Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Phonetic and orthographic properties of the basic vocabulary of five European languages

Carlson, R. and Elenius, K. O. E. and Granström, B. and Hunnicutt, S.

KTH Computer Science and Communication

journal: STL-QPSR
volume: 26
number: 1
year: 1985
pages: 063-094

http://www.speech.kth.se/qpsr

C. PHONETIC AND ORTHOGRAPHIC PROPERTIES OF THE BASIC VOCABULARY OF FIVE EUROPEAN LANGUAGES *

Rolf Carlson, Kjell Elenius, Björn Granström, and Sheri Hunnicutt

Abstract

Corpora of approximately 10,000 words have been examined in five languages: Swedish, English, German, Italian, and French. A 2-class and a 6-class "cohort" classification have been defined, and calculations made of the number of cohorts, the number of unique cohorts, and their maximum, mean, and expected sizes. The discriminatory ability of stress is also considered. Further calculations examine the predictability of a word given either its first or last letters (forward and backward prediction). The predictive capacity of known stress pattern, and, for Swedish, known parts of speech, has also been studied.

Introduction

For a number of years we have been working with text corpora of different European languages. This was originally part of our multi-language text-to-speech project (Carlson, Granström, & Hunnicutt, 1982). As a base for development of both text-to-speech rules and exceptions dictionaries, we needed large collections of the basic vocabulary in the different languages. These corpora were generally available in print only. Most of the corpora were created during the last two decades from large samples of newspaper material, typically 1 million words.

These corpora were phonetically transcribed for the text-to-speech project. Some of them have also been processed in other ways, including partial morphological decomposition and parts-of-speech labeling.

The corpora in this study represent a variety of West European languages: both Germanic (German, English, and Swedish) and Romance (French and Italian) with large individual differences. Swedish is a language with tonal word distinctions. German has a very rich inflexional structure and a strong tendency for compounding compared to, English, for example. Italian has a simple relation between orthography and pronunciation but relatively free stress, whereas French has fixed stress but a more complicated orthographic system with many letter sequences lacking correspondence in pronunciation. This suggests that the optimal strategies for handling large vocabularies in speech synthesis or speech recognition systems, for example, would be different.

In this paper we will present several different studies based on these corpora. Our main motivation for this work has been in relation to large-vocabulary speech recognition systems. Similar studies with this motivation have been carried out elsewhere, notably at Massachusetts Institute of Technology.

---

* also presented at the French-Swedish Seminar, Grenoble, April 22-24, 1985.

Coarse, and hence relatively reliable phonetic classifications, have been suggested as a means of forming small groups of possible word candidates from a large dictionary, so-called "cohorts." We have investigated the consequences of four different schemes for classification. The classifications have been made in terms of vowels and consonants or in terms of vowels and five consonantal categories. In both cases, the analysis has been carried out with or without stress and, for Swedish, word tone. At this point, we do not make any claims about the feasibility of the classifications. We see them, rather, as examples that reveal some of the lexical structure of the studied languages. Among the aspects studied are statistical distributions of cohort sizes, maximum, mean, and expected cohort size, and phonetic structure of the largest, most probable cohorts in the individual languages.

Word beginnings have been thought of as richer in structure and, hence, more useful for discrimination than word endings, at least for Germanic languages. We have studied this by seeing how far into a word one needs to go to identify the word within our corpora. This has been done both from the beginning and the end of the word. The effect of knowing the stress structure of the word or its part of speech has also been studied. Apart from its general interest, this study has applications in communication aids for non-vocal, another project we are currently pursuing. The number of keystrokes while typing could be substantially reduced when creating sentences if on-line prediction is implemented using dictionaries of different kinds and grammatical hypotheses.

The present report necessarily contains a great deal of detail of rather disparate nature in figures and tables. Only part of this is commented in the text. After this introduction, it should be possible, according to the reader's interest, to selectively read different sections. Since the results reflect the structure of well-known languages, they will frequently comform with our intuitions and sometimes appear almost trivial, at least from a qualitative point of view. We have , however, felt a need for quantitative and comparative discriptions of language structure.

The corpora

The corpora that we have used are the approximately 10,000 most frequent word forms in the different languages as reported in the studies of Table 1.

The studies are based on between .5 and 1 million words. The number of different word forms varies between about 110,000 for the Swedish one million word material to about 30,000 for the French half-million word material. Most studies are based on varied selections of text samples from newspapers. The Italian material is based on a more varied selection of printed material and the French samples are drawn from modern novels.

SWEDISH	ALLEN, S. 1970 "Frequency Dictionary of Present-Day Swedish"
ENGLISH	KUCERA, H. AND FRANCIS, W.N. 1967 "Computational Analysis of Present-Day American English"
GERMAN	ROSENGREN, I. 1972 "Ein Frequenzwörterbuch der deutschen Zeitungssprache"
FRENCH	ENGWALL, G. 1984 "Vocabulaire du roman francais 1962-1968"
ITALIAN	BORTOLINI, V., TAGLIAVINI, C., AND ZAMPOLLI, A. 1971 "Lessico di frequenza della lingua italiana contemporanea"

TABLE 1. Frequency-sorted corpora for five languages.

Phonetic transcriptions

None of the corpora were avialable to us with phonetic transcription. To facilitate the transcription process, which is not a small task, we developed a semi-automatic procedure.

The creation of these corpora were originally part of our multilingual text-to-speech project. The first step was to process the material through a preliminary version of the orthographic-to-phonetic component of the text-to-speech system. The next step was to check and edit the proposed phonetic transcriptions. This was accomplished interactively on a computer terminal with the possibility to listen to the corresponding synthesized words. This feature of the editing system greatly improved the precision and speed of the correction procedure. In the notation system we have the possibility of specifying alternative pronunciations. This facility has, however, not been used in the studies reported here.

Root corpus formation

The more productive suffixes have been removed from the Swedish and English corpora to form what we denote as the "root corpora." It should be noted that no prefix or compound word analyses have been made. The suffixes removed are mostly inflexional. No endings that affect stress have been considered. Examples can be seen in Table 2, along with the frequency of occurrence in our word corpora.

EXAMPLE OF ENDINGS IN THE ENGLISH CORPUS		EXAMPLE OF ENDINGS IN THE SWEDISH CORPUS			
-S'	2	-A	144	-EN	676
-ES'	1	-ISKA	48	-R	378
-D	347	-NA	56	-AR	78
-ED	552	-ARNA	16	-ER	472
-ABLE	34	-ERNA	126	-S	337
-ING	601	-D	90	-DES	41
-FUL	26	-DE	212	-NS	7
-'S	103	-NDE	172	-ENS	79
-S	1498	-RE	80	-ETS	25
-ES	42	-STE	19	-T	578
-MENT	59	-ISK	29	-ET	208
-LY	278	-N	248		

TABLE 2. Endings used in the "root corpus" formation.

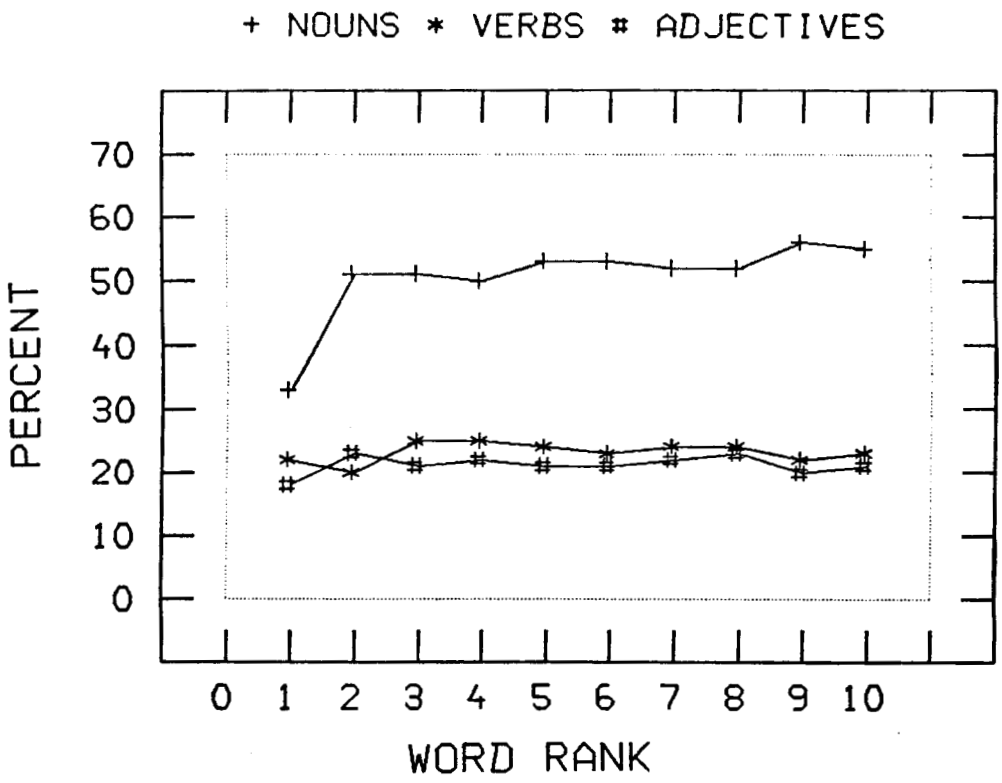
By this processing, the number of items in the corpus was reduced by 30% for English and 35% for Swedish.

Parts of speech analysis

The Swedish corpus has been subjected to a parts-of-speech labeling. As for the phonetic transcription and the suffix removal, this was done in a semi-automatic fashion. A short rule system was developed that made parts-of-speech assignment based on orthographic surface structure criteria. Only about forty rules were used. The rule-generated labels were then manually corrected.

In Fig. 1, the result of parts-of-speech analysis can be seen, along with the error probability for the main categories. The result is organized according to word frequency classes. In Fig. 1a, the bin size is 1000 words, i.e., the leftmost data points relate to the 1000 most frequent words in the corpus. With decreasing frequency, the proportion of nouns stabilizes at above 50% while verbs and adjectives each represent about 20%.

In Fig. 1b, the errors in the rule-predicted parts of speech are presented. The results demonstrate the effect of high-frequency words. For the 200 most common words, we get a total prediction error of 90%. This is obviously due to the high proportion of function words in this frequency class. With decreasing word frequency, the total error approaches 20%. We obtain the smallest error for nouns, which is also the predominant category. We regard these results as very promising. Combined with a small function word dictionary with grammatical information, a rule system like this forms a good basis for a parsing system without an extensive dictionary.



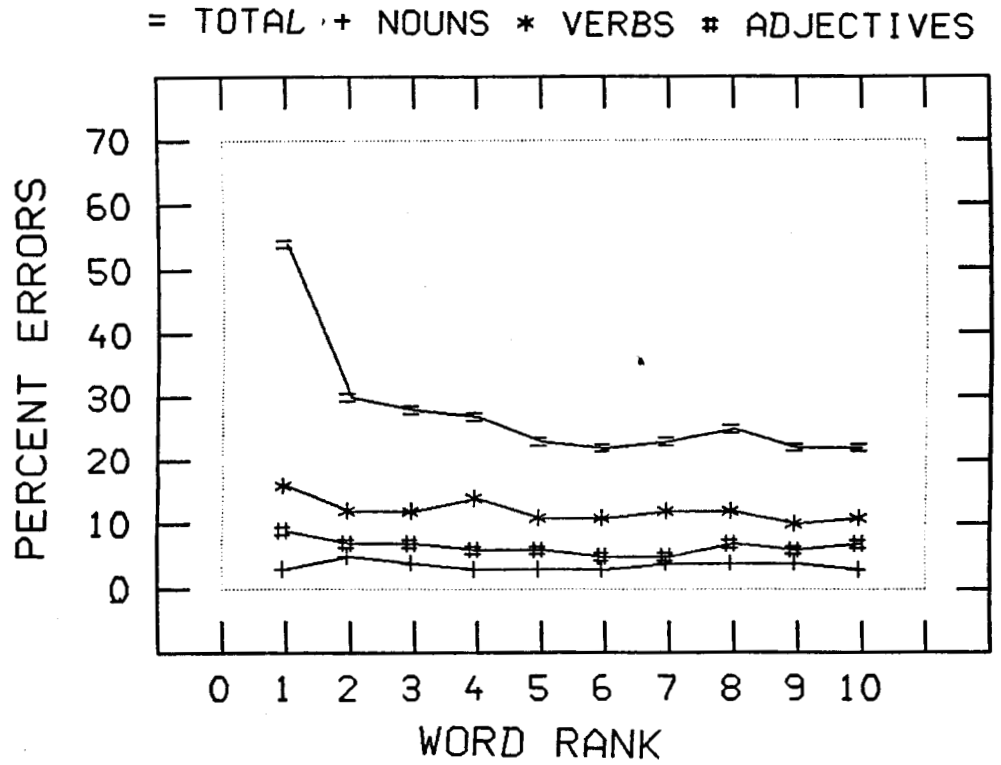


Fig. 1a (top) Distribution of parts of speech in the Swedish word corpus according to word rank (in thousands).

1b (bottom) Prediction error in the automatic parts-of-speech assignment.

Study of the Corpora: Word Length and Structure

In the five languages we have examined, mean word length differs by one or two letters, and by two or three phonemes (see Table 3). Maximum word length, both orthographically and phonetically, is found in German. The shortest words orthographically (in the mean) are in English, and phonetically, in French. Italian, Swedish, English, and German average slightly more than one letter per phoneme. French, however, averages 1.5 letters per phoneme, ranking lowest (shortest) for phonetic word length and second highest, next to German, for word length in letters. The consonants which are pronounced only in liaison are not included in these data.

Letter data	Word Length	Root Length
English	7.09	6.68
Italian	7.39	--
Swedish	7.43	6.78
French	7.62	--
German	8.69	--

Phonetic data	Word Length	Root Length
French	5.20	--
English	5.96	5.65
Swedish	6.94	6.45
Italian	6.94	--
German	7.78	--

TABLE 3. Mean letter and phonetic length of words and roots in Swedish and English, and of words in German, Italian, and French.

The mean number of letters in a root morph is nearly the same for Swedish and English, the two corpora for which we have done morphemic analysis. The length of phonetic representations for roots differs by a little less than a phoneme. Word length in the two languages differs by one phoneme.

Mean data for the distribution of vowels and consonants in words and roots is given in Table 4. The ratio of consonants to vowels in a word is lowest for the two Romance languages, French and Italian, and highest for German words. Italian words contain the most vowels per word, and German, the most consonants. The mean number of vowels for words and roots in all five languages except Italian lies between 2 and 3; consonant means lie between 3 and 5. Graphs of these distributions are shown in Fig. 2.

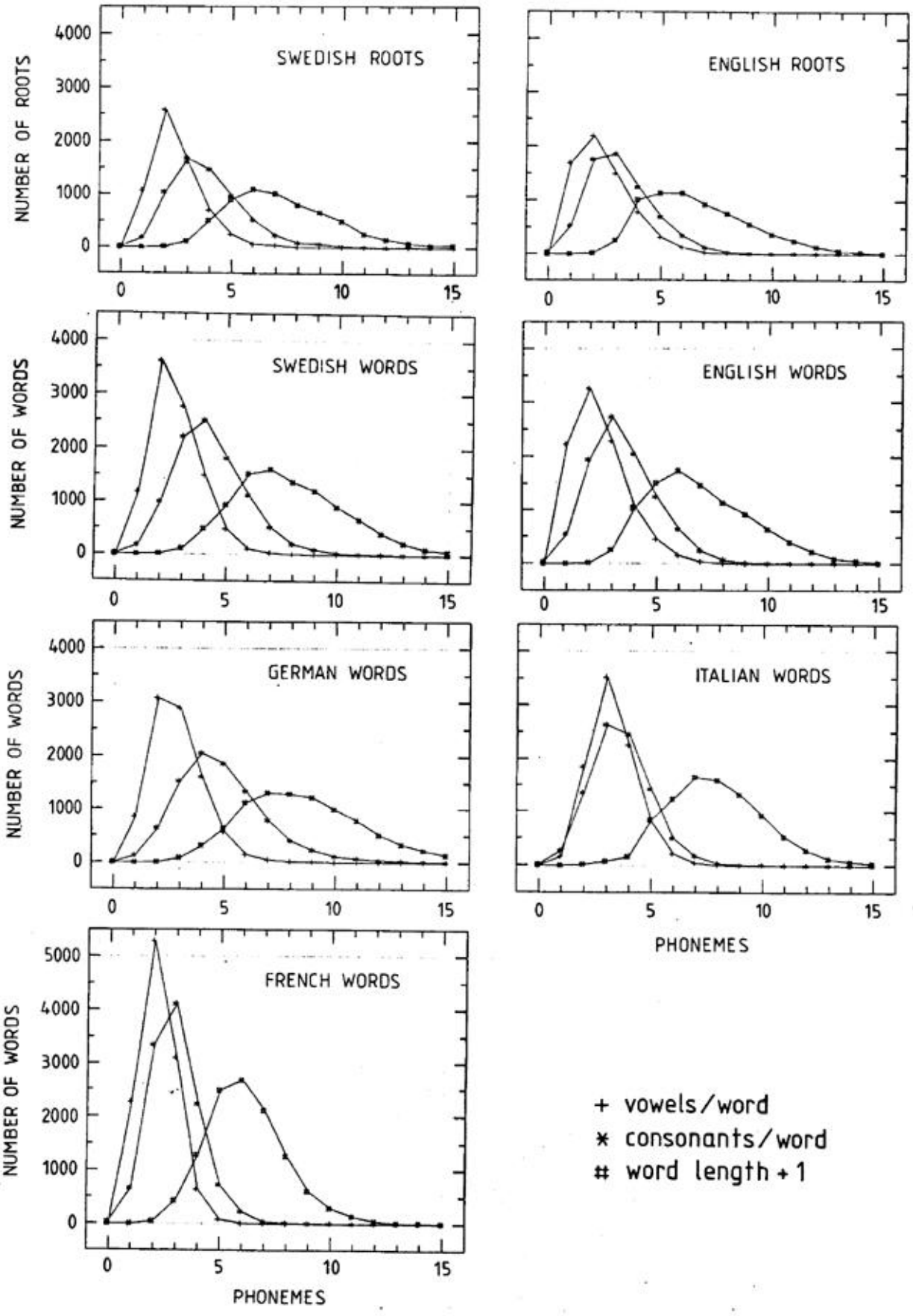


Fig. 2. Distribution of number of consonants and vowels in the word and root corpus.

Phonetic data	Vowels		C/V Ratio	
	Words	Roots	Words	Roots
French	2.21	--	1.35	--
English	2.47	2.45	1.41	1.31
Swedish	2.69	2.51	1.58	1.57
German	2.87		1.71	
Italian	3.27		1.12	

TABLE 4. Mean number of vowels and consonant/vowel ratio. Phonetic data: Swedish, English, German, Italian, and French

Study of the Corpora: Partitioning into Cohorts

In the studies described below, we have partitioned our corpora by two major classifications. The more coarse classification results from mapping the string of phonemes representing a word's pronunciation into a corresponding string in which all consonants are replaced by C, and all vowels by either 'V (stressed vowel) or V (unstressed vowel). (This description is somewhat simplified; further details will be considered in the following discussion.) A variation of this classification does not consider stress, and simply maps all vowels into V. For example, the words "plane," "stop" and "troop" are mapped into a class identified by the string C-C-'V-C, or alternately, C-C-V-C. This representation is of interest in linguistic investigations and has been referred to as the "syllabic skeleton." (See, for example, Halle & Vergnaud, 1980.) Our mapping partitions all the words in our corpora into such classes, generally referred to as equivalence classes. We will refer to the set of words mapped into a particular equivalence class as a "cohort." This term has come to be employed in some psycholinguistic and speech recognition literature in this more general sense since its introduction by Marslen-Wilson (1978) to refer to the group of all words that begin with a particular phoneme string.

The second major classification results from mapping the pronunciation of the words in the corpora into corresponding strings of 6-valued or 7-valued elements. Consonants fall into the categories nasal, other sonorant, stops, strong fricatives, and weak fricatives. Vowels are categorized as in the previously described classification. This classification has earlier been explored by the speech recognition group at Massachusetts Institute of Technology (Shipman & Zue, 1982; Huttenlocher & Zue, 1983).

Partitioning of the corpora into cohorts in this manner provides us with an approximation of the benefit that would be derived in a speech recognition system if all phonemes could be classified at least this well. Examination of the words in a particular cohort shows us which words we would have to make a decision among after this coarse classi-

fication. Using the data on cohorts as a base, we have made additional calculations designed to describe the expected speech recognition task. The number of unique (1-member) cohorts represent those words for which no further decisions need be taken: the cohort contains a single word. The maximum size of a cohort reveals the worst case. Mean cohort size is presented in the results, but is actually not as useful as another value, expected cohort size (Waibel, 1982). This measure takes into consideration the likelihood of the occurrence of a particular cohort size and the probability of a word to fall into a cohort of this size. It appears to be a more representative measure of the actual task involved in word recognition. If we look at Swedish 3-class cohorts, for example, we see that over half of the cohorts are unique. If one were to pick a cohort at random, then, it would likely have only one member. This large number of 1-member cohorts is reflected in the mean cohort size of 7. Picking a word at random, however, we would likely find it in a cohort of much larger size since the expected cohort size is 105.

The assignment of a vowel to the stressed vowel class varies somewhat for the five languages. For German, Italian, and French, only one vowel per word is marked as stressed. In the English corpus, both primary and secondary stress is marked and, when forming cohorts, vowels with varying degree of stress are mapped into separate items, either ˈV (primary-stressed vowel) or ˌV (secondary-stressed vowel). Swedish has two types of word tone. A vowel with Type I word tone is mapped into ˈV, and the vowel pair with Type II word tone into ˌV and ˋV.

Swedish 2-class and 3-class cohorts

If accuracy in spoken word recognition were limited to perceiving stress and distinguishing vowels from consonants, the Swedish corpus of 9,679 words would fall into 1,336 classes, or cohorts. Of these cohorts, 682 would contain only one word each. Of the remaining 654 cohorts, containing more than one word, the maximum size is 393, representing 4.1% of the corpus. In the mean, a cohort in Swedish would contain 7 words. The expected size of a cohort would be 105 words (see Table 5).

The five largest cohorts contain words of the following form: (C stands for consonant, ˈV for Type I word tone, and ˌV and ˈV for the primary and secondary stressed vowels in Type II word tone.)

Number of words	Cohort	Example
393	C-ˈV-C	bil
358	C-ˌV-C-ˋV-C	billig
348	C-ˌV-C-C-ˋV-C	bunden
293	C-ˈV-C-V-C	benen
289	C-ˌV-C-ˋV	bita

If stress cues were insufficient, and judgment had to be based on vowel-consonant decisions alone, information would be greatly impoverished. The corpus would be classified into 732 categories, giving only 55% of the discriminability available as when stress cues are sufficient. The number of unique cohorts, containing one word each would be only 48% of the number produced with stress discrimination. The maximum cohort size would be 1.9 times as large. This is essentially due to the collapsing of two frequent two-syllable cohorts with different word tone into one pattern (C-V-C-V-C). The expected cohort size would be 192 words. (See Tables 5 and 6.)

In order to inspect the effects of affixation on this classification, one can look at root morphs only. Converting the values in Table 5 to percentages of the total corpus, we can see that the only major effect of the non-stress-affecting suffixes that we have removed in our root corpus formation is in the discriminatory function of stress. (See Table 6.) It appears that stress discrimination, which nearly halves the maximum cohort size for words in Swedish, has no effect on the maximum cohort size of roots since this cohort contains one-syllable members. The maximum size in the classification including stress is about the same for words and roots. Since the total number of words is less for roots, it seems that affixation does not decrease the maximum size, i.e., there are not so many new categories which would take members from the group of maximum size. This must imply that, since the new affixed words must not have the same patterns as the roots, there is a near one-for-one replacement of root patterns by affixed word patterns. We see, however, that the number of unique cohorts is much less for roots, suggesting that affixation does have a discriminatory function.

English 2-class and 3-class cohorts

Partitioning our English corpus into words described by their consonants, primary and secondary stressed vowels and unstressed vowels yields 1,515 cohorts containing a total of 9,526 words. Of these cohorts, 831 (8.7%) are unique. The maximum size of the remaining 684 non-unique cohorts is 602, 6.3% of the corpus. A cohort would contain 6.3 words in the mean, and would have an expected value of 133 words. English, then, has more unique cohorts than Swedish, and larger maximum size.

	Units	Cohorts	Unique	Max Size	Mean Size	Expected Size
SWEDISH						
WORD	9679	1336	682	393	7.24	105.27
no stress	9679	732	325	748	13.22	192.57
ROOT	6334	979	505	411	6.47	94.98
no stress	6334	575	264	411	11.02	140.69
ENGLISH						
WORD	9526	1515	831	602	6.29	133.04
no stress	9526	901	425	660	10.57	178.62
ROOT	6630	1172	653	601	5.66	115.92
no stress	6630	696	331	649	9.53	154.07
GERMAN						
WORD	9219	1663	974	433	5.54	84.10
no stress	9219	1114	599	543	8.28	121.40
ITALIAN						
WORD	8857	613	246	752	14.45	223.95
no stress	8857	426	157	1002	20.79	315.55
FRENCH						
WORD	11388	293	86	1296	38.87	459.35
no stress	11388	284	82	1296	40.10	460.40

TABLE 5. 2-class cohorts: consonant/vowel discrimination with and without stress for Swedish, English, German, Italian, and French.

	Cohorts	Unique	Max Size	Mean Size	Expected Size
SWEDISH					
WORD	13.8	7.0	4.1	.07	1.09
no stress	7.6	3.4	7.7	.14	1.99
ROOT	15.5	8.0	6.5	.10	1.50
no stress	9.1	4.2	6.5	.17	2.22
ENGLISH					
WORD	15.9	8.7	6.3	.06	1.40
no stress	9.5	4.5	6.9	.11	1.88
ROOT	17.7	9.8	9.1	.09	1.75
no stress	10.5	5.0	9.8	.14	2.32
GERMAN					
WORD	18.0	10.6	4.7	.06	.91
no stress	12.1	6.5	5.9	.09	1.32
ITALIAN					
WORD	6.9	2.8	8.5	.16	2.53
no stress	4.8	1.8	11.3	.23	3.56
FRENCH					
WORD	2.6	.8	11.4	.3	4.03
no stress	2.5	.7	11.4	.4	4.04

TABLE 6. 2-class Cohorts: Percentage of total corpora for Swedish, English, German, Italian, and French

The five most common cohorts from our English word corpus are the following: (C stands for consonant, ´V for stressed vowel, and V for unstressed vowel.)

Number of words	Cohort	Example
602	C-´V-C	bob
472	C-´V-C-C	box (x = ks)
411	C-´V-C-V-C	bodies
339	C-C-´V-C	block
231	C-´V-C-C-V-C	boxes

Combining stressed and unstressed vowels into one category yields only 901 cohorts, 59% of the discriminability as when stress is a factor. The number of unique cohorts decreases by almost a factor of 2. Maximum size, however, increases by 9.6% representing the addition of (secondary-stressed) function words to the one-syllable C-V-C pattern. Mean size increases by 70%, and expected size by 30%. Stress, then, serves a similar discriminatory function for both English and Swedish. An exception is its effect on maximum cohort size: a large effect (47% reduction) is seen for Swedish, but a smaller effect (9% reduction) for English.

There are about 2900 more words than roots in the English corpus, a reduction of 30%. This is about the same reduction as for Swedish. The number of cohorts reduces by 23% (31% for Swedish). Stress discrimination has about the same effect on maximum size for roots as for words in English, unlike Swedish in which the effect is much larger for words.

German 2-class and 3-class cohorts

The number of 3-class cohorts for the German corpus of 9,219 words is 1,663. This is an increase of 2.5% in discrimination over English, and an increase of 4.2% over Swedish. The number of unique cohorts is also a larger percentage of the total for German. Maximum size is somewhat larger than in Swedish, but not so large as in English. The mean size is smaller than either Swedish or English, and expected size is also smaller.

The five most common cohorts produced by a 3-class classification of the German corpus are the following:

Number of words	Cohort	Example
433	C-´V-C-V-C	baden
359	C-´V-C-C-V-C	bahnhof
232	C-´V-C	bahn
212	C-´V-C-C	bald
194	C-C-´V-C-V-C	bleiben

Stress does not serve quite as much of a discriminatory function in German as in either English or Swedish. Lack of stress information decreases the number of cohorts by 33%, as compared to 41% in English and 45% in Swedish. The number of unique cohorts decreases by 39% as compared to 49% and 52% for English and Swedish, respectively. The maximum size increases by 25%, intermediate between the low 10% of English and the high 90% of Swedish. Mean size without stress discrimination is 8.3, somewhat lower than the other two languages. Expected size is also somewhat lower.

Italian 2-class and 3-class cohorts

Italian has less than half as many cohorts as Swedish, English, and German, and much less than half as many unique cohorts. These cohorts, correspondingly, have a much larger maximum size, mean size, and expected size. There are only 613 3-class cohorts for the Italian corpus of 8,857 words, a reduction by half in discriminability from Swedish, the least discriminating (by this classification) of the three languages so far considered. It can be noted (see Table 3 - Phonetic Length) that the number of cohorts is not correlated with average phonetic length in the languages examined. Italian, for example, ranks second in word length and fourth in number of cohorts. The number of unique cohorts (246), is small as well, being only 2.8% of the entire Italian corpus. This can be compared with 7.0% for Swedish, 8.7% for English, and 10.6% for German. Mean size is 14.5, twice as large as for Swedish, and more than twice as large as for English and German.

The five most common cohorts in the 3-class classification of the Italian corpus are the following:

Number of words	Cohort	Example
752	C-V-C-'V-C-V	badare
671	C-'V-C-V	babbo
493	C-'V-C-C-V	banda
411	C-V-C-C-'V-C-V	balcone
273	C-V-C-V-C-'V-C-V	baricate

From this short list, it is obvious that Italian has a high frequency of multisyllabic words. The most frequent pattern is trisyllabic. The most common one-syllable cohort, C-'V, ranks only 26 and contains only 59 words.

Reduction in number of cohorts due to lack of stress knowledge (31%) is similar to German, at 33%, as is the reduction in number of unique cohorts (36%; in German, 39%). The increase in maximum cohort size (33%), like German (25%), is also intermediate between the extremes of English and Swedish. Mean size, 21, and expected size, 316, as in the case including stress, is more than twice as large as for German.

French 2-class and 3-class cohorts

The results for French are in the same direction as for Italian, but doubly exaggerated. French has the smallest number of consonant/vowel cohorts of the five languages and the smallest number of unique cohorts. Discrimination by this type of classification results in only 293 cohorts in French (2.6% of the total number of words) as compared to the next highest number in Italian (751 or 8.5%) and to the highest number of cohorts in German (1663 or 18.0%). The maximum French cohort size is 1,296, the mean size, 38.9, and the expected size, 459.4. These contrast with the low values for German where expected size, for example, is only 84.1.

The most common cohorts in the 3-class classification of the French corpus are as follows:

Number of words	Cohort	Example
1296	C-V-C-'V	deja
797	C-V-C-'V-C	duquel
783	C-'V-C	donc
762	C-V-C-C-'V	monsieur
537	C-V-C-V-C-'V	balayer

Reduction in number of cohorts due to lack of stress knowledge is insignificant in French, and maximum size does not change at all since French is a fixed-stress language. Thus French differs radically from the other languages in that word stress pattern is not a discriminating factor on the word level. Of course, if stress could be detected in running speech, it would be useful in segmenting the speech into words.

Review of 2-class and 3-class cohorts

Reviewing, we see (Table 7) that the German vocabulary is best classified by a consonant/vowel discrimination. German has the largest number of cohorts for both 2-class and 3-class classifications, and leads in the number of unique cohorts for both classes as well. The maximum cohort size for German is, correspondingly, smallest (except for Swedish in the 3-class instance). English and Swedish follow, taking the middle ground between German and Italian. The French vocabulary falls well away from Italian, and is least well classified by a consonant/vowel discrimination, having the fewest cohorts and largest cohort size. The only exception in this otherwise strictly ordered chart is the previously mentioned unusual discriminability of the stressed vowel in Swedish in decreasing the maximum cohort size.

%	Number of cohorts as percentage of total number of words in corpus		Number of unique cohorts as percent of total number of words in corpus		Max. cohort size as percentage of total number of words in corpus	
	C,V,'V	C,V	C,V,'V	C,V	C,V,'V	C,V
18	German					
	English					
15						
	Swedish					
12		German				
			German		French	French+ Italian
9		English	English		Italian	
		Swedish				Swedish
	Italian		Swedish			English
6				German	English	German
		Italian			German	
				English	Swedish	
3	French		Italian	Swedish		
		French		Italian		
			French	French		
0						

TABLE 7. Comparison of percentages for Swedish, English, German, and Italian: 2-class cohorts (C = consonant, V = vowel, 'V = stressed vowel)

6-class and 7-class cohorts

Narrowing our classification by specifying the five consonant types: nasal, other sonorant consonant, stop, strong fricative, and weak fricative, produces a significant increase in discriminability for all five languages. The figures for the five languages are given in Tables 8 and 9. Increase in discriminability due to the increase in number of phoneme categories from 2 to 6 (or 3 to 7 counting the stressed/unstressed vowel discrimination) is given in Table 10. A comparison of the percentages given in Table 9 appears in Table 11.

Referring to Table 10, we see that the number of word cohorts increases by an amount corresponding to between 22.7% (French, with and without stress discrimination) and 52.7% (German, without stress discrimination) of the number of words in the total corpora. Unique cohorts increase by between 12.8% (French, with and without stress discrimination) and 46.1% (German, with stress discrimination). It is interesting to note that the increase in number of unique cohorts represents a high percentage of the increase in total number of cohorts. For

German, English and Swedish, over 80% of the increase in number of cohorts is accounted for by unique (one-member) cohorts. The corresponding percentages for Italian and French, with and without stress discrimination, are between 56% and 64%. (See Table 10 for exact values.)

Maximum size is decreased most for French by this classification, decreasing to 170 words from 1296 words in the consonant/vowel classification. This decrease corresponds to 9.9% of the corpus. The next highest decrease in maximum size is found in the category of English roots, both with and without stress discrimination. The smallest decrease is in Swedish words. Mean size is also decreased most in French by a five-consonant classification. The mean size of English root cohorts with stress discrimination decreases the least.

We find highly significant decreases in expected cohort size for all languages by this narrower classification, most decreasing by more than an order of magnitude. French displays the most extreme change, from an expected cohort size of 460 in the consonant/vowel classification to 25 in the five-consonant/vowel classification.

Referring to Table 11, we see that the relative phonemic descriptiveness of the five languages by the 6-class coding is about the same as with the 2-class coding. German is best described, having 68.2% as many cohorts as words, and French and Italian are least well described, having 25.3% and 44.2% as many cohorts as words respectively. English and Swedish words (but not roots) change places from the 2-class coding, with Swedish words being the better described with the 5-consonant type discrimination than English. The same ordering holds for the number of unique cohorts as well. The unusual discriminability of the stressed vowel in the 2-class coding of Swedish is overshadowed in the 6-class coding by the discriminability of the 5-consonant classification. The five languages, according to this 6-class coding, then, are strictly ordered within the three categories of number of cohorts, number of unique cohorts, and maximum cohort size with the single exception that Italian and English are exchanged for maximum cohort size.

It is clear from these results that the ability to distinguish between stressed and unstressed vowels is much less powerful in lexical search than the ability to make a 5-class discrimination for consonants. Whereas the 5-class consonant division decreases the expected size of cohorts by at least 80 words (German) and at most 435 words (French), the stressed/unstressed vowel distinction does not affect the expected size for French, and decreases the expected cohort size by an average of only 43 words for the other four languages. It is also interesting to note that the cohorts which coalesce without stress discrimination are not the largest cohorts. Disregarding French, in which stress does not play a discriminatory role, we can see that the number of cohorts is decreased by at least 305 (German) and at most 689 (Swedish) by lack of stress discrimination, while the maximum size increases by 7 at the most, actually remaining fixed for Swedish.

	Units	Cohorts	Unique	Max Size	Mean Size	Expected Size
SWEDISH						
WORD	9679	6100	4871	57	1.59	5.10
no stress	9679	5411	4127	57	1.79	6.53
ROOT	6334	3836	3097	64	1.65	6.70
no stress	6334	3409	2595	66	1.86	7.61
ENGLISH						
WORD	9526	5305	4187	121	1.80	9.58
no stress	9526	4725	3544	126	2.02	11.15
ROOT	6630	3671	2930	122	1.81	11.31
no stress	6630	3221	2437	127	2.06	13.18
GERMAN						
WORD	9219	6283	5228	29	1.47	3.66
no stress	9219	5978	4825	33	1.54	4.02
ITALIAN						
WORD	8857	3863	2321	91	2.29	9.13
no stress	8857	3400	1937	100	2.61	11.33
FRENCH						
WORD	11388	2878	1544	170	3.96	25.13
no stress	11388	2871	1540	170	3.97	25.22

TABLE 8. 6-class Cohorts for Swedish, English, German, Italian and French: Vowel, Nasal, Other Sonorant Consonant, Strong Fricative, Weak Fricative, and Stop Classification.

	Cohorts	Unique	Max	Mean	Expected
SWEDISH					
WORDS	63.0	50.3	.6	.02	.05
no stress	55.9	42.6	.6	.02	.07
ROOT	60.6	48.9	1.0	.03	.11
no stress	53.8	41.0	1.0	.03	.12
ENGLISH					
WORDS	55.7	44.0	1.3	.02	.10
no stress	49.6	37.2	1.3	.02	.12
ROOT	55.4	44.2	1.3	.03	.17
no stress	48.6	36.8	1.3	.03	.20
GERMAN					
WORDS	68.2	56.7	.3	.02	.04
no stress	64.8	52.3	.4	.02	.04
ITALIAN					
WORDS	43.6	26.2	1.0	.03	.10
no stress	38.4	21.9	1.1	.03	.13
FRENCH					
WORDS	25.3	13.6	1.5	.03	.22
no stress	25.2	13.5	1.5	.03	.22

TABLE 9. 6-class Cohorts for Swedish, English, German, Italian, and French in percentage of each corpus: Vowel, Nasal, Other Sonorant Consonant, Strong Fricative, Weak Fricative, and Stop Classification

	Increase in Cohorts		Increase in Unique Cohorts		Decrease in Maximum Size		Decrease in Mean Size	Decrease in Exp. Size
	(Number)	(%)	(Number)	(%)	(Num)	(%)	(Number)	(Number)
SWEDISH								
WORD	4764	49.2	4189	43.3	336	3.5	5.65	100.17
no str.	4679	48.3	3802	39.2	691	7.1	11.43	186.04
ROOT	2857	45.1	2592	40.9	347	5.5	4.82	88.28
no str.	2834	44.7	2331	36.8	345	5.5	9.16	133.08
ENGLISH								
WORD	3790	39.8	3356	35.3	481	5.0	4.49	123.46
no str.	3824	40.1	3119	32.7	534	5.6	8.55	167.47
ROOT	2499	37.7	2277	34.4	479	7.8	3.85	104.61
no str.	2525	38.1	2106	31.8	522	8.5	7.47	140.89
GERMAN								
WORD	4620	50.2	4254	46.1	404	4.4	4.07	80.44
no str.	4864	52.7	4226	45.8	510	5.5	6.74	117.38
ITALIAN								
WORD	3166	35.7	2067	23.3	506	5.8	9.53	157.70
no str.	2929	33.1	1787	20.2	698	7.9	14.28	223.38
FRENCH								
WORD	2585	22.7	1458	12.8	1126	9.9	34.91	434.23
no str.	2587	22.7	1458	12.8	1126	9.9	36.13	435.18

TABLE 10. Increase in discriminability from 2-class (consonant, vowel) to 6-class (nasal, other sonorant, strong fricative, weak fricative, stop, vowel) classification. (Increase given in both number and percentage of total corpus for first three categories, in number only for last two categories.)

%	Number of cohorts as percentage of total number of words in corpus		Number of unique cohorts as percent of total number of words in corpus		%	Max. cohort size as percentage of total number of words in corpus	
	5C,V,'V	5C,V	5C,V,'V	5C,V		5C,V,'V	5C,V
70	German						
		German					
	Swedish						
60							
			German				
	English	Swedish					
				German			
50		English	Swedish				
	Italian		English	Swedish			
40							
		Italian		English	1.5	French	French
						English	English
30					1.2		
							Italian
	French	French	Italian			Italian	
					.9		
20				Italian			
			French	French	.6	Swedish	Swedish
10							German
					.3	German	
0					0		

TABLE 11. Cohort statistics in percent for the different languages. 6-class cohorts (C = consonant, V =vowel,'V =stressed vowel)

Correlating cohort estimates to total number of cohorts

As earlier mentioned, the number of cohorts differs substantially among the five languages explored. This variation also holds for the number of unique cohorts as well as the maximum and the expected cohort size. However, if we relate the latter three measures to the total number of cohorts for each language, we may note some interesting correlations. We should keep in mind that the results below are for about equally sized sublexicons of different languages so that expressions like 'increasing the number of cohorts' should not be taken literally though they facilatate the discussion.

The upper two plots of Fig. 3 show the number of unique cohorts as a function of the total number of cohorts. We see a very strong linear relation for the pooled stressed and unstressed data. At the 95% confidence level the confidence interval of the correlation coefficient is (0.984, 0.999) for the 2-class and 3-class cohorts and (0.988, 0.999) for the 6-class and 7-class cohorts. The number of unique cohorts is thus a linear function of the total number of cohorts.

The intersection of the regression line with the abscissa is 190 for the 2-class and 3-class cohorts and 1550 for the 6-class and 7-class cohorts. This means, if it is possible to extrapolate the regression lines, that if a sublexicon of ten thousand words of a language were to have 190 (1550) cohorts or less, none of them would be unique.

The slopes of the two regression lines mentioned above are 0.63 and 1.08, respectively. This indicates that as the number of cohorts increases more of them will be unique when using a more detailed phonetic categorisation, e.g., five consonantal categories instead of one. In fact, for the five consonant case, the unique cohorts grow at about the same rate as the total number of cohorts. However, adding stress as in the 3- and 7-class cohorts, gives a rise in the number of unique cohorts that is very much along the regression line of the 'unstressed' cohorts, as reported above. It seems that including stress information has a smaller influence on the number of unique cohorts than a finer consonantal differentation.

The lower two plots of Fig. 3 show the expected cohort size as a function of the total number of cohorts. The size decreases as the number of cohorts increases, which is natural. The confidence interval of the correlation is (-0.60, -0.97) for the 2/3-class cohorts and (-0.52, -0.97) for the 6/7-class cohorts at a 95% confidence level.

Also, the maximum cohort size is negatively correlated to the total number of cohorts and exhibits about the same correlation statistics as the expected size.

Prediction

One can also investigate how lexical search is facilitated by partial knowledge of phonemic word structure or by grammatical knowledge. Presented here are several studies concerning such possible facilitation.

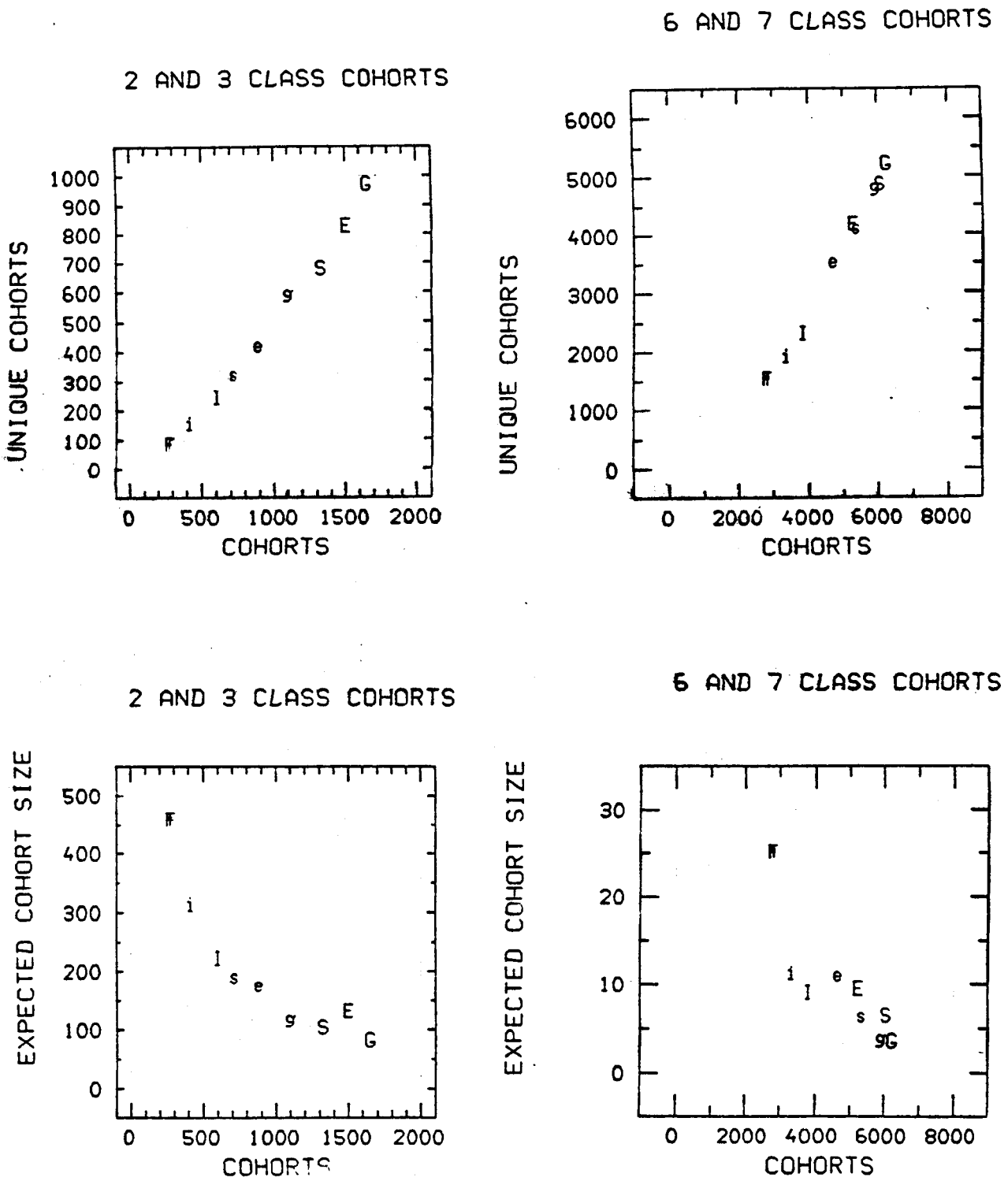


Fig. 3. Unique cohorts (top) and expected cohort size (bottom) as a function of number of cohorts. Letters in the plots are the initial letters of the different languages. Upper/lower case denotes stress considered/not considered in the classifications.

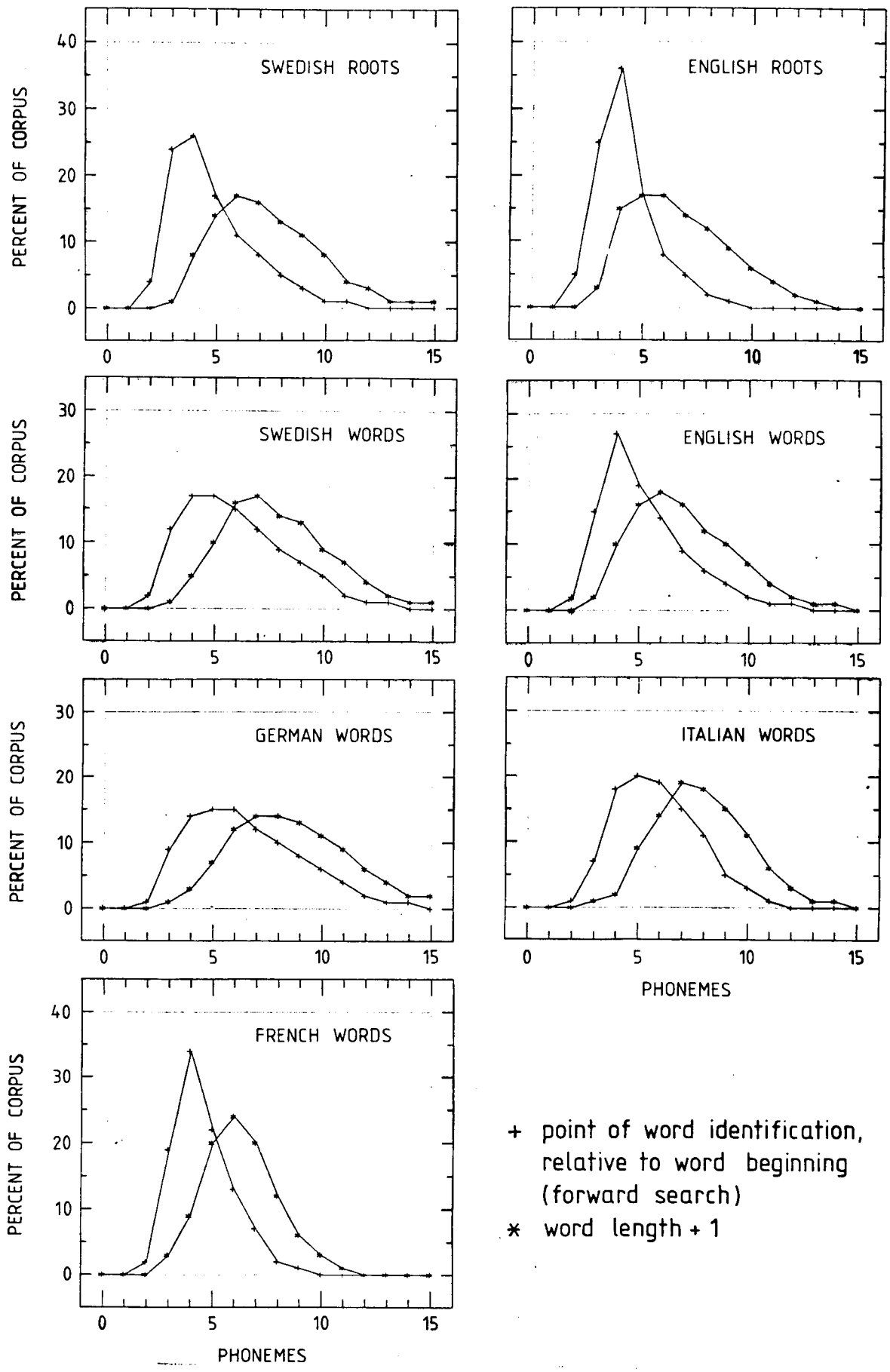


Fig. 4. Distribution of the point of identification in the word and root corpora (forward search).

Fig. 4 provides information on lexical search reduction given a phonemic transcription and assuming that word identification takes place from the word beginning. The plots here can be viewed as a representation of lexical redundancy. The rightmost curve in each plot gives the percent of words in the corpus which have a particular word length. Because an end-of-word marker, such as a space, is necessary to completely specify a word, the abscissa represents "actual word length + 1" in the graphs. The leftmost curve shows the percent of words in the corpus which are uniquely specified by their first n phonemes, n being the co-ordinate on the abscissa. A comparison of these two curves indicates lexical redundancy: the phonemes that need not be specified to determine the vocabulary uniquely.

The median (most common) number of phonemes for word identification, or unique prediction from the beginning, in all five languages is two less than the median number of phonemes for actual word specification (including an end-of-word marker). The most common length for a word is 5 for English and French, 6 for Swedish and Italian, and 6 or 7 for German. The prediction curves for English and French both exhibit noticeable peaks at 4 phonemes, differing by at least 10% of the corpus from other word identification points. For Swedish, German, and Italian, the most common points of identification form a cluster of three points around 5 phonemes.

An additional interesting point apparent in Fig. 4 is the symmetricity of French word length in the corpus. The French corpus shows an even distribution of actual word length around 5 phonemes. The corpora of the other languages have a greater precentage of long words. Referring to Table 3, we see that, indeed, mean word lengths are about one phoneme greater than the median for the four languages other than French, and both values near 5 for French.

Fig. 5 presents the point of word identification relative to the primary stressed vowel. In French, this point most commonly occurs at the stressed vowel. An inspection of the most common cohorts in the other languages suggests that most polysyllabic words are uniquely specified either by the end of the syllable containing the stressed vowel or by the vowel in the following syllable.

An examination has been made of the relative predictive power of the initial and final parts of words and roots in our corpora in view of the general expectation that word beginnings are more distinguishable from each other, i.e., phonotactically richer, more complex, than the final part of words. The results of a number of psycholinguistic experiments have led to this expectation.

Experiments leading to the conclusion that polysyllabic words are accessed via their first syllable in perception (visual tachistoscopic experiments) were performed by Taft & Forster (1976). Through experiments with interference effects, they were further led to stipulate the independent status of first syllables in the mental lexicon. Extensive work by Cole & Jakimik provides further evidence for the theory of left-to-right processing of words and the importance of the first syllable in

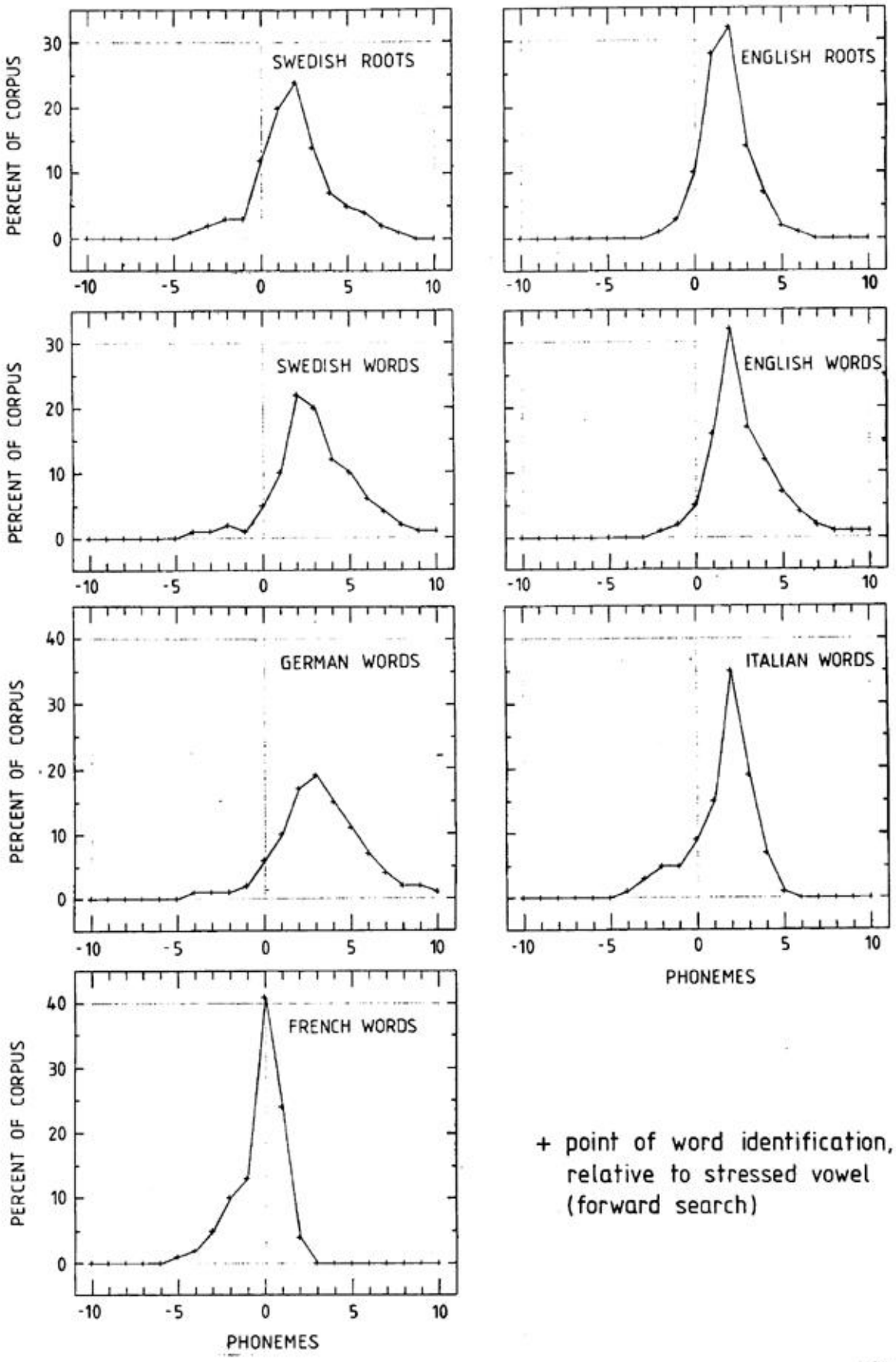


Fig. 5. Distribution of the point of identification in the word and root corpora in relation to the stressed vowel (forward search).

the procedure of lexical access. Their work revealed that for both stressed and unstressed first and second syllables, mispronunciations were detected faster in the second syllable, although more often in the stressed syllable. This evidence led to their hypothesis that word candidates, in perception, are accessed from the sounds which begin them, regardless of their stress pattern (Cole & Jakimik, 1980). A strong statement for left-to-right word processing is also supplied by Welsh (1980). His work leads him to the conclusion that a stressed-syllable based theory of lexical access is inadequate, and that speech is actually processed left to right.

In a study of lexical access via word beginnings and endings in Dutch, Nooteboom (1981) presents evidence for the superiority of word beginnings. He predicts that lexical items will generally be found to carry more information early in the word, realized in a greater variety of phonemes and phoneme combinations.

The results of our word beginning vs. word ending study appear in Table 12. For each language, two types of information is given. The number of roots or words in the corpus of each language is given at the top of the table. The remainder of the table contains information on the effectiveness of prediction given the beginning or the end of a word or root. This effectiveness is expressed in terms of the percent reduction in letters or phonemes that need to be specified in order that the word be uniquely determined in the corpus. For example, the number "19" that appears in the LETTER DATA for SWEDISH WORDS under the heading "Reduction Fore" indicates that specifying the first 81% of the letters of a Swedish word, on the average, results in its unique specification, i.e., the last 19% of the letters need not be specified. We will refer to this process as forward prediction. Likewise, the "32" in the column to the right of the "19" (Reduction Back) indicates that if one were to specify the last 68% of the letters of a Swedish word, that the remaining 32% would be uniquely determined. This process will be referred to as backward prediction.

The efficiency of this type of prediction across languages is quite similar. Orthographically, forward prediction varies between 19% (Swedish) and 23% (German). That is, on the average, the first 77% to 81% of the letters of a word specify it uniquely within the language corpora. Backward prediction varies from 29% (French) to 32% (German and Swedish). The figures are similar, but higher, for the phonetic representations. Forward prediction in words is, on the average, 26%, and backward prediction, 34%. (If stress is not to be predicted, these figures are decreased by 2% or 3%.)

One cause for the greater discriminability of backward prediction can be seen by looking at the results for roots in English and Swedish. These results are quite similar for backward and forward prediction, differing by an average of 2.5% as compared to 8.0% for words. Thus, we see that forward prediction is hampered for words by the fact that two

words having the same root with two different suffixes cannot be distinguished from each other until at least the first letter of the suffix is encountered.

The fact that forward prediction is no better than backward prediction for roots, however, indicates that our intuitions about beginnings of words being more discriminatory were unfounded, at least in the two languages examined. This may be a reflection of the large number of one-syllable roots which tend to have a symmetric structure of increasing sonority towards the central vowel. However, we note very little phonetic richness in the first syllables of the large cohorts in both English and Swedish. For example, there is only one initial cluster in the larger cohorts, that of a stop followed by a liquid or glide. It would be interesting to compare these results with a similar study for Italian and French, which have more open (and therefore non-symmetric) syllables.

If, in fact, word beginnings are superior to word endings for lexical access, as suggested by Nooteboom's results, this superiority must therefore be due to the coding or storage of lexical items in the mental lexicon and/or to the retrieval procedure. It appears not to be dependent upon discriminatory orthographic or phonetic word structure. The equally discriminatory phonetic structure of word and root endings may point to an equally important function, that of confirming and sometimes rejecting, word candidates, as well as providing grammatical information. Nooteboom has also drawn our attention to word endings, and calls the function associated with processing them the "monitoring component" of word recognition. He advises that the possibly redundant final part of a word is not to be considered superfluous.

Evidence for the importance of the stressed syllable in word recognition has been presented by a number of investigators. Garnes & Bond (1980), in a study of word errors, report that word stress is typically not in error. In addition, they report that only 4.1% of their error corpus involved misperception of stressed vowels. Experiments by Cole, Jakimik, & Cooper (1978) showed that listeners could detect a mispronunciation in stressed syllables 82% of the time, whereas such mispronunciations were only detected in unstressed syllables at a rate of 47%. Other perceptual research which has shown the relevance of word stress for the organization of the mental lexicon includes work by Engdal (1978) and Browman (1977).

In lexical studies designed to aid in speech recognition, it has also been found that word stress and other prosodic characteristics of words are very useful in limiting the number of possible word candidates, i.e., in reducing cohort size. Waibel (1982) finds syllable duration and a ratio of unvoiced to voiced segments to be complementary to the word's phonetic specification in reducing a lexical search space. Huttenlocher & Zue (1983) present a classification comprised of (possibly alternate) stress patterns plus the phonetic specification of the stressed syllable as a robust and useful representation. They also demonstrate that in their corpus of English, the phonotactic information

in stressed syllables is much more discriminatory than in unstressed syllables.

If we assume that our prosodic information is reliable, and that we can predict the number of syllables and stress pattern of a word, our results improve remarkably for Swedish, English, and German, particularly for forward prediction. For words, forward prediction improves by 12.6% with stress given, and for roots, by 15%. Backward prediction improves as well, by 8.4% for words and by 7.5% for roots. An interesting change occurs in the results for roots. Given the stress pattern, forward prediction excels over backward prediction in both English and Swedish, reaching 49% in Swedish and 51% in English. That is, if the number of syllables and stress positions are known, roots in the two languages are predictable from their beginning half. The figures are high for backward prediction also, 46% for Swedish and 42% for English.

It is generally assumed that syntactic constraints play a significant role in lexical access within sentences as well. If this is so, we should see a noteworthy increase in predictibility if a word's part of speech (word class) is known. For the Swedish corpus, in which part of speech is specified, we see that forward prediction is 30% for words and 44% for roots. This represents an increase in predictive power of 6% and 9% respectively. Backward prediction is aided by only 2% or 3% given parts-of-speech information. These figures are notably smaller than the corresponding figures for known stress pattern, 15% and 14% for forward prediction and 7% for both instances of backward prediction. Thus we see that, for Swedish, knowledge of the word stress pattern is much more helpful in lexical access than the grammatical knowledge provided by the part of speech of a word.

Conclusions

In this study, we have presented data on the structure of large dictionaries from several European languages. Subclassification of the vocabularies in terms of stress, vowel and one or five consonantal categories shows interesting differences between the languages. A simple measure, such as the number of cohorts, can be used to characterize the languages, and correlates well with other aspects such as the number of unique cohorts, and the mean and expected cohort size. This implies that the natural languages studied use the lexical space in similar ways within the structural constraints. The constraints include typical word length, syllable and stress structure, phonemic inventory and phonotactics.

One general conclusion is that the lexical space is extremely unevenly exploited. Considering a model language with a fixed word length of six phonemes and no further constraints yields (in a six-category classification) over 45,000 unique patterns (6 raised to the sixth power). In the natural languages with this classification, we observe between 1540 and 4825 unique patterns, and an expected cohort size of from 4 to 25. This indicates a tendency towards more standardized patterns in natural language, a clustering within the lexical space.

	SWEDISH		ENGLISH		GERMAN		ITALIAN		FRENCH	
Num. of WORDS	9679		9526		9233		8857		11388	
Num. of ROOTS	6334		6630		--		--		--	
	Reduction Fore	Back	Reduction Fore	Back	Reduction Fore	Back	Reduction Fore	Back	Reduction Fore	Back
LETTER DATA										
WORD	19	32	22	31	23	32	21	30	20	29
ROOT	27	33	29	31	--		--		--	
PHONETIC DATA										
WORD	24	36	25	33	25	37	25	30	29	30
no str.	22	33	23	32	24	33	21	29	23	30
ROOT	35	39	35	34	--		--		--	
no str.	31	36	32	33						
STRESS KNOWN										
WORD	39	43	35	40	38	48	32	36	32	37
ROOT	49	46	51	42	--		--		--	
PARTS OF SPEECH KNOWN										
WORD	30	38	--		--		--		--	
ROOT	44	42	--		--		--		--	

TABLE 12. Forward and backward prediction in Swedish, English, German, Italian, and French. Reduction = 1 - (units not predicted) / (word length + 1)

Another way to look at the unevenly distributed space is to compare expected and mean size. If all cohorts contained the same number of members, the two measures would be the same. In our material (the 6-class classification), the expected size is typically 4.5 times as large as the mean size, indicating that some patterns are favored.

Still, the study of forward/backward prediction of words within the corpora shows a considerable degree of redundancy, especially if the stress pattern, and to a smaller degree, parts of speech, is known. There thus appears to be a complementary balance of standardization and discriminability in the lexical space of these natural languages.

Knowledge of the lexical structure of large samples of natural languages is interesting for models of lexical access. The distribution of lexical items within the lexical space, and their similarity to one another are meaningful considerations in constructing test corpora and explaining the process by which words are retrieved. Such knowledge also has immediate application in large-vocabulary speech recognition systems since it is useful to know how many actual words can be expected to correspond to a recognized pattern and how detailed the decision process must be. A measure of similarity for such a system has been suggested by Makino, Wakita, & Applebaum (1984).

Our future plans include the development and use of model languages in which a variety of structural assumptions is made, to study the effects of constraints such as word length, distribution of word length, phonotactics, and stress systems or word tone. From these studies, we expect to learn more about how natural languages differ from a random collection of phonetic strings, and how these organizing factors might facilitate human communication.

References

Allen, S. (1970): Frequency Dictionary of Present-Day Swedish, Almqvist & Wiksell, Stockholm.

Bond, Z.S. & Garnes, S. (1980): "Misperception of fluent speech," in ed. R. Cole) Perception and Production of Fluend Speech, Erlbaum, Hillsdale.

Bortolini, V., Tagliavini, C., & Zampolli, A. (1971): Lessico di frequenza della lingua italiana contemporanea, IBM Italia.

Browman, C. (1979): "Word-internal order: Evidence from language errors," LSA, Los Angeles.

Carlson, R., Granstrom, B., & Hunnicutt, S. (1982): "A multi-language text-to-speech module," Conference Record, IEEE-ICASSP, Paris.

Cole, R. & Jakimik, J. (1980): "How are syllables used to recognize words," J.Acoust.Soc.Am. 67, no. 3.

Cole, R., Jakimik, J., & Cooper, W. (1978): "Perceptibility of phonetic features in fluent speech," J.Acoust.Soc.Am. 64, no. 1.

Engdahl, E. (1978): "Stress and rhythm in speech production and perception," workshop at U. of Massachusetts at Amherst.

Engwall, G. (1984): Vocabulaire du roman francais (1962-1968), Almqvist & Wiksell, Stockholm.

Halle, M. & Vergnaud, J.-R. (1980): "Three-dimensional phonology," J. of Linguistic Research.

Huttenlocher, D. & Zue, V.W. (1983): "Phonotactic and lexical constraints in speech recognition," MIT Speech Communication Group Working Papers, Volume III.

Kucera, H. & Francis, W.N. (1967): Computational Analysis of Present-Day American English, Brown University Press, Providence.

Makino, S., Wakita, H., & Applebaum, T. (1984): "Lexical analysis for word recognition based on phoneme-pair differences," written version of paper presented at Oct. ASA meeting, Minneapolis.

Marslen-Wilson, W.D. (1978): "Sequential decision processes during spoken word recognition," MS from papers pres. at Psychonomic Society, and Amherst Workshop on "The Mental Representation of Phonology."

Nooteboom, S.G. (1981): "Lexical retrieval from fragments of spoken words: beginnings vs. endings," J. of Phonetics 9, pp. 407-424.

Rosengren, I. (1972): Ein Frequenzworterbuch der deutschen Zeitungssprache, CWK Gleerup, Lund.

Shipman, D.W. & Zue, V.W. (1982): "Properties of large lexicons: implications for advanced isolated word recognition systems," ICASSP '82, IEEE ASSP, pp. 546-549.

Taft, M. & Forster, K.I. (1976): "Lexical storage and retrieval of polymorphemic and polysyllabic words," J. of Verbal Learning and Verbal Behavior 15, pp. 607-620.

Waibel, A. (1982): "Towards very large vocabulary word recognition," CMU-CS-82-144.

Welsh, A. (1980): "The effects of word accent and syllabic stress on the processing and perception of continuous speech," U. of Chicago dissertation.