, has been used in all three languages, and the *no*, meaning "have not any," or "none," or "without" has been removed. Standard word No. 12, the word for "two," shows contact with India; however the word has been derived from the Austronesian *duwa* (cf. Francisco 1965, 44). The Ilk. form has been changed to reflect a dialect pronunciation of the word expressing a more suitable Austronesian sound.

In the Tag. list word No. 13, *malaki*, has been removed, for it resembles too closely the word for "man" in Indo., and thus is likely borrowed. Of the two words originally listed as possible Tag. words with the meaning appropriate for No. 13, the more Tag. word *dakila* was left. Word No. 15, the Bik. *saditsadit*, which resembled the Indo. *sedikit*, was removed and replaced with Bik. *saday*, also meaning "smallness," leaving its original meaning intact. Word No. 17, *lalaki*, in all three languages has been totally removed from the list in all three languages because it was borrowed from the Indo. word *lalaki*. The standard list word No. 22 in all three languages has been replaced with the older more mature meaning of the word for "louse," *kuto*, versus the borrowed young immature meaning of "louse," *lisa*, taken from Sak. *liksa*.

Word No. 23 has been removed from the list in all three languages because the Bik. *poon* is borrowed from the Indo. *pohon* and the Tag. *punong kahoy*, meaning "tree of wood," with the root *puno* meaning "tree," was also borrowed. This leaves only the Ilk. *kayo* left in the standard list that is not borrowed. Having no other language with which to compare the Ilk. form, standard list No. 23 was removed. Tag. No. 24, *binhi*, borrowed from Sak. *hiji*, was replaced with another form, *buto*, meaning "bone" and "seed." Bik. word No. 25, *saka*, borrowed from Sak. *sakha*, has been removed as no suitable replacement was found. Tag. word No. 33, *grasa*, borrowed from Spa. *grasa*, has been replaced by the Tag. *taba*, meaning "fat" in both Tag. and Ilk. The Ilk. word *manteka*, borrowed from *manteca*, meaning "butter," has been replaced by *taba*.

Standard list word No. 34 has been totally removed from the list in all three languages because of the borrowing of the Ilk. *sara* from Sak. *sara*, meaning "horn." This left only one word in Tag. and none in the Ilk. or the Bik. list. With nothing remaining with which to compare the Tag. word, No. 34 was removed in all three languages. Standard word No. 44 has been totally removed in all three languages because *dila* is borrowed from the Sak. *lidha*. The entry row for standard word No. 46 has been removed since both the Tag. and Ilk. words are borrowed from Sak. Standard word No. 48, Ilk. *ima*, possibly borrowed from Sak. *lima*, meaning "five," has been left even though it could be a borrowed word (one possible argument for it as a Sak. borrowing is that it could be that the Sak. word meaning "five" became a representation of the word for "hand," which consists of five fingers).

The entry row for word No. 53 has been removed from the list because the Tag. *atay* is borrowed from Indo. *hati*, thus leaving only one word for comparison. Word No. 54 is borrowed, in both Tag. and Ilk. from Indo. *minum*, and the entry row has therefore been removed from the list. Tag. word No. 61, *mamatay*, is borrowed from Indo. *mati* and was replaced with *yumao*, meaning to "pass away," yet no replacement could be found for the Ilk. form, which was also borrowed from Indo. This left insufficient data for comparing the three languages. Because of this, No. 61 in both the Tag. and Ilk. languages was removed.

Tag. No. 71 *magsalita* and *magwika* were borrowed from Sak. *carita* and *vaka.* The Ilk. *sarita* was also borrowed from Sak. They have been replaced with Tag. *magsabi* and Ilk. *kuna.* Standard list word No. 77 appears to be borrowed from Indo. *batu,* yet it was not removed in all three languages for the same rationale as applied in the case of word No. 12. Tag. word No. 79, *mundo,* was borrowed from Spa. *mundo* and was replaced with Tag. *daigdig,* an older Tag. word also meaning "world" or "earth." The Tag. *apoy* and Ilk. *apuy* words for No. 82 appear similar to the Indo. *api,* yet they will not be considered to be truly borrowed from the Indo. language. In all three languages, No. 83 is removed from the list because all forms are borrowed from the Indo. *abu.*

Tag. word No. 88, *berde,* is borrowed from Spa. *verde,* meaning "green" and was replaced with an older Tag. form *lunti.* The Ilk. *berde,* borrowed from Spa., was replaced with *nalangto,* also an older synonym for the word. All three languages use word No. 89 as *amarilyo* or *amarilio.* Both are borrowed from Spa. *amarillo,* meaning "light grey." They were replaced with the Tag. *dilaw,* the Ilk. *kiaw,* and the Bik. *darag.* No. 90 was removed from the list in all three languages. The Tag. *maputi* is believed to be borrowed from Sak. *pudi,* meaning "honorable, pure, virgin" and was thus replaced. This left only one word for comparison. No. 95 has been removed from the list in all three languages because it is borrowed from Indo. *penuh,* meaning "full." Tag. No. 96, *mabuti,* is borrowed from Sak. *bhuti* and was therefore replaced with the Tag. word *maganda,* which means "beautiful."

Missionary Language Mangling

Lynn Henrichsen

Missionaries of the Church of Jesus Christ of Latter-day Saints have a reputation that is celebrated in many ways. In the nineteenth century, for instance, there was a popular song that claimed, "None can preach the gospel like the Mormons do." Here is how it went:

> We're going to preach the gospel to all who want to hear.
> A message of salvation unto the meek we'll bear.
> Jehovah has commanded us, and therefore we must go.
> For none can preach the gospel like the Mormons do, like the Mormons do.
> (Kaufman 1980, 44)

Missionary Language Learning Success

In the early days of the Church, Mormon missionaries taught primarily in their native language—English. As years went on, however, the gospel was taken to nations where English was not spoken (for additional details see Henrichsen 1999). Under these circumstances, missionaries began learning foreign languages as part of their missionary duty. This was in fulfillment of modern-day scripture: "For it shall come to pass in that day, that every man shall hear the fulness of the gospel in his own tongue, and in his own language" (D&C 90:11).

Speaking of missionaries and their investigators in foreign lands, Joseph Smith proclaimed, "Let the Elders preach to them in their own tongue, whether it is German, French, Spanish, or Irish, or any other" (Smith 1976, 195).

Later Brigham Young added, "We should be familiar with the various languages, for we wish to send missionaries to the different nations and the islands of the sea. We wish missionaries who may go to France to be able to speak the French language fluently, and those who may go to Germany, Italy, Spain, and so on to all nations, to be familiar with the languages of those nations" (Richards 1974, 8:49).

In this endeavor, missionaries have been marvelously successful. Throughout the world, our missionaries have a reputation for being good language learners, learning to speak the local language fluently.

A few years ago, *U.S. News and World Report* carried an article on Mormon missionaries in Japan. In this article, the author wrote that "Most missionaries become somewhat to very accomplished in their host country's language. . . . Of the Americans I've met in Asia who can operate deftly and successfully in the local language, a disproportionate number have been Mormons" (Fallows 1988).

Perhaps we should update the old missionary song to go like this: "For none can learn a language like the Mormons do, like the Mormons do"

The Downside

This success is not without its downside, however. One problem is the impression many people have that young missionaries enter the Missionary Training Center (MTC) knowing nothing of their mission language and emerge two months later speaking it with total fluency and 100 percent accuracy. This illusion not only trivializes the difficulty of learning a foreign language but may also lead to impatience with others who are learning languages (such as immigrants who have "been here six months and still can't speak English"), leading to questions like "What's wrong with them? My son went into the MTC and learned his mission language perfectly in just two months!" Another problem with this illusion is that it may lead other language learners to have a poor self-concept as they struggle with the complexities of their new language. After years of study, most still have limited fluency and ask themselves, "Why can't I master this language like the missionaries do? Why do I struggle so hard? What's wrong with me?"

The Reality of Missionary Language Learning

In most cases, the answer is "Nothing!" The reality is that, despite their great successes, missionaries also struggle with language difficulties.

Many missionaries (myself included) can recount how they left the MTC thinking they had mastered their mission language only to arrive in the field and find that they could not understand the natives and the natives could not understand them. Although they "lived their language" at the MTC and managed to communicate with their companions in a slowed-down, pidginized form of their target language, communicating with natives in a native-like fashion is quite another thing. Once in the field, many can empathize with Mark Twain, who after visiting France said, "In Paris they simply stared when I spoke to them in French. I never did succeed in making those idiots understand their own language" (cited in Benderson 1983, 1).

More seriously, research at the MTC has found that after two or three months of intense language study, most missionaries leave with only an elementary level of proficiency in their mission language. On the Foreign Service Institute language proficiency rating scale, they generally score at about the 1+ level (Eric Ott, pers. comm., 20 March 2001). At this level, a missionary "can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience [a new missionary] can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase" (Wilds 1975, 37).

After a humbling, frustrating experience or two, most missionaries quickly realize that they have a lot to learn before they will be able to communicate the way they want to.

Even after two years of using their mission language, most missionaries reach only the 2+ FSI level, "Limited Working Proficiency." They "can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information; [they] can [also] handle limited work requirements" (Wilds 1975, 37) especially on gospel and Church-related topics. This is not to say, however, that they do not occasionally stumble and miscommunicate—whether they are talking about the gospel or wandering into less familiar territory. These mistakes can be humbling, embarrassing, even frustrating for a dedicated missionary trying to represent the Lord. They can also be humorous—not necessarily at the time they occur, but at least in retrospect.

Stories of Missionary Language Mangling

The linguistic errors that missionaries commit may involve a small slip that leads to a big miscommunication, or they may entail outright language mangling. The examples that follow are true stories that I have collected over the years. Some of these are personal, some come from friends and acquaintances, and others come from the Mormon-Humor listserv (All quotations from this list are given largely as they appeared. I have edited them lightly, mainly for punctuation and spelling).

I would like to warn readers that, although innocuous-sounding to untrained ears, some parts of this paper may be unintentionally offensive to speakers of various languages.

My Own Story: Being "Embarrassed" in Spanish

I'll start with my own story. I had an experience that many Spanish-learning missionaries have had. It happened a couple of nights after I arrived in Mexico. We went to a ward social, and I was trying to speak Spanish with a group of members. I struggled and was humbled by the mistakes I knew I was making left and right. Not knowing the correct Spanish word for "embarrassed" (*avergonzado*), I made my best guess and explained, "Cuando yo hablo español, estoy muy embarazado." Of course, this brought more laughs. I had fallen into the false cognate trap. I thought that *embarazado* meant "embarrassed." Little did I realize (until it was explained to me a little later) that in saying "estoy embarazado" I was saying that I was pregnant.

This is a common mistake, and many returned missionaries have stories similar to mine. On the Mormon-Humor listserv, Heber Ferraz-Leite (17 December 1998) recounts, "It is told that a new sister missionary was called to bear her testimony. She felt embarrassed, because she didn't speak that well yet, so she went up and the first thing she said was 'I'm pregnant, and it's the bishop's fault!'"

Other Stories Involving False Cognates

I recently learned that English-speaking missionaries trying to communicate in Spanish are not the only ones who make false cognate mistakes like this. In Portuguese, the word *embaraçada* has a meaning very close to the English word *embarrassed.* My wife knows of a Brazilian sister who was called to serve as a missionary in Uruguay, where she had to speak in Spanish, a close relative to her native Portuguese, but in some key points quite different. She emerged from a meeting with some of the elders in her area—a meeting in which something funny had happened—and shocked the other sisters by telling them, "Me dejaron embarazada." Of course, the close Portuguese equivalent, "Me dejaram embaraçada," her intended meaning, meant only that she was embarrassed, not pregnant.

In a contribution to the Mormon-Humor listserv, Richard B. "Andy" Anderson (18 December 1998) explains false cognates with a French twist:

> The French language, after all, is the very breeding ground (forgive me!) of what linguists call the "faux ami"—"false friend": you're sure you know what it means, because it looks or sounds so much like something you know. . . . My beloved companion's greatest achievement along these lines . . . may have been his attempt to compliment a family upon the attractive apartment into which they had kindly invited us. He said, "Ah, que vous vivez dans la luxure ici!" He *thought* he was saying, "Ah, you live in luxury here!" *La luxure,* however it may seem, does not mean "luxury," but rather "abominable lewdness." The French word for "luxury" is *luxe.* As in "deluxe." We were not invited back.

Mispronunciation

Another category of slips involves mispronunciation. Missionaries have to learn that small differences in a sound or two can lead to big differences in meaning.

Mensaje/Masaje

For instance, you may have heard the story of the two Spanish-learning elders who, when speaking to a woman on a doorstep, said, "Estamos aquí para darle un masaje" (We are here to give you a massage). Of course, they intended to say *mensaje* (message), but they didn't realize their mistake until after she gave them a shocked look and quickly closed the door in their faces.

Muy casado

On the Mormon-Humor listserv, Mike L. Hardy (20 December 1998) shared this story of a similar mistake:

> Here's a foreign language funny that happened to me. In Spanish, *Cansado* means "I'm tired," and *Casado* means "I'm married." While in Mexico, someone asked me if I was tired, to which I responded in a very tired voice, "Yes I'm very married."

Makes you wonder what they thought about Mormons and polygamy after that, doesn't it?

Hambre/Hombre

On the same listserv, Steven Leuck contributed this story of confusion and humor resulting from the substitution of an [o] for an [a] sound:

> One of my personal favorites was at lunch one day with several Elders and a pair of sisters. One of the sisters was relatively new in the mission and struggling with the language. Her slip up involved the non-interchangeability of the words *hombre* (man) and *hambre* (hunger). As she came into the room full of Elders and the family serving our meal she let out a big sigh and said rather loudly: "Wow! What hunger I have!!!" Unfortunately for her but to our ample chuckleability factor it came out as "Wow! What a MAN I have!!!" This was funny enough but then we were rolling in the aisle when the young lady serving us didn't miss a beat by exclaiming "Where, where, I want one too!!"

Pecado/Pescado

Thomas Stoddard (17 December 1998) contributed this story to the Mormon-Humor listserv:

> The word *pecado* means "sin." The word *pescado* is "fish." Many a time a young elder has stated to a contact or investigator, "We do not believe in original fish."

Ordenar/Ordeñar

Stoddard continues with another funny example:

> Or, the two words *ordenar* and *ordeñar* (with a tilde over the n̲ such that you say the "ny" sound) . . . *ordenar* means "to order, or ordain," as in confer the priesthood; *ordeñar* means "to milk," as in to milk a cow. Imagine the horror when an investigator learns that "Joseph and Oliver were *milked* by John the Baptist."

Ånd/And

And now, lest you think that all my stories involve Spanish, here's one about a missionary learning Danish (reported by a sister missionary on the listserv, 22 March 2001):

> The Danish word for "spirit" is *ånd;* the word for "duck" (web-footed bird) is *and.* One missionary testified to an investigator that if he prayed about what they had been talking about, that the "holy duck" would manifest the truth unto him!

Lexical Substitution (Similar Sounding Words)

Sometimes errors that missionaries make involve using the wrong word, not just the wrong sound. In fact, the substitution of one word for another is a common mistake. In many cases it is difficult to determine whether these are lexical substitutions or mispronunciations because the two words typically sound similar—especially to the ears of the unsuspecting missionary.

Going Ichigaya/Kichigai

For instance, in the days before the MTC or Language Training Mission (LTM, the forerunner to the MTC), my friend Earl Wyman served in the Northern Far East Mission, which included the Philippines, Hong Kong, Taiwan, and Japan, and he struggled with a new language every time he was transferred. When he was serving in Japan, he found himself one day on a subway in Tokyo headed for the Ichigaya station. He used his elementary Japanese to strike up a conversation with a pleasant Japanese person near him. When that person asked Earl where he was going, Earl replied, "I am going *kichigai.*" He didn't even realize his mistake until the pleasant Japanese person stepped back startled that he had substituted the word for "crazy" (*kichigai*) for the name of his station (*ichigaya*).

Hilse/Hest

A sister missionary who served in Denmark (18 December 1998) tells the following stories about the confusion that resulted when she and her companion substituted one similar-sounding word for another in Danish.

> Soon after I began my mission in Denmark, my trainer and I went to visit an inactive sister. The sister wasn't home, so we talked to her teenage son for a few minutes before leaving. As we were getting ready to leave I decided to try out a phrase I'd heard a lot since I'd been in Denmark: *Hilse,* which means "Say hi." I tried to say, *Hilse din mor for os,* which means "Tell your mother hi for us." Instead of saying, *Hilse,* I said, *Hest din mor for os.* The boy gave me a strange look and went into his house. I turned to my trainer and asked why he'd given me such a funny look. She laughed and said I'd just told him to "Horse your mother for us."
>
> Several months later I was a trainer. My trainee wanted to say, "Because of [the] atonement we can receive forgiveness for our sins." Instead of using the word, *Tilgivelse* (forgiveness), she used the word, *Tilladelse* (permission). So what she actually said was, "Because of [the] atonement we can receive permission for our sins." Luckily we were in our room studying and not teaching an investigator when she said that.

Tengo Miedo/Mierda

On the Mormon-Humor listserv, Marc Page (17 December 1998) shared this story:

> In Spanish, *miedo* means fear; *mierda* means, well, it's a vulgar term for excrement. A sister missionary was trying to say she was afraid (or in Spanish grammar, she had fear), but it didn't come out that way.

Estornudar/Desnudar

In the same list discussion, Rik Andes (17 December 1998), a Spanish-speaking returned missionary, described a fellow missionary's blooper:

> The verb for "to sneeze" in Spanish is *estornudar,* while the verb for "to undress; get naked" is *desnudar.* While trying to explain his allergies to the family he was visiting, he said, "Every time a cat enters the room I get naked."

Rama/Ramera

Rik Andes also reports the following (17 December 1998):

> As if that wasn't enough, the word for "branch" is *rama,* and the word for "whore" (as in "whore of all the earth," a well-known phrase from the Book of Mormon) is *ramera.* During a rather interesting discussion about the wide variety of religions in the world, my companion agreed with the man, saying, "You're right, there are a lot of whores in the world."

Paz y gozo/Gas y pozo

A lexical mix-up of a different kind is illustrated in the following story, contributed by Steven Leuck (17 December 1998), who served in the Argentina Rosario mission:

> During a lesson one day I was explaining about the happiness and joy that come from being a celestial family. The question I posed from the memorized discussions was: "Brother Brown, how would you feel to know that your family could live in peace (*paz*) and happiness (*gozo*) together forever?" Unfortunately, because I juxtapositioned the first letters of the words "peace" and "happiness" it came out as "Brother Brown, how would you feel to know that your family could live in a pit with gas forever?"
>
> My companion told me later that evening why I was getting such strange stares over that one.

Paloma/Plomo

Steven Leuck (17 December 1998) told another funny story about lexical mix-ups:

> I had a companion who accidentally used the word *plomo* (lead, as in heavy metal) when he meant to say *paloma* (dove). His fractured sentence came out: "and then the Holy Ghost descended upon Jesus like lead."

Johannes der Täufer vs. Johannes der Teufel

In German, according to Mark A. Schindler (18 December 1998), the terms *Johannes der Täufer* and *Johannes der Teufel* sound very similar to someone learning the language, but the former is John the Baptist, and the latter is John the Devil. Big theological difference.

Erdbeben vs. Erebeeren

Schindler continues:

> I'm convinced the following is an apocryphal story, but it's too good to let historical accuracy get in the way. The story goes that a junior companion was talking with an investigator about the signs of the times and meant to say, *wenn das große Ende kommt, wird es große Erdbeben überall geben,* which means "when the End [of the world] comes, there'll be big earthquakes everywhere." Of course, what actually came out is *wenn die große Ente kommt, wird es große Erdbeeren überall geben,* which means "when the giant duck comes there'll be big strawberries everywhere," which may not be theologically correct, but appeals more to the palate, you have to admit.

Lexical Substitution (Wild)

Sometimes the substitution of the wrong lexical item can only be explained as the result of a mental process akin to drawing a wild card. The substitute word bears no apparent relationship to the correct word. It's just what popped into the missionary's mind while he or she was searching for the right word in the course of a discussion. The result can be some astounding new Church doctrine.

Two Carrots in the Sacred Grove

When serving in Taiwan, my friend, Earl Wyman struggled to communicate in Chinese (in pre-MTC days). He finally memorized the Joseph Smith story to where he thought he could share it with investigators. He recalls bearing fervent (but unintentional) testimony that in the sacred grove Joseph Smith really did see two *carrots* in a pillar of light. In Chinese the word for "carrot" (*húluóbo*) bears no resemblance to the word for "personages" (*rén*). For some strange reason, however, it was the word that came to his desperate mind at the time.

Incorrect Semantic Domain for a Lexical Item

Occasionally lexical mistakes are not the result of incorrectly substituting one word for another. Rather, they involve choosing the wrong word for the intended meaning. In some cases, these mistakes are not just casual slips but come after careful preparation.

Love vs. Lust After

Such was the case with a mistake made by my friend Earl Wyman. After spending months in Taiwan, he had gained some proficiency in Mandarin Chinese. Of course, he still depended heavily on his bilingual dictionary. As he was preparing to leave Taiwan, he diligently prepared a farewell speech. In it he intended to say, "In my time in China I've come to love the Chinese people." In fact, however, he didn't get quite the right meaning of *love* from his dictionary. Instead of *ài*, which means "to love" in the pure sense of a mother loving her child, he used *tanqiú*, which means "to lust after." We can only hope that the members laughed.

Problems at the Syntactic Level

Of course, mistakes that missionaries make are not limited to the sound or word level. Sometimes they involve incorrectly formed phrases or larger structures. This mistaken formation may be the result of translating word for word from the missionary's native language. It may also involve the insertion or deletion of a key element in a phrase or sentence.

Dio a luz

Mark Page (17 December 1998) tells a story that illustrates this point:

> I had a companion who was trying to say that the Lord brought light to the people in the Americas, but used the phrase *dio a luz*, which literally translates to "give to light" but means "to give birth." So he ended up saying instead that He gave birth to the people in the Americas.

American English Speakers Miscommunicating in Other Varieties of English

Interestingly, it is not necessary for missionaries to be speaking a foreign language to make embarrassing mistakes, unless you count other varieties of English as foreign languages. For example, in England, apartments are called *flats*, and elevators are called *lifts*.

Stuffed

Rob Herr (17 December 1998) told this story on the Mormon-Humor listserv:

> Two sister missionaries, one from the U.S. and the other an Aussie herself, were serving together in Australia. After a large dinner appointment the American sister exclaimed, "I am so stuffed!"
>
> Her Aussie companion was a little shocked at first, knowing that *stuffed* is Aussie slang for being pregnant!

Pants

Another list member (17 December 1998) told a true story that happened to an American elder serving in the UK.

In England, a new missionary complimented a woman who entered the church on the nice looking "pants" she had on because the weather was so bitterly cold. She stared at him to see if she heard right. (*Pants* in Britain refers to undergarments.) He thought she did not hear him, so he repeated it. She promptly slapped him! The poor missionary was quickly educated by his more seasoned companion, and never repeated the mistake!

English Language Teaching Blunders by Missionaries

A final category of missionary language mangling stories involves not language learning but language teaching—particularly, English language teaching. In many missions around the world, young, untrained missionaries set up and teach English classes. In most cases, their only qualification is the fact that they are native speakers of English. Of course, this is hardly sufficient. Thinking, "I speak English natively; therefore, I can teach English to nonnatives" is akin to saying, "I have teeth; therefore, I am a dentist." Nevertheless, missionaries still set up English classes and provide language lessons that are sometimes humorous.

A Spud in Spain

The story is told of a missionary from Idaho serving in Madrid, Spain. He ended up teaching English classes to a rather sophisticated group of learners—lawyers, doctors, educators, etc. One night he had the idea of teaching a lesson on fruits and vegetables. Using his best teaching methods, he brought in some realia—a sack of fruits and vegetables. His teaching procedure was to pull an item out of the sack, name it, and have the class repeat. The teacher would reach into the sack, pull out a carrot, and say, "This is a carrot." Then, the class would repeat, "Carrot. Carrot." The teacher would grab an apple and say, "This is an apple." The class would then say in chorus, "Apple. Apple." The class went reasonably well, until this elder pulled a potato out of the sack and seriously, almost reverently (perhaps he was homesick), modeled, "This is a *spud*."

CONCLUSION

I began with a popular nineteenth century Mormon folk song, and a language-learning variation on it—"For none can learn a language like the Mormons do." It may be appropriate for me to end with another popular nineteenth century hymn—one that is no longer in our hymnbook. As early converts to the Church gathered to Zion, many of them had unrealistic expectations and were disappointed with what they found when they finally arrived in Nauvoo or Salt Lake City. To help them understand, Eliza R. Snow penned the words to "Think Not When You Gather to Zion" (Snow and Tullidge 1948). It goes like this:

Think not when you gather to Zion,
Your troubles and trials are through,
That nothing but comfort and pleasure
Are waiting in Zion for you:
No, no, 'tis designed as a furnace,
All substance, all textures to try,
To burn all the "wood, hay, and stubble,"
The gold from the dross purify.

If it's not too irreverent, I would like to conclude by sharing with you my own language-learning version of this song:

Think not when you learn a new language,
Your troubles and trials will be few,
That nothing but communicating
Is what without trouble you'll do.

No, no, languages can be tricky,
Their fine points will cause you to cry.
To become proficient will take effort
And you're bound to mess up by and
by.

REFERENCES

Benderson, Albert. 1983. Foreign languages in the schools. *Focus 12* [Published by Educational Testing Service], 1–12, 14–21, 24.

Fallows, James M. 1988. The world beyond Salt Lake City. *U. S. News and World Report*, May 2, 67.

Henrichsen, Lynn E. 1999. Foreign language training for LDS missionaries: Historical antecedents and foundations for current Church policies and institutions. *Proceedings of the Deseret Language and Linguistics Society 1999 Symposium*, ed. Alan D. Manning et al., 103–122.

Kaufman, William I., ed. 1980. *The Mormon pioneer songbook*. Bryn Mawr, PA: Theodore Presser.

Richards, F. D., and S. W. Richards. 1854–1886. *Journal of discourses*. Liverpool. Repr., Salt Lake City: Deseret Book, 1974.

Smith, Joseph Fielding, ed. 1976. *Teachings of the prophet Joseph Smith: Taken from his sermons and writings as they are found in the Documentary history and other publications of the Church and written or published in the days of the prophet's ministry*. Salt Lake City: Deseret Book.

Snow, Eliza R., and John Tullidge. 1948. Think not when you gather to Zion. In *Hymns of the Church of Jesus Christ of Latter-day Saints*, no. 21. Salt Lake City: Church of Jesus Christ of Latter-day Saints.

Wilds, Claudia P. 1975. The oral interview test. In *Testing language proficiency*, ed. Randall L. Jones and Bernard Spolsky. Arlington, VA: Center for Applied Linguistics, 29–44.