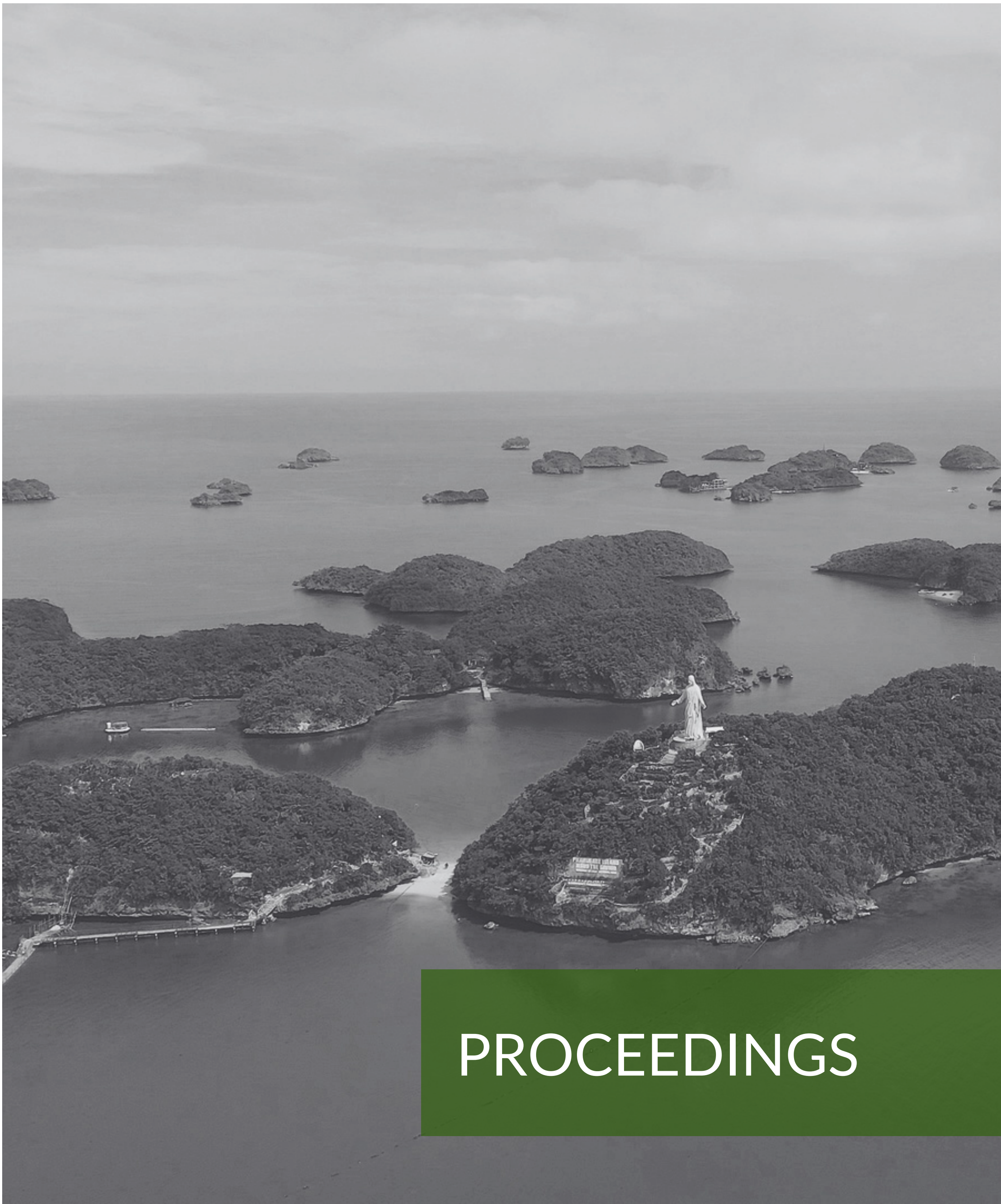
7TH PHILIPPINE LINGUISTICS CONGRESS

1995



PROCEEDINGS

Department of Linguistics
Palma Hall
Quirino Avenue
University of the Philippines
Diliman, Quezon City 1101
Philippines

https://linguistics.upd.edu.ph/

linguistics.upd@up.edu.ph

LINGUISTIC STATISTICS:
A Commentary

Jonathan Malicsi
Professor of Linguistics
UP

Even basic statistics can give the linguist and language teacher some form of security in numbers. It is simply the numerical representation, or reduction, of sets of data and the discovery of relationships among them. It rests on the notion of "countability"--basic to statistics is knowing what to count, how to count, and what for.

The last factor is the most important. At one dissertation defense I participated in, the proponent was rattling of a series of averages, yet, when asked what all the numbers mean, could not give a coherent answer. In another study, the statistician took over all the data and spewed out several sophisticated measurements which only left the project leader baffled. In fact, whether master's and doctoral students should be allowed to have their statisticians around to help them during their defense is, to my mind, objectionable.

And when all the numbers have been processed, the terrible question remains, "Now that I know that, what do I do?"

There are at least two basic ways by which statistics can be used in linguistics. One is in purely linguistic analysis, where statistics can be used to understand, even appreciate, linguistic phenomena at a scientific and, perhaps, aesthetic plane. The other is in applied linguistics, or cross-linguistic studies, where statistics can be used to serve as a guide to language teaching and language planning, such as in the identification and application of an intervention measure to address a language difficulty, or in deciding on language policy and formulating its effective implementation.

LINGUISTIC ANALYSIS:

Counting phonemes:

Perhaps you have wondered why Tagalog does not have a monosyllabic native word apart from the grammatical particles. This phenomenon can be better understood by referring to the relationship of size of phonemic inventory to the canonical form of morphemes--a smaller phonemic inventory entails longer morphemes.

This ultimately can be traced to the signalling and duality principles involved in language. For example, if you want to send a signal to another person using colored paper, you begin by agreeing on a code which arbitrarily assigns a meaning to a particular color. Thus, if you have seven colors available, you could signal seven meanings.

To expand the range of meanings, you could try expanding the range of colors, but you will be limited by their availability and effective perceptual differentiation: the person you are signalling to might not be able to distinguish between lilac and lavender, or between pearl white and dirty white. This could mean that your signalling system will not be able to expand beyond a small inventory of colors and meanings.

The duality principle offers a tremendous explosion of possibilities. Instead of a single-color signalling system, you could agree on a code where the position of the color also matters, thus, a red-white sequence, seen left to right by the viewer, could be different from a white-red sequence. Signalling alone, you can hold up a maximum of only two colored pieces of paper, but in this code, even seven colors in two positions give you 49 sequenced combinations, and this is certainly more efficient than trying to find 49 colors as you would have to if you could only signal with one color at a time.

This is the same duality principle in language--structure (or sequence) and system (or options per position in the sequence). At the syllable level, the structure would be a sequence of consonants and vowels, and the systems would be the inventories of consonants and vowels, i.e., the phonemic inventory itself. Given a language with 15 consonants and 5 vowels, that language could signal 5 meanings with just the 5 vowels in a V syllable, 75 meanings for the 75 sequences in the CV syllable, another 75 meanings for the VC syllable, 1125 for the CVC, etc.

But natural language evolved in such a way that some types of sequences are restricted, and many possible sequences even among the types allowed are meaningless. In Tagalog, the CV combinations *ba*, *ka* (variant of *ikaw*), *ha*, *na*, *nga* (variant of *nga?*), *pa* and *sa* are meaningful; some dialects use *ga* as a variant of *baga*; but *da*, *la*, *ma*, *ra*, *ta*, *wa* and *ya* are meaningless. The latter are sometimes called "junk" sequences. This simply means that the sequences used for signalling in natural language are but a small subset of the total number of possible sequences, and such limitation actually heightens the ability of the perceiver to distinguish a particular sequence he hears.

This is where the phonemic inventory becomes relevant. To establish a high level of perceptual distinctiveness for meaningful combinations, natural language creates a very wide range of possible combinations against which to set those used to signal meanings.

The 1,125 possibilities for the CVC syllable in Tagalog do not seem to be enough. Thus, even if *sak*, *sik* and *suk* can be differentiated easily, Tagalog repeats the syllable, to become *saksak*, *siksik* and *suksok* which, set against a wider field of 1,265,625 possible CVCCVC combinations, become more "informative," given that the degree of "informativeness" of an entity is directly proportional to the size of the field from which it derives.

In contrast to the Tagalog examples, the English *seek*, *sick*, *sack*, *suck*, and *sock* exist as individual words, because the 24 consonants and at least 9 vowels of English set these 5 words against a field of 5,184 CVC combinations. If your phonemic count in English has 11 vowels, then the CVC structure has 6,336 combinations.

Expect Sino-Tibetan languages, with around 75 phonemes, to have a large number of monosyllables, and South Pacific languages, with around 12 phonemes, to have very few. You may count phonemes to appreciate their number's effect on the morpheme size.

This does have practical value. When I was commissioned to render some native folk songs into English, it was obvious to me from the start that a long Philippine word, corresponding to many notes, may translate to an English monosyllable, e.g., Tag. *kapayapaan* to Eng. *peace*. What to do then with the rest of the notes? Do I slur the English monosyllable to stretch over all the notes of the original word? This would have made for a lot of wailing sounds. Thus, I had to write a rhythmic version in English of the sense communicated by the original line or phrase, e.g., Tag. *kapayapaan* as Eng. *a peaceful country*, or *the peace of heaven*, or *untroubled waters*. Sometimes, it became necessary, to produce a rhythmically correct English version, to add a few corollary ideas to the original, even if the resulting text runs dangerously close to becoming redundant.

Counting words:

"Of all the world's languages (which now number some 2700 [Guinness estimate is 5,000]), it [English] is arguably the richest in vocabulary. The compendious Oxford English Dictionary lists about 500,000 words [and the current Addendum project would cover about 60,000 more]; and a further half million technical and scientific terms remain uncatalogued. According to traditional estimates, neighbouring German has a vocabulary of about 185,000 words and French fewer than 100,000, including such Franglais as *le snacque-barre* and *le hit-parade*." (McCrum, p. 19)

Often, statistics like these are bandied about to emphasize the limitation of Filipino. After all, the Diksyunaryo ng Wikang Filipino has only about 31,250 word-bases (*batayang salita*), in contrast to the Merriam-Webster Dictionary, pocket edition, with around 60,000 entries. Thus, the natural reaction to this seems to be the various efforts to enlarge the vocabulary of Filipino by producing glossaries in various academic fields. If one discipline could contribute 3,000 words to Filipino, a concerted effort to mine 23 disciplines would mean an addition of 69,000 words added to the Diksyunaryo's 31,000, thus getting Filipino up to the French vocabulary size.

It should be noted, however, that ordinary written English is estimated to contain only about 10,000 words, and spoken English about 5,000 words. Thus, the big numbers quoted by dictionaries are way beyond the usual communication needs of a single person.

It should also be noted that the ability of a language to communicate information is not necessarily limited by the inventory. The amount of information that can be communicated by a limited inventory is enormous. The 850 words of Basic English listed by Ogden and Richards have been calculated to carry 12,425 meanings. Also, the duality principle must be considered. If one language has a word for a concept, that same concept could be rendered in several words put in collocation by another language. Of course, the language with less words would have to make up for the lack by producing longer structures, and the language with more words could at least claim ease and accuracy of expression as its advantages.

Still, the production of glossaries does not seem to be the answer, for after the glossaries have been compiled and published, the words therein will remain--therein. These words are not unlike all the German, French, and Spanish words in the dictionaries which we have never opened up after taking the corresponding foreign language course. The words lie buried--unknown and unused.

One should remember that dictionaries are based on existing texts; the 10th edition of the Merriam-Webster's Collegiate Dictionary was based on the company's collection of 14,500,000 citations culled from various publications, that the entries in the OED all came from citations culled from print. Thus, the enrichment or intellectualization of Filipino should be done through the production and propagation of texts, such as newspapers, journals, and textbooks. Scientific terms could be introduced in a Filipino newspaper's science page; after all, nearly all English newspapers have a science page once a week. The textbooks will ensure that the students learn and use the Filipino terms.

The glossaries being prepared now need not be published; instead, their printouts should serve as guides to the writers and editors. When several books on a particular subject shall have been produced and disseminated, then a glossary covering the terms used in these books should see print as it then becomes a tool for understanding the terms and their usages in a variety of texts.

Counting words and their frequency:

About 45 years ago, the linguist George Zipf took long texts in several languages, measured the frequency of all the words, ranked the words in order of frequency for each text, and discovered a "surprisingly tight link in all the languages between frequency and word rank: if the most common word in a text appeared 10,000 times, then the 10th most common word would appear roughly 1,000 times, and the 100th most common word roughly 100 times."

This should add to our appreciation of the symmetry of human language, akin to the appreciation of the Greeks for the diatonic scale of music as indicative of the ratios to be found in natural harmonics. Fine.

But Zipf's discovery has recently been reported to parallel what Boston researchers discovered about DNA. The puzzlement about DNA is that, thus far, it seems that all the information needed to keep us growing and alive is contained in only 3% of our DNA. The remaining 97% is called "junk DNA," yet now it appears to have some subtle linguistic feature.

As reported in Discover (April 1995) a group of Boston researchers "applied Zipf's analysis to junk DNA. They took DNA sequences from organisms as diverse as viruses and nematodes and divided the sequences into 'words' built from an 'alphabet' of four nucleotides--the individual links in the DNA chain. Since three nucleotides code for an amino acid--the building block of proteins--the shortest word was three nucleotides long. The longest word considered was only eight nucleotides long; the accuracy of the test depended on having enough sample words, and longer words would have made for too few samples.

"The researchers calculated the frequency in junk DNA of words built from all possible combinations of the nucleotides. To their surprise, they found that junk DNA obeyed Zipf's law: the most common nucleotide string was ten times more common than the tenth most common string, and so on."

But the work thus far has not been able to determine what this feature means.

"It may not mean much; junk DNA may still be junk. But if junk DNA has the features of a language, then it begins to seem more likely that it contains information--perhaps telling the cell when to make a protein and how much of it to make. 'We think we've found a language," says [Boston U physicist Rosario] Mantegna, 'but we don't know what it's saying.'"

Other possibilities in counting for linguistic analysis:

Counting positive and negative terms:

If you select any semantic field in any language, and count the number of positive and negative words therein, you will mostly likely get more negative words than positive words. There are more terms for ugliness than for beauty, more for negative emotions than positive ones, more for crime than for virtue. The Ten Commandments has more of the "thou shalt not" than of the "thou shalt." Research in this area can work out a model to explain this. My own hypothesis is that the cultural norm is generally defined in the negative; thus, the negative is expressed.

Counting words in a semantic field:

In understanding a culture through language, the lexical inventory may be regarded as reflective of the things and concepts that have acquired cultural significance. Thus, Tagalog has a word for male circumcision, but none for female circumcision, which is prevalent in Africa; Tagalog has *palay, bigas, kanin* while English only has *rice*, etc. Research in this area can determine the vocabulary size of a given semantic field, interpret that as an indication of its elaboration, and, in turn, interpret elaboration as an indication of level of cultural significance.

A few items are bothersome here. For example, if "hospitality" is indeed an outstanding trait of the Filipinos, why is there no specific native term for it, not even an idiomatic expression?

Counting cognates:

The counting of cognates, or lexicostatistics, is used to determine the closeness of relationship between languages. The first problem here is which words and how many words should go into the eliciting materials. If we take the estimate that ordinary spoken English has around 5,000 words, would 100 be representative enough, or should it be 1000, or more?

The second problem is in deciding on cognation. Since this is a largely personal judgment call, two researchers may end up having different figures, thus making the languages in question closer to, or farther apart from, each other, depending on who does the counting.

The third problem comes up when the percentage of cognate is translated into a figure for the time when the two languages under study bifurcated. Thus far, only the rate of change for Indo-European languages has been computed, based on visible changes in the forms of texts from ancient to modern times. If we consider that the writing system actually freezes a linguistic form and displaces it in time, it can be argued that the rate of change for Indo-European, which has libraries of written texts, should be slower than for American Indian or Malayo-Polynesian, which has hardly any extant manuscript from before the 16th century.

Counting rules:

In transformational-generative grammar, the rules needed to produce a sentence are formalized, made explicit, and sequenced, making it possible to count the number of rules involved in one type of construction.

The posited kernel sentence in English is the simple, declarative, active, positive sentence. This may be transformed into the passive, and both active and passive sentences may be transformed into negative ones.

Some psycholinguistic testing has already been done on these TG rules in English, demonstrating that, indeed, the processing time for the negative, passive simple sentence is a tad longer than that for the positive, passive sentence, which, in turn, takes a longer processing time than the positive, active sentence. This suggests that the rules--at least their number--may have "mental reality," that the TG formulations are, indeed, fair approximations of the workings of the brain in processing language, and not merely algorithmic exercises on paper.

CROSS-LINGUISTIC ANALYSIS (Applied Linguistics):

Counting respellings:

A few semesters ago, I was teaching a class how to do a survey. A group of Linguistics majors decided to find out the extent to which UP students respell loan words using Latin letters other than those in the Abakada, meaning *c, ch, f, j, q, v, x,* and *z*. Basically, they wanted to find out if respellings such as *cherman* and *saykoloji*--which approximate the English sound--are gaining acceptance or popularity outside the official communications of some UP units.

To get a number of words for testing, the students decided to cull from Filipino texts in the Philippine Collegian (because it is a student publication) and in some tabloids (because these target the masses, and reflect words commonly used in the print and broadcast media).

A straightforward procedure would have been to ask a representative sample of students to write down some English words dictated by the investigator. The pronunciation of the investigator, however, might cue the respondent into writing the word in the original English spelling. Thus, the group decided to prepare the eliciting words in English, put them in complete sentences, and ask the respondents to translate these sentences into Filipino. The respondents were then cued only to producing written forms in their own varieties of Filipino, thinking that this was a test of their ability to translate.

After tabulating the responses, the group found out that while the majority of the target words were rendered in Filipino, i.e., translated or respelled, nearly all respellings were limited to the Abakada--the English *ch* was still rendered as *ts*, the *j* as *dy*.

We can conclude from this small study that propagation efforts for the respelling of loan words has so far been limited in its reach. The students--and the masses--are necessarily influenced by the Filipino texts to which they are exposed. If a serious student or writer decides to consult a reference work, he will find Diksyunaryo ng Wikang Filipino spelling out *adyenda* (or *ahenda*) and *sikolohiya* (which psychologists themselves use).

The move to respell loan words in Filipino using the full complement of Latin letters has to find converts among editors and publishers of books and newspapers; without support from the mass media, it will remain an idiosyncrasy in official documents.

Counting loan words

"It is impossible--unless you go in for tortuous circumlocution--to write a modern English sentence without using a feast of Anglo-Saxon words. Computer analysis of the language has shown that the one hundred most common words in English are all of Anglo-Saxon origin. ... Anyone who speaks or writes English in the late twentieth century is using accents, words and grammar which, with several dramatic modifications, go all the way back to the Old English of the Anglo-Saxons." (McCrum, p. 61)

"When, in 1940, Winston Churchill wished to appeal to the hearts and minds of the English-speaking people it is probably no accident that he did so with the plain bareness for which Old English is noted: 'We shall fight on the beaches; we shall fight on the landing grounds, we shall fight in the fields and in the streets, we shall fight in the hills; we shall never surrender.' In this celebrated passage, only *surrender* is foreign--Norman-French." (McCrum, p. 62)

But as a postscript to the discussion above on vocabulary size, one should note, too, that a great number of the words in an English dictionary are actually loan words.

While the 100 most frequently used words in English are of Anglo-Saxon origin, the picture changes when one looks at the 1,000 most frequently used words, around 60% of which are Anglo-Saxon, and 30% are Norman French. And as one goes down the ladder of words listed according to frequency, more and more languages become represented.

The 10th edition of the Merriam-Webster's Collegiate Dictionary lists the following 75 ancient and modern languages as the source for its English entries, apart from Anglo-Saxon:

> Afrikaans, Albanian, American, American French, American Indian, American Spanish, Anglo-French, Arabic, Aramaic, Armenian, Australian, Avestan, Belgian, Bengali, Brazilian, Brazilian Portuguese , Breton, British, Bulgarian, Canadian, Canadian French, Catalan, Celtic, Chinese, Coptic, Cornish, Danish, Doric, Dutch, Egyptian, English, Finnish, Flemish, French, Frisian, Gaelic, German, Germanic, Gothic, Greek, Hebrew, Hittite, Hungarian, Icelandic, Indo-European, Irish, Italian, Japanese, Javanese, Latin, Latin, Lithuanian, Louisiana French, Low German, Mexican, Mexican Spanish, Norse, Norwegian, Pennsylvania German, Persian, Philippine Spanish, Polish, Provençal, Prussian, Romanian, Russian, Sanskrit, Scottish, Semitic, Slavonic, Swedish, Syriac, Tagalog, Tocharian, and Welsh.

In light of this characteristic of English, and, to a lesser extent, other world languages, the argument that Filipino becomes "watered down" by the introduction of loan words "holds no water." Loan words are but natural consequences of culture contact, and for as long as users of Filipino encounter concepts and things coming from other cultures, with their own terminologies, such users will naturally borrow the foreign term. Necessarily, loan words are adjusted to Filipino phonology. As to whether the loan words are to be respelled, this is a question for professional writers and editors to answer, for their own publications.

Counting letters, words and sentences:

One practical application of counting letters, words and sentences is in determining the readability of a particular text, i.e., establishing its readability index. For English texts, there are at least four such indices widely used by reading experts, and all of them are based on the following assumptions--that short words are easier to read than long words, that short sentences are easier than long sentences, and that the active sentence is easier than the passive sentence.

The Flesch Reading Ease (FRE) index takes the average number of syllables per word, and average number of words per sentence, and translates these to a score from 0 to 100, with "standard writing" at 60-70. The higher the score, the greater the number of people who can readily understand the document.

The Flesch-Kincaid Grade Level (FKGL) translates the FRE to grade-school level. Thus, if a text is determined to have an FKGL of 8.0, that means a typical eighth grader would understand the document. (The American 8th grader would roughly correspond to the Filipino 2nd year high school student.) An FKGL of 7 to 8 refers to "standard writing."

The Coleman-Liau Grade Level and Bormuth Grade Level use word length in characters and sentence length in words to determine grade level.

The Gunning "Fog" index uses the average number of words per sentence, and the number of difficult words--those with more than two syllables and are generally not used in casual conversation--in establishing the grade level.

For those of you who are computer literate, there is an easy way to establish the first four of these indices for a particular text--all of these are automatically produced at the end of the grammar check sub-routine of Microsoft Word for Windows. Thus, all you have to do is type in a text, run the grammar sub-routine, and you will get something like the following onscreen:

```
                                      FSI Memo        Time Essay
                                      re Vietnam      (1st 4 par)
Counts:
      words =                           439              366
      characters =                    2,295            1,989
      paragraphs =                       14                4
      sentences =                        17               20
Averages:
      Sentences/paragraph =               1.2              5
      Words/sentence =                   25.8             18.3
      Characters/word =                   5.2              5.4

Readability:
      Passive S =                        29%              18%
      Flesch RE =                        39.4             42.4
      Flesch-Kincaid GL =                13.9             11.4
      Coleman-Liau GL =                  14.4             14.0
      Bormuth GL =                       11.2             10.6
```

(The Fog index can be computerized but would mean developing a data bank of "words commonly used in casual conversation." For now, a researcher will have to exercise his own judgement as to which words are "difficult.")

Normally, the Word program would be giving only one set of numbers corresponding to the text you were checking. The above tabulation is for comparison. Obviously, the Flesch-Kincaid, Coleman-Liau, and Bormuth indices indicate differing grade levels owing to their different computational procedures.

Still, the statistics inform us that an official memo about a language program for Vietnamese diplomats seems to be a lot more difficult to read than a Time essay; while the essay can be easily read by an American senior high school student, the memo can be easily read by an American college sophomore.

What do these readability indices mean for us? We have to be cautious in interpreting the American grade level as corresponding on a one-to-one basis to the Philippine grade levels. After all, our assumption here is that our 14 years of schooling for a baccalaureate degree are qualitatively equivalent to the 16 years of American schooling for the same degree.

Further, we have to remember that English is a foreign language to us, and that our reading of an English text is affected by our background knowledge on the subject matter and our skills in the language.

Thus, while these readability indices could serve as a guide in choosing texts for various grade levels in English, or in editing "upwards" or "downwards" as in making a text suit the reading capabilities of a higher or lower grade level, we can not rely solely on these.

A subject for possible research is actually determining the reading abilities in English of Filipino students and professionals. For example, in a reading test of 25 items (part of a 170-item English proficiency test) based on an essay published in the Far Eastern Economic Review, with a Fog index of 12, 580 employees at the Department of Foreign Affairs and the Foreign Service Institute tested thus far, ranging from clerks to consuls, averaged only 51%, with the highest score at 92%, and the lowest at 8%. Nearly all of the respondents are college graduates. Does this mean that a Fog index of 12 should be recalibrated to 16 for Filipinos? Perhaps all the readability indices will have to be adjusted upwards for Filipino users of English.

Counting speakers:

Counting the "native speakers" of a language seems to be an unambiguous affair. All one has to do is ask the respondents what their "mother tongue" is. The US experience in the late '70s is enlightening. When schools were ordered to hire teachers who could use the "mother tongue" of a particular percentage of the pupils as an auxiliary language for teaching, in response to the civil rights clamor of the Hispanics (the right to be educated in one's own language), the school registrars had to get the language statistics from the pupils' parents. Inadvertently, the term "mother tongue" was understood by some respondents as the language of the mother, which was not necessarily the language in which the child was raised. Thus, in the Detroit area, for example, a school had to put out an advertisement for history teachers who are fluent in Ilocano, even if, as it found out later on, none of the pupils understood Ilocano--all were raised in English.

Counting the speakers of a language, especially in a multilingual country like ours, always has such social and political repercussions. The 1970s estimate that Cebuano speakers have begun outnumbering the Tagalog speakers, even if both were hovering about the 24% figure, shored up the arguments against Tagalog "linguistics imperialism." Tagalogs were quick to point out the estimated 55% who speak Filipino, as this was propagated more by the mass media than the required Filipino classes in school or the use of Filipino for teaching some courses. In the early '90s, a Cebuano edition of the Philippine Star was piloted, then withdrawn for lack of customers. Mere headcount, and estimates at that, is quite tricky, and should be related to other demographic data to be useful for any type of decision-making.

The following example shows an attempt to relate the number of speakers to voting behavior.

In last Sunday's issue of The Sunday Times, Dr Temario Rivera of UP's Political Science department reported the 1990 census estimate of the population size of eight Philippine languages, as follows:

Tagalog	= 18.9 M
Cebuano	= 14.7 M
Ilocano	= 5.8 M
Hiligaynon	= 5.8 M
Bicolano	= 3.5 M
Waray	= 2.4 M
Pampango	= 1.9 M
Pangasinense	= 1.2 M

Using the exit poll data gathered during election day, May 8, in a project jointly conducted by ABS-CBN, Inquirer, and the Social Weather Station, Dr Rivera concluded that:

"Ilocanos, Bicolanos and Ilonggos provided the highest percentages of support for their 'favored' candidates." The Ilocanos put Marcos in 1st rank, with 71% of votes; the Ilonggos put Santiago as 1st, with 63%; the Bicolanos put Roco as 1st, with 71%, Tatad 2nd, Arroyo 3rd, Honasan 4th (38%)

The Tagalogs voted mostly for non-Tagalogs--"there is evidence that the Tagalogs are the least solid voting constituency."

The Cebuanos also showed some bias for fellow Cebuanos: though Arroyo came 1st, with 50%, she was followed by Fernan as 2nd with 44%, and Osmeña as 3rd with 43%.

The Pampangos, too, ranked their fellow Pampango high: after Flavier, Arroyo came 2nd.

The Pangasinenses, having no homegrown senatorial candidate, went instead for Arroyo (63%) and Flavier (52%)

Because the statistics were presented in the text, quick comparisons were not easy to do, yet a tabular rendering of some of the statements, while showing the lack of support for Marcos, Honasan, and Santiago by the linguistic groups they do not belong to, also shows two apparent contradictions:

	Arroyo	Marcos	Honasan	Santiago	
	Rnk Per	Rnk Per	Rnk Per	Rnk Per	
Tagalog	1 (61%)	6 (38%)	11 (35%)	8 (38%)	[?]
Cebuano	1 (50%)	16 (19%)		14 (24%)	
Ilocano	2	1 (71%)	3 (57%)	11 (35%)	
Hiligaynon	2	17 (17%)	11 (28%)	1 (63%)	
Bicolano	3	17 (9%)	4 (38%)	13 (14%)	
Waray					
Pampango	2	16 (8%)		8 (22%)	
Pangasinense	1	7 (40%)	17 (7%)	18 (23%)	[?]

The source of the problems are the following statements on Marcos:

> "He [Marcos] had a respectable showing among Tagalog voters (number six with 38% of their exit votes) ...."
>
> vs.
>
> "She [Santiago] placed eighth among Tagalog (38%) ... voters."

and the following on Honasan:

> "His [Honasan's] achilles heel [sic] proved to be the Pangasinenses where he got a measly 7% at 17th place."
>
> vs.
>
> "[Santiago] trailed far behind as 18th placer among Pangasinenses with 23% of the exit votes."

Possibly, given the nature of the medium, these are but clerical errors, perhaps not of the type being encountered by the COMELEC.

Nevertheless, Dr Rivera's conclusion is that "based on the May 8 exit poll data, there is strong evidence for the claim that voting choices continue to be strongly influenced by one's language groups and identities. Pending the more systematic analysis of other voting factors such as socio-economic class, religion, party affiliation, gender, age, educational attainment, candidate performance, etc., the language affiliation factor continues to stand out as a very strong variable on Filipino's [sic] electoral choices."

Now, the front-runners in the 1998 presidential elections are both Tagalogs--Vice President Estrada and Senate President Angara. Will they take a Cebuano running mate, or unify the Tagalogs, or court the Cebuanos, Ilocanos, and Ilonggos as early as now, or will they play down the language factor and play up other factors such as popularity, intelligence, gender, and youth? Watch for their moves. And watch for the next estimate of language speakers.

Counting language users:

The number of "language users" covers first-language as well as second-language speakers, where the term "first language" refers to the language one is first exposed to and raised in up to age 7-10, while the term "second language" refers to all languages other than the first. Thus a quadrilingual would have, necessarily, one first language, a first "second language," a second "second language," and a third "second language." The terminology has to be exact, especially when one has to count.

At any rate, the number of language users is often used to argue for the dominance of a particular language over a given area. An example is the need to count the number of Filipino speakers over time, in order to measure its spread and the success, or failure, of propagation measures. Presumably, native Tagalog speakers will be counted as first language speakers, and all non-Tagalogs as second language speakers of Filipino. The objective, of course, is to determine what percentage of the population should be reflected in the statistics as Filipino users for its propagation to be considered a success (70%? 95%?), and later, what levels of fluency, or functional literacy, can be established for such users. Other sociolinguistic factors, such as status of users, types of texts, language varieties, etc., would also be important.

On the part of English, one can easily be overwhelmed by such statistics. McCrum writes:

> "About 350 million people use the English vocabulary as a mother tongue: about one-tenth of the world's population, scattered across every continent and surpassed, in numbers, though not in distribution, only by the speakers of the many varieties of Chinese. Three-quarters of the world's mail, and its telexes and cables, are in English. So are more than half the world's technical and scientific periodicals: it is the language of technology from Silicon Valley to Shanghai. English is the medium for 80 per cent of the information stored in the world's computers. Nearly half of all business deals in Europe are conducted in English. It is the language of sports and glamour: the official language of the Olympics and the Miss Universe competition. English is the official voice of the air, of the sea, and of Christianity: it is the ecumenical language of the World Council of Churches. Five of the largest broadcasting companies in the world (CBS, NBC, ABC, BBC, CBC) transmit in English to audiences that regularly exceed one hundred million."

One should also include, now, CNN and its worldwide reach, and influence on international newscasting in English.

All this point to the importance of English in the international scene, and can be used to argue that English should be taught in our schools to serve as our students' and graduates' international lingua franca--a line of thinking already enunciated in UP's language policy.

But this can not be used to argue for the use of English as a medium of instruction in courses other than the English language. The effectiveness of this teaching strategy is debatable; in fact, given the English language proficiency of teachers nationwide, there is reason to believe that using English as instructional language will only contribute to its deterioration.

Of what other value, then, are statistics on language users? For communication specialists, such numbers could spell out the size of their target clientele. A book publisher would, therefore, be more inclined to publish books in English as he could then tap the international market as well. The Filipino songs in English of Jose Mari Chan are popular even in Malaysia; Gary Valenciano performs his English songs in Las Vegas.

Yet, even if 99% of the world's population are English language users, the various native languages would still be dominant in their respective spheres, and should be nurtured.

Counting errors:

First, an "error" in language is a form which is not accordance with the prescribed rules and forms for the particular language variety involved, such rules and forms being based largely on the practice and preferences of language professionals, and leaders of society.

In the teaching of a foreign language, such forms and rules are prominent, such that the linguistic performance of the students, such as written compositions, are marked for errors. Since teachers may not have the same grasp of the rules and forms of the language, the same text may be marked differently by various teachers.

Thus, counting errors has its pitfalls. Still, such a tactic is useful for research on the foreign language teaching process, and as feedback to the students.

My own contribution to this tactic is the error index and its transmutation to a grade. At the Foreign Service Institute, a written composition submitted for grading in any of the various English classes is marked for errors, and, as a measure of the composition's language component (apart from its ideas and structure), the total number of errors is divided by the total number of clauses (not sentences). Easily, an excellent grade should go to an error index of 0. Because of the nature of official communications in the institution, I chose an error index of 0.4 for a passing mark, and 1.0 for a failing mark. The latter implies a probability of one error per clause, while the former implies a probability of two errors per five clauses.

The transmutation table for the error index I use at the FSI is as follows:

Error Index	Grade (UP)	Error Index	Grade (UP)	Error Index	Grade (UP)	Error Index	Grade (UP)
0.00	1.0						
0.02	1.1	0.22	2.1	0.43	3.1	0.73	4.1
0.04	1.2	0.24	2.2	0.46	3.2	0.76	4.2
0.06	1.3	0.26	2.3	0.49	3.3	0.79	4.3
0.08	1.4	0.28	2.4	0.52	3.4	0.82	4.4
0.10	1.5	0.30	2.5	0.55	3.5	0.85	4.5
0.12	1.6	0.32	2.6	0.58	3.6	0.88	4.6
0.14	1.7	0.34	2.7	0.61	3.7	0.91	4.7
0.16	1.8	0.36	2.8	0.64	3.8	0.94	4.8
0.18	1.9	0.38	2.9	0.67	3.9	0.97	4.9
0.20	2.0	0.40	3.0	0.70	4.0	1.00	5.0

This puts the grading of the language component of the composition on a more objective footing.

The effect on the students is interesting, in that they treat the error index as a challenge, and they make a game of improving their error index through one or two revisions of the same work. Incidentally, the errors are only identified in their papers, using a code that tells them which chapters of a grammar book to consult. I do not write in the correct forms. These are for them to find out from their grammar books and/or dictionaries and put in their revisions.

The compositions having been marked for errors, and catalogued, the frequency of errors may be determined. Such figures point out the relative difficulties of the students in particular forms or rules, and the teacher could respond by calibrating his materials to directly address such difficulties.

Counting responses to a language test:

item analysis for evaluation of the respondents:

In 1991 and 1992, the UP English Proficiency Test developed by a team headed by Prof Carmelita Ramirez (UP Department of English and Comparative Literature) was administered to students in all UP campuses registered in Communication I, as well as in the Filipino course Komunikasyon I, to gauge the preparedness of the UP freshmen for college English. A total of 4,688 students were tested in 1991, and 5,218 in 1992. This was not a sample, but practically the "universe," since the intent was really to test all Freshmen. The test had 145 items, and was administered within a 90-minute limit.

The scores are as follows:

	1991	1992
Highest score	91.03%	89.66%
Lowest score	10.34%	13.10%
Mean score	58.76%	56.89%
Q1	67.59%	65.52%
Q2 (Median)	58.62%	56.55%
Q3	50.34%	48.97%

It is fairly obvious from these scores that the 1992 freshmen scored lower than the 1991 freshmen. Even a 1.87% difference in mean scores is significant, considering the large number of respondents. This was the first objective indication of a gradual decline in English proficiency among high school graduates. In fact, in 1993 and 1994, a modified version of the same test, administered to all UP freshmen, had results that continued the downward line. This should really raise alarm bells in the English faculties of all schools, and should make college English teachers rethink the design of their English courses which, in light of the increasing use of Filipino as the collegiate instructional language, are actually the students' terminal courses in the language.

To get a better view of this language inadequacy, I computed the mean correct responses per category in the proficiency test, with the following results:

CATEGORY	NO. OF ITEMS	MEAN OF CORRECT RESPONSES	
		1991	1992
Paragraphing	3	77.91%	76.29%
Grammar	60	63.47%	60.58%
Reading	19	55.32%	55.87%
Mechanics	4	57.80%	55.31%
Idiom	29	55.87%	52.70%
Diction	28	54.89%	51.65%
(no error)	2	47.75%	44.72%

These figures demonstrate that the decrease in scores of the 1992 group was pervasive. Except for reading, all categories had lower scores in 1992 than in 1991.

In grammar, an examination of the items receiving a mean correct response of 50% or lower shows a mixture of elementary and advanced rules and forms.

Elementary items:	**1991**	**1992**
1 use of "these" as pronoun for a non-personal noun	18%	20%
2 correct use of "in which" vs. "wherein"	20%	20%
3 plural forms:		
3.1 plural of the non-count noun "equipment"	25%	24%

3.2	recognition of "media" as plural	32%	28%
3.3	plural of "advice" after the phrase "one of the best"	48%	46%
4	agreement of a verb with a distant subject	45%	35%
5	reference of the relative pronoun "which"	47%	48%
6	use of the verb "be" vs. "have" and "come," and agreement with subject	52%	49%
7	use of present perfect	55%	49%
8	agreement of subjective complement and subject	51%	45%

Advanced items:

1	parallel clauses and phrases:		
1.1	parallel clauses connected by "either ... or"	20%	15%
1.2	parallel clauses connected by "not only ... but also"	43%	46%
1.3	parallel verbal phrases connected by "and"	51%	45%
2	subjunctives:		
2.1	the "if" clause in the second conditional ("if we stood")	21%	22%
2.2	the subjunctive in a provision ("that every male 50% citizen undergo military training")	50%	
3	distinction among "thus," "although," "therefore," "but"	35%	31%
4	embedding an interrogative in a declarative, involving the shift in position of the modal or auxiliary verb ("how many people can the grand ballroom accommodate?" becoming "how many people the grand ballroom can accommodate")	46%	41%

These items underscore the deficiency in pre-collegiate teaching of English grammar.

In idioms, again the common Filipino errors were pervasive.

	Correct Form	1991	1992	Incorrect Form	1991	1992
1	"result in"	12%	8%	"result to"	82%	84%
2	"insight into"	n.d.	n.d.	"insight in"	69%	67%
3	"prefer to"	21%	16%	"prefer than"	67%	70%
4	"pretend that"	25%	24%	"pretend as if"	46%	50%
				"pretend as though"	26%	22%
5	"cope with"	28%	20%	"cope up with"	68%	76%
6	"we'd better change"	32%	33%	"we better change"	48%	43%
7	"pass on from mouth to mouth"	34%	30%	"pass from mouth to mouth"	36%	37%
8	"taken care of"	39%	37%	"taken cared of"	54%	55%
9	"the reason was that"	41%	40%	"the reason was because"	57%	57%
10	"prove of value"	42%	32%	"prove with value	34%	38%
11	"good at his job"	47%	41%	"good in his job"	25%	28%
				"good with his job"	18%	19%
12	"abide by regulations"	47%	42%	"abide with regulations"	39%	44%
13	[prove of value] "to" s.t.	49%	46%	"in" s.t.	18%	17%
				"on" s.t.	18%	20%
14	"may we request permission"	50%	51%	"may we request for" permission"	39%	37%
15	"interested in"	53%	50%	"interested with"	38%	40%

The form "insight into" comes from an item where the correct response is to mark "insight in" as erroneous. Thus, the test does not yield data on how many students consider "insight into" as correct.

In diction (or accuracy in word usage), 15 of the 28 items received scores of 50% or lower. Below are some of them. Since the items need some discussion, the data can not be presented neatly in tabular form; thus, for comparative purposes, the scores from the 1991 EPT are simply put in parentheses.

1 the non-redundant "enclosed is" got 22% (24%); "enclosed herewith is" and "enclosed herein is" got 35% (36%) and 30% (28%), respectively.

2 "raise" in reference to "standard of living" got 27% (32%); 59% (59%) chose "uplift". This item could be re-classified under idioms.

3 "[He was always] obsessed by [the problem of graft and corruption and all his life he fought to] eradicate [it]" got 35% (38%); 43% (45%) chose "anxious about ... abolish"

4 "[... the Philippines ... will determine what the answer] will be" got 42% (43%); 26% (24%) chose "would be" and 16% (19%) "should be." The context of this test item calls for the modal verb "will."

5 "[not] defensive" in reference to one's attitude about poor performance in class got 46% (47%); 29% (28%) chose "pretending"

6 that one can "infer" disappointment from a speech got 40% (50%); 28% (24%) chose "conclude" and 18% (15%) "imply." This test section was asking for the best option to complete the sentence.

7 one's ignorance and "prejudice" after the word "strong," paired off with "discriminate against" got 49% (56%).

If the UP English teacher is therefore getting the feeling that his teaching becomes more difficult with each batch of freshmen, the statistics above suggest the extent to which pre-collegiate English teaching has failed.

Other uses of statistics for language tests:

item analysis for test evaluation:

Item analysis is necessary to determine the relative ease or difficulty of the test items to a sample of the target universe. Items which receive a correct response score of 100% should be omitted from the test, as these are not discriminating. Other items receiving correct response scores between 90% and 99% must be examined one-by-one, to determine whether these are too easy for the target universes. On the other hand, items receiving correct response scores from zero to the probable score of random answers (e.g., 25% for a 4-option test item) should also be examined. Some of these may turn out to be defective--or the respondents are right but the answer key is wrong.

Here are some recent applications. Through an item analysis using 200 sophomores, the 170-item draft of the UP English Proficiency Test of 1991 was whittled down to 145 items. In another case, using a batch of UP Architecture students for an item analysis, 10 of the 120 items of a pilot Filipino proficiency test had to be disregarded in the computation of the respondents' scores. In a third case, an analysis of the scores of all those who took the admission examination of a prestigious management school led to the test's being replaced.

mean and standard deviation:

The mean and standard deviation of the test scores of a group could be used to classify them, inasmuch as, by definition, the area of the normal curve bound by the points "mean + 1 standard deviation" and "mean - 1 standard deviation" is about 68% of the total area. The scores here may be considered as indicating the "average group."

In some schools and UP campuses using the UP English Proficiency Test, the move is now afoot to use the test scores of incoming freshmen this June 1995 in classifying them, to make for more or less homogeneous classes, the easier for the English teacher to prepare himself, and his teaching materials. This necessitates the scheduling of 4 to 5 parallel sections, i.e., sections of freshman English in the same time period. The students whose test scores fall below the "average" range will be assigned to one section, while those whose test scores fall above the "average" range will be assigned to another section. The "average" group can then be distributed in the remaining sections.

At the Foreign Service Institute, the employees who scored higher than the "average" range are exempt from attending the English Language Review course, and may go direct to the Writing Memos and Letters course, or the Reading for Organizational Feedback course. Those who fall within the "average" are made to take the Review course, while those who fall below the "average" are deemed unsuited for the Institute's English language program, as currently designed.

Student's t test:

The mean scores of two or more groups are often compared to indicate which group is better or best, but because the mean is subject to the "tug-of-war" between high and low scores, and the size of the group, it is important to test the significance of the difference in means. The Student's t test is one such measurement. (The author, William Gosset, published his work at the turn of the century under the pseudonym *Student*.)

Applying this measurement to the results of the UP English Proficiency Test in 1992 showed how different the eight UP campuses are from each other, in terms of the English of their freshmen.

1992 FRESHMEN

	Total Participants	Highest Score	Lowest Score	Mean Score	Median Score
Baguio	344	74%	32%	52%	52%
San Fernando	208	77%	30%	51%	50%
Diliman	1,639	90%	18%	63%	64%
Manila	588	88%	26%	65%	66%
Los Banos	1,336	83%	34%	56%	54%
Iloilo	637	84%	18%	48%	47%
Tacloban	277	74%	13%	42%	41%
Cebu	189	77%	20%	53%	53%

The clustering of the campuses, based on a t test of the above, is as follows (at confidence level 0.001):

1st - Manila, Diliman
2nd - Los Banos
3rd - Cebu, Baguio, San Fernando
4th - Iloilo
5th - Tacloban

Manila and Diliman students have generally the highest proficiency levels, and Tacloban students the lowest. This points to the need to calibrate instructional materials to cover the range of levels in the campuses, and for the English faculties to explore the teaching strategies appropriate to their respective campuses. What is effective in Diliman may work for Manila, but may not do so for Tacloban.

As this seems to be the trend in incoming UP freshmen, the challenge is to formulate the kind of intervention--a combination of teacher training, teaching/learning materials, and protocol for teaching English both inside and outside class hours--which should succeed first in the campuses with lower proficiency levels.

CONCLUSION:

Number-crunching is not linguistic analysis. While it offers a degree of scientific security in drawing generalizations and formulating hypotheses about language and language-related phenomena, it is still only a tool for analysis, and must be guided by solid linguistic theorizing.

Even then, it certainly lends credibility to linguistic conclusions, allows the leaders of society to make informed decisions on language policy and its implementation, gives language professionals--teachers, writers, editors--a firmer grasp on their media, and, on the whole, opens up the mind to other relationships hitherto unexplored.

Paper presented at the 7th Congress in Philippine Linguistics, 18 May 1995, Faculty Center. Sponsored by the UP Department of Linguistics.

BIBLIOGRAPHY:

Carter, Ronald. 1987. Vocabulary. London: Allen and Unwin.

Diksyunaryo ng Wikang Filipino. 1989. Mandaluyong: Linangan ng mga Wika sa Pilipinas.

Glazier, Stephen. 1992. Word Menu. New York: Random House.

Guinness Book of World Records. 1994. New York: Bantam.

"Junk Debunked." 1995. Breakthroughs section, Discover (Apr 95), pp. 16-18.

Malicsi, Jonathan. 1992. The 1991 English Proficiency Test Report. Quezon City: UP English Language Project.

_____. 1993. The 1992 English Proficiency Test Report. Quezon City: UP English Language Project.

_____. 1994. The U.P. English Manual. Quezon City: UP English Language Project.

McArthur, Tom. 1992. The Oxford Companion to the English Language. Oxford: Oxford University Press.

McCrum, Robert, William Cran, and Robert MacNeil. 1986. The Story of English. New York: Penguin.

Merriam-Webster's Collegiate Dictionary. 1993. 10th Edition. Springfield, Mass.: Merriam-Webster.

Nation, I. S. P. 1983. Teaching and Learning Vocabulary. English Language Institute: U of Wellington.

Rivera, Temario C. 1995. "The language factor in the elections," The Sunday Chronicle (14 May 95), p. 6.

Pangasinan Anaphors: A Preliminary Study[1]

Nieves B. Epistola
University of the Philippines

1. Introduction

2. Anaphors and Anaphora

3. Anaphora and Binding

4. <u>Three Principles of Binding</u>

4.1. <u>Principle A</u> : an anaphor is A-bound in its governing category;

4.2. <u>Principle B</u> : a pronominal is A-free in its governing category;

4.3. <u>Principle C</u> : an R-expression is A-free (everywhere).
(Note: This paper is not concerned with Principle C.)

5. Application of the Binding Theory

5.1. In Tuki (Biloa, 1991)

(1) a. vatu$_i$ va m(u) ena vamwamate$_i$ na ngene
men SM pt see themselves in mirror[2]
'the men$_i$ saw themselves$_i$ in the mirror'

b. vatu$_i$ va mu dza ee /vamwamate$_i$ va n(u) aram/
men SM pt say that themselves SM ft come
*'the men$_i$ said that /themselves$_i$ would come/'

c. vatu$_i$ va mu dza ee /ngu m(u) ena vamwamate$_i$/
men SM pt say that I pt see themselves
*'the men$_i$ said that /I saw themselves$_i$/'

[1] Paper written for the Sixth Philippine Linguistics Congress held at the University of the Philippines, 02-04 December 1992 (Outline)

[2] Symbols used in the glosses: SM=subject marker; pt=past tense; ft= future tense.

(2) a. */vatu$_i$ va mu w(u)$_i$ ena/

men SM pt them see

*'the men$_i$ saw them$_i$'

b. vatu$_i$ va mu dza ee / /e/$_i$ va n(u) aram/

men SM pt say that SM ft come

'the men$_i$ said that they$_i$ would come'

c. vadzu$_i$ va mu dza ee /Mbara a mu w(u)$_i$ ena/

children SM pt say that Mbara SM pt them see

'the children$_i$ said that Mbara saw them$_i$'

5.2. **In Pangasinan** (cf Amurrio, 1970; Benton, 1971 a, b, c)

5.2.1 Reflexive pronouns

(3) a. pinatey toy inkasikato.

-in-+ patey to$_i$ + /-y inka-+ sikato /$_i$

p kill s/he OM RP s/he [3]

's/he$_i$ killed /herself/himself /$_i$

b. pinatey toy laman ton dili

-in- + patey to + /-y laman to + -n dili/

p kill s/he OM body her/his RM self

's/he$_i$ killed /herself/himself/$_i$

c. *pinatey koy laman ton dili

-in- + patey ko + /-y laman to + -n dili/

p kill I OM body her/his RM self

*'I killed /herself/himself/'

[3] Symbols used in Pangasinan glosses: p =past; OM =object marker; RP= reflexive prefix; RM= reflexive marker; i=co-index reference ; lp= linking particle ; CP= causative prefix ; OM$_{pl}$=object marker plural

5.2.2. reciprocal anaphors

(4) a. arumog da$_i$ la /so balang sakey ed sikaran arum/$_i$ya atawtaw/ /

a- + domog da$_i$ la /so balang sakey ed

p find they already OM each one of

bawat isa sa

sikara + -n arum/$_i$ /ya a- + tawtaw/ /

kanila sila

they OM others lp p lose

I already found each of those who were lost.

b. *?arumog da$_i$la/so balang sakey/ /ed sarayan arum/$_i$ [ya atawtaw/ / /

a- + domog da$_i$ la /so balang sakey/ed

p find they already OM each one of

saraya + -n arum/$_i$ ya a- + tawtaw/ / /

these OM others lp p lose

c. ?*arumog ko$_i$la/so balang sakey/ /ed sikaran arum/$_i$ [ya atawtaw/ / /

a- + domog ko$_i$ la /so balang sakey/ ed

p find I already OM each one of

sikara + -n arum/$_i$ya a- + tawtaw/ / /

5.2.3. NP-traces

(5) a. nen inatey si Berto$_i$ inpakabat mi la /$_{NP}$so inpatey to$_i$ /

ed saray kanayon to$_i$

nen in-+ patey si Berto$_i$ inpa-+ kabat mi la

Nong namatay si Berto sinabihan na namin

when p die SM Berto CP know we already

/$_{NP}$so in- + patey to$_i$ / ed saray kanayon to$_i$

na namatay siya ang mga kamaganak nya.

OM p die he to OMpl relatives his

[illegible]

'wnen Berto died, we did inform his relatives about his death"

b. nen inatey si Berto$_i$ inpakabat mi la t$_{NP}$ ed saray kanayon to$_i$

nen in-+patey si Berto$_i$ inpa-kabat mi la/trace ed saray kanayon to

'when Berto died we already informed his relatives

c. ? inpakabat mi lad kanayon nen Berto

inpa-+kabat mi la + -d kanayon nen Berto

CP know we already to relatives of Berto

6. Conclusion