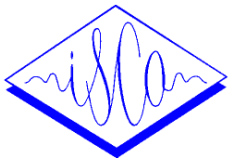

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

# FACTORS IN HUMAN LANGUAGE IDENTIFICATION

*Ian Maddieson* and Ioana Vasilescu***

*Department of Linguistics,
University of California, Berkeley, USA
ianm@socrates.berkeley.edu

**LIMSI-CNRS,
Orsay, France
ioana.vasilescu@limsi.fr

## ABSTRACT

Listeners' skill at identifying five target languages from short samples and discriminating the targets among a set of other languages was tested. Subjects in two groups, US and French, were generally quite successful in identification. Individual variation is poorly explained by degree of prior casual exposure to the target languages or academic linguistic training. Linguistic training does predict some greater success in discrimination. The implied language groupings which emerge from identification errors and in the discrimination phase are more informative from the largely linguist US subjects.

## 1. INTRODUCTION

Language identification raises many issues of theoretical and practical interest. For example, how soon does an infant begin to discriminate between its 'native' language and other languages [1], and how can automatic systems be devised to connect a telephone caller to an operator speaking the right language [2]? Despite these interests there is still comparatively little research on how adult human listeners perform on language identification tasks and what factors affect the relative degrees of success of different individuals [3, 4].

Obviously a speaking knowledge of particular languages predicts the ability to identify those specific languages [5]. But does more casual exposure to a language in everyday life also predict greater skill in identifying it correctly? Does language identification skill in general increase as a function of the number of languages known? Does involvement in the academic discipline of linguistics predict good language identification skills?

In this paper we particularly examine the potential contributions of prior language exposure and of academic linguistic knowledge to success in language identification tasks. Because academic linguists can be expected to know characteristics which are specific to certain languages and have the conceptual framework to label and keep track of these characteristics we expected this to confer an advantage over non-linguists. In addition, because they know the genetic, areal and typological categories according to which languages are classified in systematic ways, linguists' implied or explicit judgements of similarity between languages is expected to be more consistent and more readily interpretable than judgements by non-linguists.

## 2. DESIGN OF THE EXPERIMENT

We conducted an experiment consisting of three phases, all presented via a self-guided computer interface using Psyscope. The speech samples used are paragraphs from recordings of the fable of the "North Wind and the Sun," a passage long used by the International Phonetic Association (IPA) for illustrations of the phonetic characteristics of different languages [6].

In the first phase samples of five diverse target languages, Amharic, Romanian, Korean, Morroccan Arabic and Hindi, were presented for familiarization. A single speaker was presented as a model. The samples were presented over headphones in a good listening environment. The text was the first paragraph of "North Wind and the Sun". Subjects were instructed to play the samples as many times as they felt was necessary to form an impression of the characteristics of the language represented. Most chose to listen relatively few times to each sample, often only once or twice.

The second phase of the experiment was an identification test of samples of the target languages, spoken by different speakers than those used in the familiarization phase These samples used the second and final paragraphs of the "North Wind and the Sun". This design provides listeners with some opportunity to identify repeated lexical items resulting from the continuity of topic between the samples, in addition to any segmental and prosodic cues to language identity present. Each languages was presented four times, using different speaker/paragraph combinations. Order of presentation was varied randomly across subjects to eliminate any consistent effects of proximity of one language to another.

The third phase of the experiment was a test of the ability to recognize the target languages among a set of 'distractor' languages. Single samples of each of the target languages and 18 other languages were presented. The target language samples were spoken by different speakers from those used in prior phases. The 'distractor' samples were primarily excerpts from the illustrations of languages available from the IPA [6]. Subjects were asked first to decide whether or not the sample represented one of the five target languages. If they responded "yes" they were asked to identify which of the five languages they thought it was. If they responded "no" they were asked to indicate if they felt the language was similar to one of the target languages, or was unlike any of the five. This enables both explicit groupings among the languages as well those implied by errors in identification to be examined. The relative uniformity of similarity judgements across the subjects can also be examined for the insights this presents.

Because the interest in computational approaches to language identification is in reaching rapid decisions, all the samples used in this experiment are of quite short duration. Average sample duration is on the order of 20 seconds.

## 3. SUBJECT GROUPS

Two groups of subjects participated in the experiment, one of 22 engineers, computer scientists and psychologists at the LIMSI laboratory in Orsay, France, and the other of 28 faculty members, visiting scholars and students of linguistics at the University of California, Berkeley. The majority of subjects were native speakers of the local national language, but the US group includes several speakers whose first language is not English. Both populations cover broadly similar ranges of age and overall educational level.

All subjects are assigned to a level on the scale in Table 1 indicating their degree of exposure to academic linguistics. The US subject group cover steps 2-6 of the scale with at least three subjects representing each level. Seven of the subjects are professors of linguistics and three are senior graduate students close to completion of a doctorate. Twelve of the subjects tested in France had no training in the discipline, so are assigned to level 1. Most remaining members of this group have a limited exposure to linguistic ways of thinking through such related disciplines as cognitive science or speech synthesis. This educational background does not cover the same knowledge base as a degree in general linguistics but it seems necessary not to treat such subjects as 'pure' non-linguists: they are mostly classed at level 2. One is a doctoral candidate in speech and classified as step 4.

| Level | Description |
|---|---|
| 1 | No formal linguistic training |
| 2 | Some undergraduate linguistics courses |
| 3 | Undergraduate degree/senior student in linguistics |
| 4 | Some graduate study/Master's degree in linguistics |
| 5 | Senior graduate students in linguistics |
| 6 | Doctorate/Professor of Linguistics |

*Table 1*: Scale of expertise in linguistics

Subjects with more than a passing familiarity with any of the five target languages were excluded. Each subject's familiarity with the target and distractor languages, as well as several additional major world languages not included in the experiment, was assessed by a post- (US) or pre-experiment questionnaire. Table 2 gives the descriptive labels used for levels of familiarity. These labels are designed to suggest appropriate incremental steps in familiarity. A weakness of this assessment method is that it relies on self-evaluation rather than an objective measure of exposure to a language or competence in speaking it; different individuals may well use different thresholds for the various levels of language familiarity in the table. However, no other method is practicable.

| Level | Description |
|---|---|
| 1 | I don't think I've ever heard this language |
| 2 | I've heard this language spoken around me |
| 3 | I can understand a few words of this language |
| 4 | I can speak this language a little |
| 5 | I can speak this language well or perfectly |

*Table 2*: Scale for self-evaluation of language familiarity

## 4. RESULTS

### 4.1. Identification test

In the identification phase subjects were asked to hit a numbered key corresponding to which of the five languages they believed the sample being played represented. The next sample was presented after a keystroke was registered. In a very few instances a subject hit an erroneous key, resulting in no decision being recorded for that sample. There are 6 such instances in the entire response set.

The mean success rate in the test of identification of the five target languages across the entire subject group is 65%, corresponding to 13 correct identifications among the 20 exemplars presented (4 samples x 5 languages). A chance result is 20% correct. The mean for the subject group tested in France is 60% correct, 56% for those at Level 1 of Table 1 (this includes the one subject who responded at chance level). The group of 7 linguistics professors tested in the US score a mean of 71% correct, representing about 14 correct responses. The 3 students close to completing a doctorate in linguistics also have a mean of 71% correct. Differences between sets of subjects divided by location and level of linguistic expertise do not reach statistical significance.

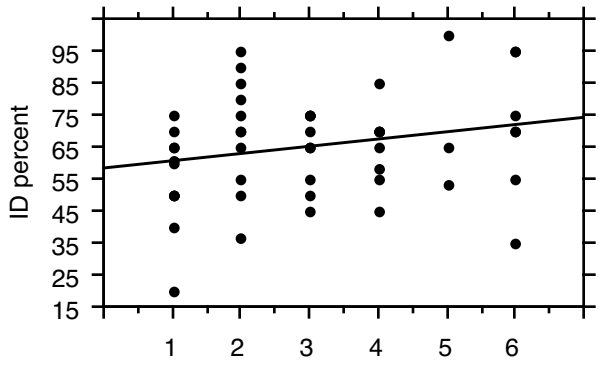


*Figure 1*: Percent correct in identification test vs level of academic linguistic expertise, 48 subjects, $R^2$ = .056

The scatter of percent correct scores by level of linguistic expertise is plotted in Figure 1 (a dot may represent multiple subjects with the same scores). A simple regression line is fitted. Clearly, very little of the variation between individual scores is explained by the level of linguistic expertise.

An index of reported familiarity with the target languages for 45 subjects was computed from the questionnaires (3 did not complete one). Each step above 1 in the scale shown in Table 2 for any target language adds .5 to this index. Good speaking ability in all five would therefore score 10. Most subjects reported no more than having heard one or more of the languages "spoken around them", so scores cluster around 0.5-2. The scatter of percent correct identification scores by values on this index is plotted in Figure 2. There is a relatively weak relationship between prior exposure to the target languages and success in identification, $R^2$ =.207. If the three individuals with index scores of 3.5 or higher are excluded from the analysis $R^2$ drops to .144.

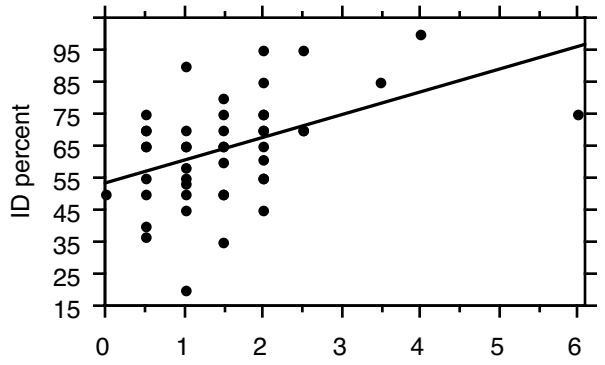


*Figure 2*: Percent correct in identification test vs index of familiarity with target languages, 45 subjects

Analysis of the nature of the errors in identification is also of interest. A matrix of the aggregated responses for the subjects tested in the US is given in Table 3. 100% correct is 108 responses in the correct box; correct identification of Korean samples almost achieves this level, and there are fewer false identifications of other languages as Korean (9) than false identifications as one of the other four languages (32-52)

| | **Identified as....** | | | | |
|---|---|---|---|---|---|
| **Played ↓** | **Amh.** | **Rom.** | **Korean** | **M.Arab.** | **Hindi** |
| Amharic | **64** | 11 | 2 | 20 | 10 |
| Romanian | 8 | **66** | 6 | 8 | 20 |
| Korean | | | **107** | 1 | |
| M. Arabic | 30 | 2 | | **73** | 3 |
| Hindi | 14 | 19 | 1 | 10 | **63** |

*Table 3*: Identification results for 28 subjects in US

Over half the errors are attributable to two pairwise confusions, namely between Amharic and Arabic, and between Romanian and Hindi. Possibly the prior conceptual framework provided by knowing that the members of these pairs of languages are classed together in the same family tempted these linguistically-trained subjects to form super-categories of Semitic and Indo-European languages.

The pattern of responses from the subjects tested in France differs somewhat, as shown in Table 4. 100% correct is 88 for this subject set. The Korean and Morroccan Arabic samples are both well-identified by these subjects, at about 90% correct. Korean appears simply to be the most distinctive of the set of languages tested. The greater familiatity of Arabic in France due to the large number of residents of North African origin surely accounts for the greater success in identifying Arabic for this group (18 of 19 respondents report having at least heard some Arabic). However, the errors on the remaining languages are more randomly distributed than those of the more linguistically knowledgeable US subjects. The two most poorly identified languages, Amharic and Hindi are only identified correctly at about 35%. There is a great number of false identifications of other languages as Amhar-ic (65), and Hindi does not particularly pair with Romanian.

| | **Identified as....** | | | | |
|---|---|---|---|---|---|
| **Played ↓** | **Amh.** | **Rom.** | **Korean** | **M.Arab.** | **Hindi** |
| Amharic | **32** | 12 | 7 | 16 | 19 |
| Romanian | 18 | **47** | 5 | 0 | 17 |
| Korean | 7 | 2 | **70** | 0 | 9 |
| M. Arabic | 12 | 0 | 0 | **73** | 0 |
| Hindi | 28 | 26 | 2 | 2 | **30** |

*Table 4*: Identification results for 22 subjects in France

### 4.2. Discrimination/Similarity test

In the third phase, subjects heard samples of the five target languages together with samples of 18 other languages. Results are shown in Table 5 for the subjects tested in the US and in Table 6 for the subjects tested in France. Recall that subjects were asked to report if they considered the sample to *be* one of the five target languages, *like* one of the target languages, or *unlike* any of the target languages. In these tables the name of the language in the sample is in the first column. The total of subjects who responded that the sample either *was* or was *like* one of the target languages is tabulated in the five columns headed by abbreviated names of these languages. The next column is the number of subjects who responded that the sample was *unlike* any of the target languages. Below the target languages at the top of the table, the other languages are listed in increasing frequency of 'unlike' responses. A small number of non-responses occurred.

| | Amh | Rom | Kor | M. Ar | Hin | Unl. | Disp |
|---|---|---|---|---|---|---|---|
| Amharic | 9 | 3 | 0 | 4 | 5 | 7 | 2.19 |
| Romanian | 3 | 10 | 0 | 1 | 5 | 9 | 2.10 |
| Korean | 0 | 0 | 25 | 0 | 0 | 3 | 1.27 |
| Arabic | 4 | 0 | 0 | 21 | 1 | 2 | 1.70 |
| Hindi | 0 | 1 | 3 | 2 | 16 | 6 | 2.00 |
| Portuguese | 1 | 16 | 1 | 2 | 3 | 5 | 2.15 |
| Persian | 4 | 3 | 3 | 8 | 3 | 6 | 2.40 |
| Sindhi | 3 | 3 | 6 | 3 | 6 | 7 | 2.41 |
| Hausa | 11 | 5 | 0 | 3 | 1 | 7 | 2.10 |
| Catalan | 5 | 16 | 0 | 0 | 0 | 7 | 1.68 |
| Croatian | 3 | 17 | 0 | 0 | 0 | 8 | 1.64 |
| Turkish | 5 | 2 | 0 | 4 | 6 | 11 | 2.16 |
| Hebrew | 1 | 2 | 0 | 14 | 0 | 11 | 1.79 |
| Czech | 0 | 14 | 0 | 1 | 2 | 11 | 1.79 |
| Slovenian | 0 | 16 | 0 | 1 | 0 | 11 | 1.57 |
| Cantonese | 0 | 0 | 15 | 0 | 0 | 13 | 1.41 |
| Chichewa | 5 | 2 | 2 | 1 | 4 | 14 | 2.23 |
| Bulgarian | 1 | 11 | 0 | 1 | 1 | 14 | 1.90 |
| Hungarian | 6 | 3 | 2 | 1 | 1 | 15 | 2.17 |
| Thai | 0 | 0 | 11 | 0 | 0 | 16 | 1.41 |
| Irish | 1 | 2 | 0 | 3 | 3 | 19 | 1.93 |
| Swedish | 2 | 3 | 0 | 2 | 0 | 20 | 1.74 |
| Dutch | 0 | 3 | 0 | 0 | 0 | 25 | 1.27 |

*Table 5:* Similarity results for 28 subjects in US

| | Amh | Rom | Kor | M. Ar | Hin | Unl. | Disp |
|---|---|---|---|---|---|---|---|
| Amharic | 13 | 2 | 0 | 2 | 4 | 1 | 1.51 |
| Romanian | 3 | 8 | 2 | 0 | 5 | 4 | 1.83 |
| Korean | 2 | 0 | 19 | 0 | 0 | 1 | 1.44 |
| Arabic | 1 | 0 | 0 | 21 | 0 | 0 | 1.09 |
| Hindi | 3 | 2 | 1 | 0 | 13 | 3 | 1.54 |
| Cantonese | 1 | 0 | 20 | 0 | 0 | 1 | 1.41 |
| Sindhi | 4 | 0 | 5 | 1 | 9 | 3 | 1.66 |
| Thai | 1 | 2 | 16 | 0 | 0 | 3 | 1.70 |
| Hausa | 12 | 3 | 1 | 1 | 2 | 3 | 1.70 |
| Persian | 6 | 5 | 3 | 2 | 2 | 4 | 1.96 |
| Czech | 1 | 12 | 2 | 1 | 1 | 5 | 1.95 |
| Chichewa | 6 | 0 | 4 | 0 | 7 | 5 | 1.53 |
| Bulgarian | 2 | 15 | 0 | 0 | 0 | 5 | 1.48 |
| Hungarian | 3 | 11 | 0 | 0 | 3 | 5 | 1.65 |
| Croatian | 1 | 15 | 0 | 0 | 0 | 6 | 1.40 |
| Slovenian | 0 | 16 | 0 | 0 | 0 | 6 | 1.29 |
| Portuguese | 1 | 12 | 0 | 0 | 3 | 6 | 1.61 |
| Turkish | 3 | 4 | 0 | 1 | 5 | 9 | 1.64 |
| Hebrew | 4 | 1 | 1 | 7 | 0 | 9 | 1.54 |
| Irish | 2 | 3 | 1 | 3 | 3 | 10 | 1.76 |
| Catalan | 1 | 9 | 0 | 0 | 1 | 11 | 1.47 |
| Dutch | 1 | 2 | 0 | 1 | 0 | 18 | 1.37 |
| Swedish | 0 | 2 | 0 | 0 | 0 | 20 | 0.94 |

*Table 6*: Similarity results for 22 subjects in France

The final column contains an index of the diversity of the responses to each sample. It is the sum of the square roots of the totals in each of the six response columns divided by the square root of the total number of responses: if the responses uniformly fall in one response category, the index = 1 and it rises as the responses become closer to being equally dispersed across the response categories. Values of this index are moderately well correlated across the two subject populations (correlation coefficient = .527) — although not necessarily indicating the two groups classify a given sample in similar ways, this index highlights which languages were sources of agreement vs confusion or indecision. The dispersion index results are plotted in Figure 3.

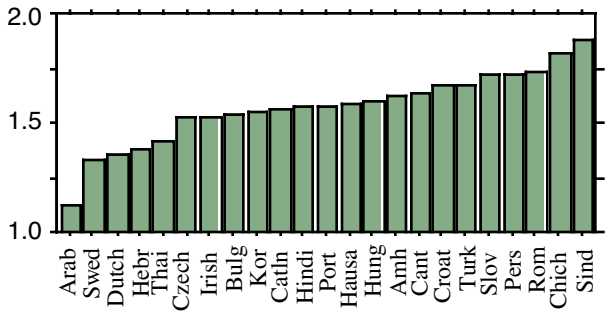


*Figure 3*: Dispersion index (mean of 2 populations)

We analyze the results of this phase primarily for the value the expressed and implied similarities may have in approaching a partial identification of a language, for example by placing it in an appropriate genetic or geographic group [7].

Both subject groups were quite consistent in judging Korean and Arabic as targets or target-like, but were relatively unsuccessful with the Amharic and Romanian samples (high dispersion index). However, Hausa is frequently classed with Amharic. This is the only other language in the sample with ejective consonants, and shares membership in the Afro-Asiatic language family. Both groups of subjects tend to assimilate the Balkan Slavic languages, Bulgarian, Croatian and Slovenian, to Romanian. These languages have a long history of contact and the Balkans are a classic example of a linguistic convergence area. Czech, and, for the French subjects, Hungarian, also join this areal grouping. The US linguist group also strongly associate the two other Romance languages, Catalan and Portuguese with Romanian; the non-linguists make this association a little less frequently.

Thai and Cantonese are associated with Korean for many subjects in both groups, but especially by the non-linguist group. These three languages are the three languages from the eastern part of Asia included in the experiment. Both Korean and Thai have been influenced by varieties of Chinese and to some degree have converged to the linguistic type.

The more linguistically-oriented US subject group are more likely to classify a distractor language as unlike any of the targets than the French group. In this they show a better discrimination of the differences between the languages as a whole. The Germanic languages (Dutch, Swedish) are regarded as unlike other languages by majorities in both groups. But only the linguist group casts a majority vote for other distractor languages (Hungarian, Thai, Irish) being unlike any of the targets. 12 languages are viewed as 'unlike' by 40% or more of the US group but only 6 by the French group.

## 5. FINAL REMARKS

The present experiment provides less evidence than we had expected that the cognitive tools available to a trained linguist assist in language identification, although they appear to provide a firmer basis for discrimination and similarity judgements.. Further analysis of these results and future research will aim at understanding what factors may underlie the individual differences in language identification skills.

## 6. REFERENCES

[1] Ramus, F., Nespor, M., Mehler, J. Correlates of linguistic rhythm in the speech signal, Cognition 73: 265-92, 1999.

[2] Muthusamy, Y.K., Barnard, E., Cole, R.A.: Automatic Language identification: A review/tutorial.", IEEE Signal Processing Magazine 11(4): 33-41, 1994.

[3] Stockmal V., Muljani D., Bond Z. Perceptual features of unknown foreign languages as revealed by multi dimensional scaling, ICSLP, Philadelphia, 1996.

[4] Marks, E.A., Bond, Z.S., Stockmal, V., The effect of proficiency in a specific foreign language on the ability to identify a novel foreign language, ICPhS, San Francisco, 1999.

[5] Vasilescu, I., Pellegrino, F., Hombert, J.M. Perceptual features for the identification of Romance languages, ICSLP Beijing, 2000.

[6] IPA . Handbook of the International Phonetic Association, University of Cambridge Press, Cambridge, 1999.

[7] Hombert, J-M, Maddieson, I. 1999. The use of 'rare' segments for language identification. Eurospeech '99, Vol 1: 379-82. Budapest. 1999.