file Papers Rosenberg

Lingvisticæ Investigationes III:2. 323-339 (1979).

THE HARDEST NATURAL LANGUAGES

ARNOLD L. ROSENBERG
IBM Research Center

> "Come, let us go down and there confound their language, that they may not understand one another's speech."
>
> *Genesis* xi. 7

1 Introduction

In the mid-1950s, Noam Chomsky 1956, 1957 revolutionized the linguistic world by introducing mathematical models suitable for studying the process of generating and checking syntactic validity of purported sentences in a language. Little time elapsed before workers in the then-embryonic field of theoretical computer science began studying Chomsky's classes of so-called *formal* languages (see, *e.g.*, Bar-Hillel, Perles, and Shamir 1961) which promised to have more relevance to the study of computer languages than of natural languages. Almost immediately, the theoretical computer scientists began discovering intimate relationships between Chomsky's classes and certain variants of Alan M. Turing's (Turing 1936) mathematical models for computational processes (Evey 1963, Schützenberger 1963, Kuroda 1964). The two fields of formal language theory and so-called automata theory were thereby united by a bond that has yet to be sundered. As time went by, researchers began to investigate the computational complexities – *e.g.*, time and space requirements – of the computational processes associated with classes of formal languages. In the course of such investigations, certain languages in a given class were on occasion found to be maximal in consumption of resources among languages in the class; that is, in a certain precise sense, these languages were the *hardest* ones in the class. To cite but two such results (many of which are surveyed by Stockmeyer 1977), A. R. Meyer and L. J. Stockmeyer 1972 discovered a language that was a hardest one in Chomsky's class of so-called context-sensitive languages; and S. A. Greibach 1973 found a language that, in a somewhat different sense

of hardness, was a hardest one in Chomsky's class of so-called context-free languages. Our purpose here is to return to Chomsky's original intent of studying natural languages, but to bring to our investigation the more recent complexity-oriented point of view. We propose to set out on a quest for the hardest natural languages.

The nature of our quest puts us at a disadvantage relative to the seekers of hard formal languages. Whereas these formal linguists obtain their results by exploiting the syntactic "encoding" abilities of their respective classes of languages, we are faced with our relative ignorance concerning the syntaxes of natural languages, coupled with the likelihood that all natural languages are equivalent in expressive power.

> [*Aside*: We have, of course, no assurance that all natural languages have similar expressive powers, especially if we are willing to countenance the languages of antiquity. A potential counterexample issues from one of our informants (our informants, to all of whom we owe a deep debt of gratitude, are listed at the end of the paper) who tells us that the entire known fragment of the ancient Medean language comprises the single word
>
> *σπακα*
>
> meaning "horse." If the remainder of the Medean language is neither syntactically nor semantically excessively more complicated than the already known fragment, then Medean may, in fact, be a candidate for the "easiest" natural language. Indeed, Medean may even be the language sought by K. Gibran 1926 when he observed,
>
> > "We shall never understand one another until we reduce the language to seven words."]

At any rate, it seems clear that we do not have recourse here to the tools used by the formal linguists. How, then, can we hope to find a hardest natural language? R. A. DeMillo, R. J. Lipton, and A. J. Perlis 1977, when quoted out of context, give us the needed hint:

> "A large measure of credit for the continued success and growth of *mathematics* belongs to the *social* mechanism of "proving" theorems; . . . mathematics is . . . an ongoing social process"

Here, then, is the solution to our problem: we shall determine the hardest natural language by heeding the pronouncements of the speakers of natural

languages. (Relative to this course of action, we have, for once, an advantage over students of formal languages.) The authority vested in the social consensus we shall rely on does not reside solely in the speculations of DeMillo, Lipton, and Perlis, as interpreted by us. A much higher Authority bolsters our confidence in our proposed course of action. In the words of the noted scholar Alcuin 800,

"*Vox populi, vox Dei.*"

The people thus speak with high authority indeed. It remains unclear, though, in which domains Alcuin's *dictum* endows the People with this authority: there may be areas of inquiry of such little interest to God that His pronouncements and, all the moreso, those of His surrogates do not merit immediate acceptance. A glance at the Bible, though, assures us that natural languages, and particularly the unintelligibility of one language to the speaker of another, have been among His direct concerns since early times: we find, for instance, in *Genesis* xi. 7,

> "Come, let us go down and there confound their language, that they may not understand one another's speech."

Although we personally find the foregoing argument lending Absolute Authority to the Voice of the People in our investigation, we recognize that there are men of honor who might still question the impact of *vox populi*. As but one example, we find the renowned man of letters A. Pope pondering in Pope 1733-1738, *Epistle* I *bk.* I.

> "The voice of the people is odd,
> It is, and it is not, the voice of God"

While we can (and must) ignore completely those who scoff at the Voice of the People, we must treat with sympathy and concern men of good will, such as Mr. Pope, who have sincere reservations about the authority of the People's voice. As a concession to these ambivalent souls, we shall, whenever the means at our disposal permit us, supplement the pronouncements of the laity with supporting citations from notables of yore and/or of antiquity. These voices from the past must convince all doubters and, indeed, quiet even the most strident cynic, for as Bernard of Chartres reminds us through the voice of his student John of Salisbury (12th cent.), as translated and condensed by G. Sarton 1935, as reported by R. Merton 1965:37ff.:

> "In comparison with the ancients, we stand like dwarfs on the shoulders of giants"

Thus armed with a technique for verifying the "hardness" of one language relative to another, which will convince all men of good will, we turn to the only remaining methodological precursor to our investigation. While we shall have faith in the pronouncements of the People and, even more, in the writings of the notables of days past (really, of course, we mean years or even centuries past), what shall we expect these sources to say to us? Shall we infer, for instance, that Turkish is a harder language than Persian from the somewhat arcane Persian proverb,

> Arabic is a language, Persian is a sweetmeat, and Turkish is an art.

Decidedly not, for such "free-form" proverbs are all too often poorly translated and/or subject to alternative interpretations. (Could the cited proverb, for instance, be referring to the grammatical looseness of Turkish rather than its intricacy? Not knowing any Turkish, I have no basis for interpreting such a proverb). We can also not trust the individual opinion of self-styled experts such as C. C. Colton 1820 who made the (in context, disparaging) observation,

> "A literary (*sic*) Chinese must spend half his life in acquiring a thorough knowledge of [the Chinese language]"

No, we must be certain that it is the Voice of the People that we are hearing and that they are clearly telling us that language A is harder than language B. Our formal relation of linguistic difficulty is suggested by no less august a personage than the Bard himself (Shakespeare 1599: I, ii, 287) who put these words in the mouth of Casca:

> "But, for my own part, it was Greek to me"

Here we see an expression in common usage even after almost four centuries – hence clearly in the mouths of the People – whose meaning is not a matter of dispute (see, for example, the eighth meaning of "Greek" in *Webster's New International Dictionary of the English Language*, 2nd ed., unabridged, Springfield, MA: G. & C. Merriam Co., 1958). We generalize from this example to define formally,

the language A *is harder than* the language B if
(1) in language B, the assertion

"It is A (to me)"
or some minor syntactic variant thereof (to accommodate local tastes) is synonymous with the assertion
"It is unintelligible (to me)"
or
(2) there is a language C such that
(i) A is harder than C via clause (1)
(ii) C is harder than B

While we have clearly made an airtight case for the reasonableness of our *modus operandi*, we recognize that some specialists in the areas of linguistics and cultural anthropology may be so upset at the simplicity and neatness of our framework and at our incursion into their bailiwicks that they may cast stones at our impregnable edifice. To such *acini-acerbi*-ists, we say, in the too-eloquent-to-be-paraphrased words of Merton 1965:175,

> "I often prefer to rely on my own feeble resources rather than to turn invariably to the scholars on every side who could set me straight when ignorance threatens to lead me astray. . . . I regard an original error as better than a borrowed truth."

With this dogma thus established, to those who still detract from this scholarly work, I say with the immortal words of Edward III (1349),

"*Honi soit qui mal y pense*"

2 The Relative Difficulties of Languages

With the assistance of good-willed helpful informants, listed in their own section among our references, and with the information we have been able to glean from numerous reference books, also listed in a designated section of our list of references, we have studied at great length (and, where feasible, depth) the relation

is harder than

on pairs of natural languages. To our great joy and surprise, and to that of our informants – we made no effort, however, to elicit the reactions of our original sources, to whom a section of our list of references is devoted, nor of the authors of our several reference books, nor, in the case of the prior death or in-

capacity of the aforementioned, of their legal heirs, assignees, agents, or r presentatives – the studied relation is almost a partial order, that is a "one-wa relation. Only one cycle has been found, namely, the Turkish-Arabic-Persia Turkish cycle depicted in Figure 3. And the largest collection of mutuall related languages, including most of the Western languages, is a partial order!

[The results of our study are encapsulated in the attached four figures, thre of which the reader will recognize as collectively depicting a partial orde (The multiplicity of figures is used only for expositional and presentation convenience).]

Indeed, considering the fact that our searches have covered what my admitted ly chauvinized Western ear would consider all of the major languages, and ther some, we have the temerity to assert that all the "dead ends" in our relatio are languages that are the *hardest natural languages.*

Although we have found numerous hardest languages in our quest, we must acknowledge the special position of Chinese among hardest languages. If we were backed into a corner and forced to select a single language that deserved the designation "hardest," then, in terms of popular consensus, of geographical consensus, and of cultural consensus – all of which are inferrable by comparing Figure 1 with our other figures – Chinese would be the hands-down winner.

The remainder of this section is dedicated to augmenting the information in our figures by citing the exact aphorism when such is available to us and, in a very few cases, to trace the (various versions of) the aphorism back to their commonly-acknowledged sources.

[Before beginning this elaboration-*cum*-elucidation, let me make explicit a point you may have missed several sentences back. It was anything but a slip of the pen that led to my apprising you of my *chauvinized* Western ear. I have chosen to confess to you, my reader, that I suffer, to some extent at least, from that ubiquitous malady one might term aural chauvinism: the languages I appreciate most are those I know best. But – and here is where my contribution to my native language must be acknowledged – I ask if the guilt of this linguistic parochialism must be borne by me alone without ascribing to my forebears, teachers, and environment their share of the blame. No! I refuse to refer to my "chauvinistic" Western ear, for one might infer from that phrase that I have formed and fashioned my own earmuffs. I insist with Pope 1733: Ep. i, line 150 that

"Just as the twig is bent, the tree's inclin'd"

so that those chauvinizers who bent my linguistic twig must stand with me to receive their share of the blame. I shall do no less when my children call me to task for chauvinizing them.]

English: Not surprisingly, our most extensive documentation centers around the author's native language. Speakers of English seem to bewail the difficulties of both Greek and (Double or High) Dutch, the latter seeming to prevail in Britain. As noted in Section 1, our present reference to Greek seems to derive from the Bard of Avon, but he deserves at most credit for its form, not its substance.

> "But, for my own part, it was Greek to me"
> (Shakespeare 1599: I, ii, 287)

An earlier rendering took the form,

> "This geare is Greeke to me"
> (Gascoigne 1573: I)

On the "Dutch" side of the coin, we find

> "Why, 'twas just all as one as High Dutch"
> (Dibdin 1789: II)

And, lest their be warfare among the Greek-o-phobes and Dutch-o-phobes (please forgive the word forms!), we find the conciliatory

> "The preacher preached double Dutch or Greek or something of the sort"
> (Spurgeon 1879: vol. 25, 297)

Of course, this olive branch 'twixt the Greek-ery and Dutch-ery (pardon, once more) is not likely to thrill the clerg-ery.

Although English is unique in its exaltation (in our relation) of Dutch, it is not unique in its elevation of Greek.

Afrikaans:

Dis Grieks vir my

(All unattributed citations come from dictionaries or from informants).

Latin:

Graecum est; non potest legi
(It's Greek; it can't be read)
(Anonymous medieval saying)

Portuguese:

É grego para mim

Polish:

To jest dla mnie greka

The Poles honor not only the Greeks in this way, but also the Chinese, who thus appear for the first time, but decidedly not the last, in our saga.

To jest dla mnie chińszczyzna

It is now time to let some of the "second plateau" languages have their say. The time is ripe, since both of the thus far exalted languages join with the Poles in giving at least part of their adulation to the Chinese.

Dutch:

The Dutch exalt *three* languages (a record in our study). Included here are Chinese (of course), Spanish (a custom the Dutch share with their neighbor Germans), and, most commonly of all, Latin (thus raising that language to the second plateau):

Dat is Latijns voor mij

Most exalters of Chinese do so single-tonguedly.

Greek:

μοῦ φαίνεται κινέζικα
(It appears Chinese to me)

Hebrew:

Nishma c'moh sinit
(It sounds like Chinese)

Romanian:

Parca e Chineza
(It looks like Chinese)

Russian:

Eto dlya menya kitaiskaya gramota
(It's a Chinese document to me)

Serbo-Croatian:

To je za mene kineski
(It's Chinese to me)

A number of other languages exalt only Chinese, but do so in terms that have resisted our discovery efforts:
Estonian, *Flemish*, *Hungarian*, *Swiss-German*, and *Tagalog* (a/k/a *Filipino*).

Although the Dutch set a record for the number of languages beside Chinese that they consider difficult, they are decidedly not alone in having some alternative.

Finnish:

Onpas Kiinalainen juttu
(What a Chinese thing)

The Finns seem not to understand the Hebrew language either (which as we just saw a few lines back merely affords the poor Finns a way-station on the road to not understanding Chinese).

Se on minulle täyttä hepreaa
(It is totally Hebrew to me)

In fact, I am informed that this latter phrase exalting Hebrew connotes an even higher degree of unintelligibility than does the former reference to Chinese. Were our ground-rules for adjudging relative difficulties of languages not firmly and immutably fixed, this information would raise the specter of a cycle in this main component of our relation, if only from the vantage point of a third party. Fortunately for all lovers of order (and of orders, even partial ones), our firmly established constitution has no provision for amendment and, so, our partial order remains, lifting its branching ramifications toward Heaven. (We shall see later that this reference to Heaven is not entirely out of place here).

The Finns are not alone in appreciating the complexities of the Hebrew language.

French:

"C'est de l'hébreu pour moi"
(Molière 1653: III)

Completing Figure 1, we find the Czechs and Germans looking up to the Spanish who in turn are looking up toward the Chinese.

Czech:

To je pro mne španělská vesnice
(It's a Spanish village to me)

German:

Das kommt mir spanish vor
(That seems like Spanish to me)
Es waren mir spanische Dörfer
(That would be Spanish villages to me)

["Wait!" I hear you saying, dear reader! "Whence come these *villages* into our discussion?" I am relying here on one of my older and more historically minded informants, a native German, who maintains that these village-oriented expressions and the familiar (and popular)

Es waren mir böhmische Dörfer

arise from the unintelligibility (to the speaker) of the *names* of the mentioned villages. The references are, thus, linguistic ones and, hence, merit a place – indeed one of honor for the color they bring with them – in our study.]

Back to the level of prose, we find the Spaniards saying:

Spanish:

Para mi es chino

Our elaboration on Figure 2 is even shorter than that on Figure 1 for two reasons. First we have had to rely on informants rather than on literary sources even more heavily with this second group of interrelated languages than with the first group. Second, we find the presence of the cycle Turkish-Arabic-Persian-Turkish so painful to contemplate, so destructive of what would otherwise be an ordered universe, so anomalous in the context of the other relationships we have uncovered, that we have had to muster every iota of scholarly integrity at our disposal to resist the temptation to sweep under the carpet this offensive cycle. Indeed, we cannot resist, even after mustering all this integrity, to relate a circumstance that somewhat extenuates one of the links in this cycle, but this will await the proper moment. We begin by noting that the Italians elevate the Turks in our relation.

Italian:

Questo è turco per me

And I am informed that Italians exalt Arabic also. But, even if they didn't, the position of Arabic above Italian would be assured by the Turks.

Turkish:

Anladimsa arab olayim
(If I understood that, I'd be an Arab)

The Arabs, I am told, have difficulty with both Persian and Hindi. We do not have access to the Hindi-elevating phrase; but the Persian-oriented expression goes as follows.

Arabic:

Kalam ajami
(It's Persian to me)

[Here is where we must expose a weakness in our edifice. The word *ajam*, according to my informants, originally connoted to the Arabs a person incapable of intelligible speech. As the peninsular Arabs came in contact with the Persians, *ajami* came to mean the Persian language. Hence, in sharp contrast to the foreign-language-becoming-synonymous-with-unintelligibility origins of our other phrases, this Arabic phrase followed the reverse gestatory path. This anomaly looms all the larger since Arabic-Persian is one of the links in our only cycle.]

I have dawdled over esoterica long enough. I can no longer in good conscience keep from your eyes, dear reader, the dreaded back-reference, that cruel hint of perversity in nature, that grim reminder of ununderstanding and misunderstandings that have plagued our race since the days of Babel. (Oh, how I put off the fateful moment!) Speakers of Persian have trouble understanding Turkish!

Persian:

Turki gofti?
(Did you say it in Turkish?)

The remainders of Figures 2, 3, and perforce 4 (which displays only unrelationships) rely solely on the words of informants, with nary a dictionary entry nor a citation to bolster them. Perhaps you, dear reader, can do me – nay,

the scholarly world – the service of supplying some reference or some quotation that may flit before your eyes. Our appreciation will know no bounds.

Before leaving you, dear reader, let me share with you three gems whose disposition has troubled me. I present you with three expressions that do not quite fit in the prescribed framework but which come so close that they cry out to be mentioned.

Bulgarian:

Tova za mene sa ieroglifi
(It's hieroglyphics to me)

How is one to interpret this reference? Are the "hieroglyphics" mentioned a reference to some "holy engravings," an interpretation suggested by etymological considerations? Are these "hieroglyphics" intended to refer to any writing of alien form? Or, are the Bulgarians really exalting here some ancient language in our partial order? If so, would it be Egyptian? Hittite? Mayan? The number and diverse characters of the possible antecedents of the term "hieroglyphics" here makes this expression of the Bulgarians a strong candidate for the most unintelligible connoter of unintelligibility.

Chinese: Having seen so many turning to Chinese as their symbol of unintelligibility, one must wonder where the Chinese turn. To Heaven! The Chinese analog of our long-studied expression is (roughly translated),

It's Heavenly script to me

[Going back momentarily to the Bulgarians, a possible Greek rendering of "Heavenly script" would be "hierograph," a short hop from the Bulgarian "ieroglifi" for "hieroglyphics." Does this coincidence strengthen our initial putative interpretation of the Bulgarian expression?]

There is no ambiguity in this Chinese expression, but we are nonetheless stymied in how to incorporate it into our study. We have been quite unable to obtain certification that the mentioned script is the written version of a natural language. Lacking such certification, we have used scholarly discretion to include this expression as an oddity rather than as a legitimate entry in our relation-catalog. We have thus opted to preserve for Chinese its exalted and unapproachable position among languages.

Danish:

Det er det rene volapyk for mig
(It's pure Volapük to me)

I must utter a decidedly nonscholarly "Wow!" at this expression. As you certainly know, dear reader, Volapük is an *artificial* language promulgated at the end of the 19th century (Schleyer 1887) and subsequently eclipsed by competitors such as Esperanto 1887. Thus the Danes exalt – in the sense of our partial order at least – a language that is at once artificial and unsuccessful. But there is a piece of history beneath the surface here, for Volapük has been all but forgotten *because of its excessive complication.* Perhaps the Danes have picked the best expression of all, referring to a language that is so hard that it has been abandoned.

In fact, maybe the Danes have chosen so well, that this is the place to stop.

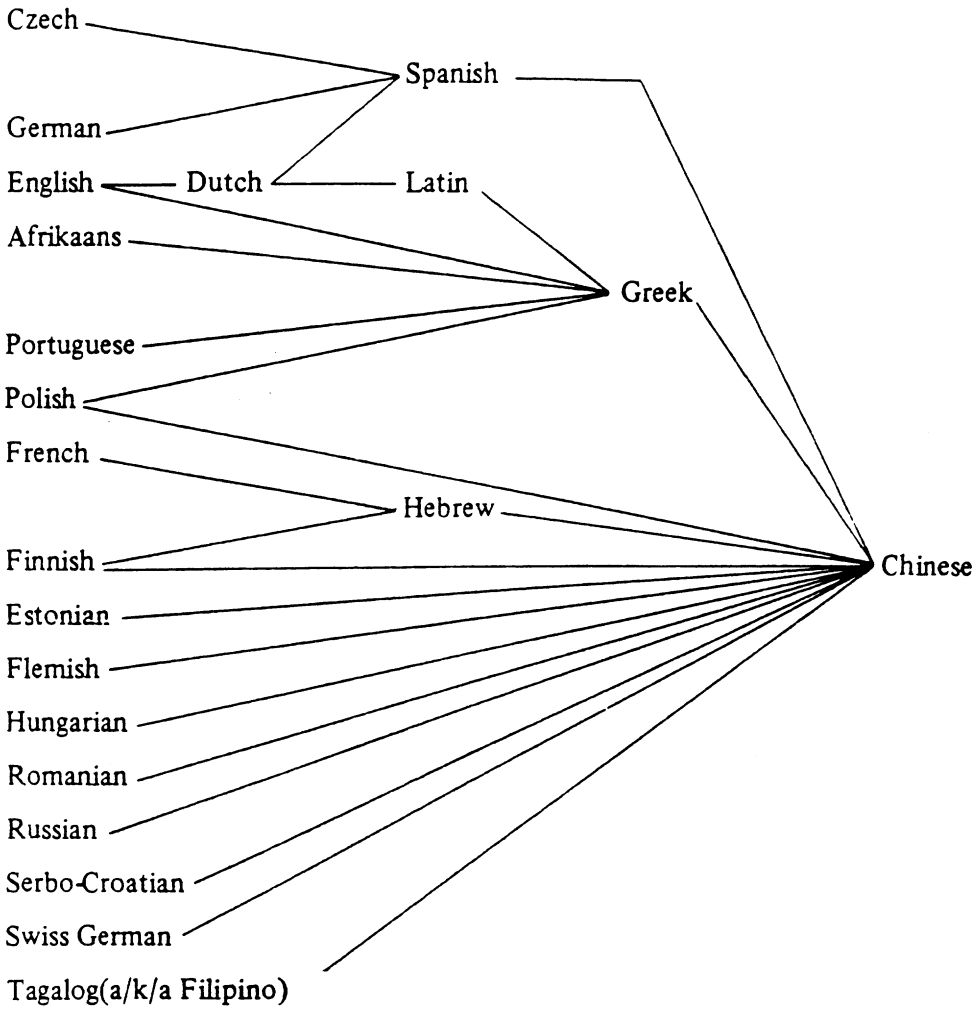

Figure 1

Figure 2

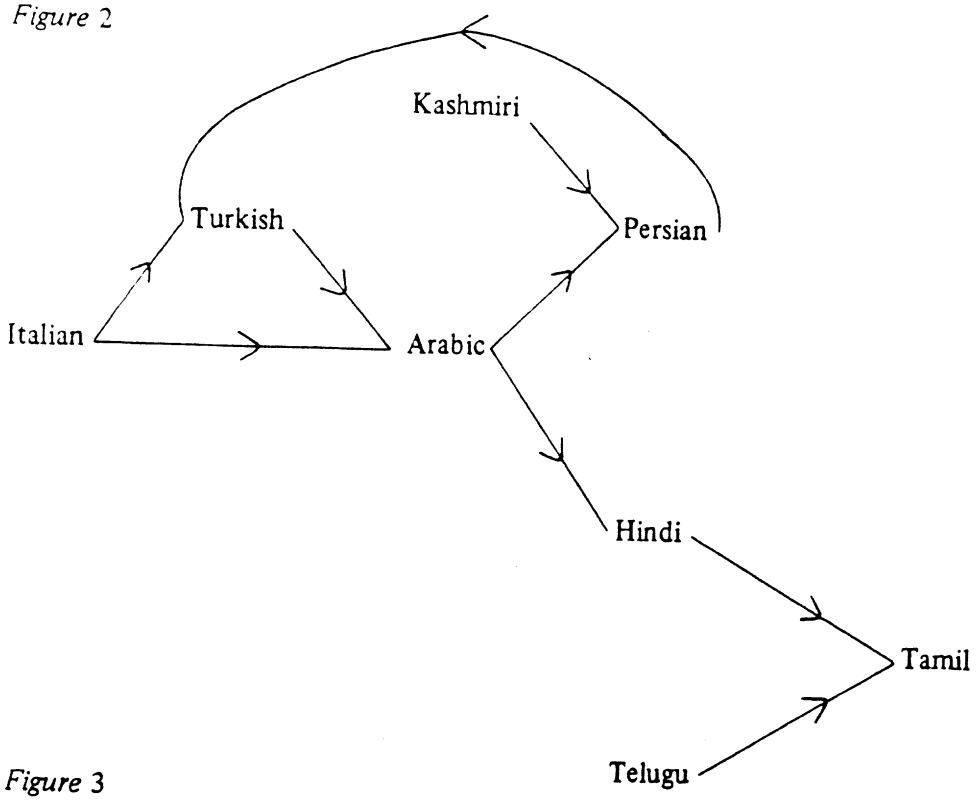

Figure 3

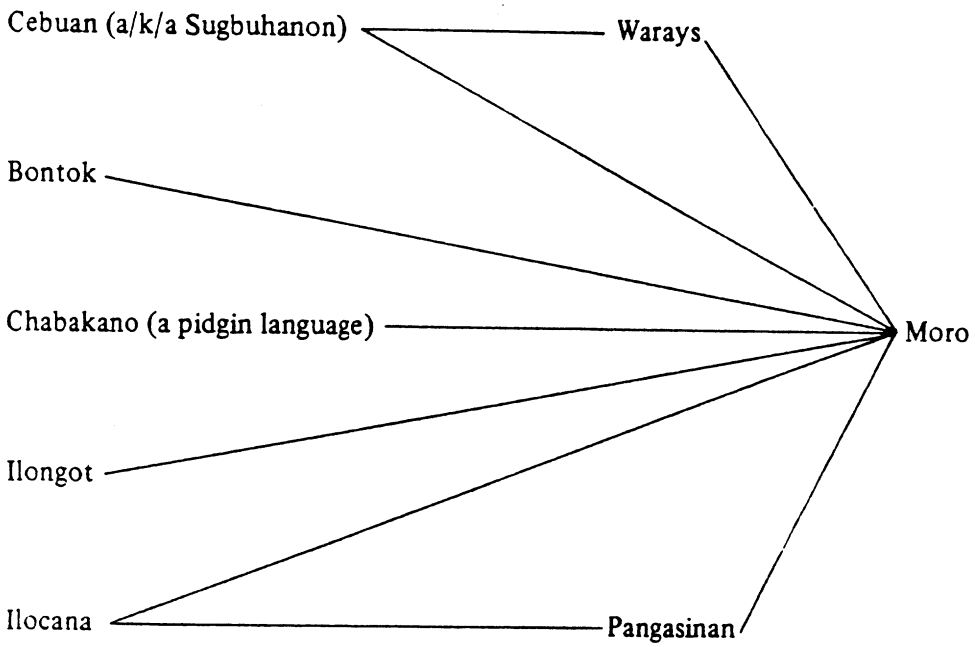

Figure 4

Japanese

Norwegian

Swedish

REFERENCES

A. *Original Sources*

Alcuin. 800. *Letter to Charlemagne.*

Bar-Hillel, Y.; Perles, M.; and Shamir, E. 1961. On formal properties of simple phrase structure grammars, *Z. Phonetik, Sprachwissen., Komm., 14*, Reprinted as cap. 9, Y. Bar-Hillel, *Language and Information*, Reading MA: Addison-Wesley, 1964.

Chomsky, N. 1956. Three models for the description of language, *IRE Trans. Inf. Th. IT-2*, 113-124.

Chomsky, N. 1957. *Syntactic Structures* 's-Gravenhage: Mouton.

Colton, C. C. 1820. *Lacon.*

DeMillo, R. A.; Lipton, R. J.; Perlis, J. A. 1977. Social processes and proofs of theorems and programs, *Proc. 5th ACM Symp. on Principles of Programming Languages.* (*Commun. Assoc. Comput. Mach.*, 22 (1979):271-80).

Dibdin, C. 1789. *Poor Jack.*

Edward III. 1349. *Motto of the Order of the Garter.*

Esperanto [pseud. of L. L. Zamenhof]. 1887. *Internacia Lingvo.*

Evey, R. J. 1963. The theory and application of pushdown store machines, *Mathematical Linguistics and Automatic Translation*, Harvard U.: Computation Lab. Rpt. NSF-10.

Gascoigne, G. 1573. *Supposes.*
Gibran, K. 1916. *Sand and Foam.*
Greibach, S. A. 1973. The hardest context-free language, *SIAM J. Computing 2*: 304-310.
Kuroda, S. Y. 1964. Classes of languages and linear-bounded automata, *Information and Control 7*: 207-223.
Merton, R. K. 1965. *On the Shoulders of Giants*, New York: Harcourt, Brace and World.
Meyer, A. R.; Stockmeyer, L. J. 1972. The equivalence problem for regular expressions with squaring requires exponential space, *Proc. 13th IEEE Symp. on Switching and Automata Theory*: 125-129.
Molière [*pseud. of J. B. Poquelin*]. 1653. *L'Étourdi.*
Pope, A. 1733. *Moral Essays.*
Pope, A. 1733-1738. *Imitations of Horace.*
John of Salisbury. 12th century. *Metalogicon.*
Sarton, G. 1935. In *Isis 24*: 107-109.
Schleyer, J. M. 1887. *Grammar with vocabularies of Volapük* (Eng. trans. by W. A. Seret).
Schützenberger, M. P. 1963. Context-free languages and push-down automata, *Information and Control 6*: 246-264.
Shakespeare, W. 1599. *Julius Caesar.*
Spurgeon, C. H. 1879. *Sermons.*
Stockmeyer, L. J. 1977. Classifying the computational complexity of problems, *Proc. 2nd IBM Symp. on Mathematical Foundations of Computer Sciences*, Kansai, Japan: 151-197.
Turing, A. M. 1936. On computable numbers, with an application to the Entscheidungsproblem, *Proc. London Math. Soc., Ser. 2-42*: 230-265.

B. *Reference Books Other than Dictionaries*

A New Dictionary of Quotations (on Historical Principles), H. L. Mencken, ed., Alfred A. Knopf, NY, 1952.
F. P. A.'s Book of Quotations, F. P. Adams, ed., Funk and Wagnalls, NY, 1952.
The Oxford Dictionary of Quotations, G. Cumberlege, ed., Oxford U. Press, London, 1953.
Dictionary of Foreign Phrases and Classical Quotations, H. P. Jones, ed., John Grant Ltd., Edinburgh, 1958.
Dictionary of Quotations, B. Evans, ed., Delacorte Press, NY, 1968.
Familiar Quotations (by J. Bartlett), 14th ed., E. M. Beck, ed., Little, Brown and Company, Boston, 1968.

C. *Informants* (with gratitude)

A. V. Aho
T. C. Ancheta
E. Arjomandi
A. K. Chandra
P. Cohen
J. Cooley
A. Ehrenfeucht
S. Even
L. Guibas
L. Herman
F. Jelinek
S. R. Kosaraju
L. Kou
W. Liniger
I. Meilijson
J. A. Moyne
N. Neergaard
O. Nevanlinna
F. Odeh
W. Paul
N. Pippenger
F. Preparata
M. Ronay
G. Rozenberg
J. Simon
T. Sundheimer
V. K. Vaishnavi
J. van Leeuwen
V. Vianu
S. Winograd
D. Wood
A. C.-C. Yao

Received 17 October 1978 – Revised 20 May 1979