Measuring the Usefulness (Functional Load) of Phonological Contrasts

Dinoj Surendran*
University of Chicago

Partha Niyogi †
University of Chicago

Frequency counts are a measure of how much use a language makes of a linguistic unit, such as a phoneme or word. However, what is often important is not the units themselves, but the contrasts between them. A measure is therefore needed for how much use a language makes of a contrast, i.e. its functional load. We generalize previous work in linguistics and speech recognition and propose a family of measures for the FL of several phonological contrasts, including distinctive features (e.g. voicing, aspiration) and suprasegmental features (e.g. tones, stress). Furthermore, we empirically test it for robustness e.g. to changes of corpora. Examples are provided in several application areas (historical linguistics, child language acquisition, speech recognition) and languages (Cantonese, Dutch, English, German, Mandarin).

1 Introduction

> "The term functional load is customarily used in linguistics to describe the extent and degree of contrast between linguistic units, usually phonemes. In its simplest expression, functional load is a measure of the number of minimal pairs which can be found for a given opposition. More generally, in phonology, it is a measure of the work which two phonemes (or a distinctive feature) do in keeping utterances apart – in other words, a gauge of the frequency with which two phonemes contrast in all possible environments" – King (1967)

This paper describes a method to measure how much use a language makes of a contrast to convey information, i.e. the functional load (FL) of the contrast.

The concept of FL goes back to the 1930s. However, existing definitions are so limited that researchers who want to measure FL often cannot. For example, Pye, Ingram and List (1987) speak of the need to make explicit a phonological model of acquisition which "predicts that children will attempt to build phonemic contrasts on the basis of maximal opposition within the language". They go on to say :

> "We need a rigorous definition of maximal oppositions that specifies the relative strengths of different features within any language. . . . The frequency of consonants across lexical types is an imperfect guide to children's phonological systems because it refers to isolated segments rather than oppositions." — Pye, Ingram and List (1987)

Ingram (1989) suggests a method of computing FL, based on counts of minimal pairs, but as So and Dodd (1995) point out, it "does not include other aspects of phonology

* Computer Science Department, 1100 E58th St, Chicago IL 60637. Email : dinoj@cs.uchicago.edu
† Computer Science Department, 1100 E58th St, Chicago IL 60637. Email : niyogi@cs.uchicago.edu

that might contribute, relatively, to the functional loading of consonants: vowel, syllable structure, stress and tone."

The framework we propose *does* measure the FL of consonant oppositions, and several other contrasts, while taking into consideration word and syllable structure, stress and tone. The use of the term 'contrast' in this paper is broader than standard, encompassing phoneme oppositions (binary or not), distinctive features (again, binary or not), suprasegmental features and even phonological rules such as phoneme deletion in certain contexts. This permits researchers with the appropriate corpora to answer questions like these:

- Is it more important to correctly hear the tone or the vowel in Cantonese?
- Does Hindi make more use of aspiration or voicing?
- How much information is lost due to vowel reduction in unstressed syllables?
- If second-language speakers have trouble learning contrasts that are not present in their native language, e.g. the [l]-[r] distinction in English for Japanese speakers, how badly off are they?

.

Section 2 summarizes the history of FL in linguistics and related work in speech recognition. Section 3 defines our FL measure and Sections 4 and 5 demonstrate the range of its applicability with several examples. Section 6 tests its robustness to the approximations required to compute it. Section 7 is similar, investigating whether corpora that are not representative of continuous speech, such as word-frequency lists with citation form pronunciations and written frequencies, give usable FL values. Sections 8 to 11 give detailed examples of applications in linguistic typology, historical linguistics, language acquisition and speech recognition. Applications come with actual computations with corpora for Cantonese, Dutch, English, German and Mandarin. Section 12 discusses the interpretation of FL values, especially in light of their being relative values rather than absolute.

As we have not managed to eliminate enough notation from them, readers may wish to skim Section 3 and skip Section 6 on a first reading.

2 Previous Work

2.1 FL in the Linguistics community

Languages use contrasts of features to convey information. The concept of 'amount of use a language makes of a contrast' arose in linguistics early in the 20th century, and the term *functional load* for a measure of it can be found in the writings of the Prague School (Mathesius, 1929; Trubetzkoy, 1939). The term 'contrast' was nearly always taken to mean 'binary opposition of phonemes'.

Martinet (1955) popularized the concept, positing it as an important factor in sound change. This has been disputed; a quantitative corpus-using study by King (1967) found no evidence for FL playing a role in the context of phonological mergers. But finding no evidence for X and finding evidence against X are different things, and the reader interested in the debate is referred to Peeters (1992), Lass (1980; 1997), and to the example in the case of a recent merger in Cantonese in Section 9.

Meyerstein (1970) notes, in his survey of the topic, that FL is easy to define intuitively but hard to define precisely. The first person to propose a formula for it was Hockett (1955). His formula was only meant for the FL of the opposition of a pair of phonemes, say x and y, in a language L[1]. The absence of this opposition creates a language L_{xy} just like L but with x and y collapsed into a single phoneme. For example, in $English_{bp}$ the verbs 'bat' and 'pat' have the same pronunciation.

Hockett assumed that any language could be modelled by a sequence of phonemes, and its informational content represented by the entropy H of a language. (The definition and computational details of H are described in Section 3. For now, we just need to know that H is the number of bits of information transmitted by the language.) The closer $H(L)$ and $H(L_{xy})$ are, the less the information lost when the $x - y$ opposition is lost from L, and hence the less the reliance of L on it. Therefore he proposed :

$$FL_{Hockett}(x,y) = \frac{H(L) - H(L_{xy})}{H(L)} \tag{1}$$

The crucial part of the definition is the numerator, which clearly illustrates the notion of 'Functional Load as Information Loss'. The denominator is a normalizing factor that makes it interpretable as the *fraction* of information lost when the opposition is lost.

Other definitions of FL were also proposed by linguists, some information theoretic e.g. Kučera (1963) and some not e.g. Greenberg (1959), King (1967).

2.2 Measuring constrasts' use in the Speech Recognition community

Interest in FL among linguists waned after 1970. When it arose in a different guise in the automatic speech recognition (ASR) community in the 1980s, nobody noticed — in either community. Ironically, several linguists had previously predicted that FL would be useful for ASR research.

One reason that the connection was not spotted was due to the very different way the concept originated in ASR. We now describe this. It was thought possible to build broad-class recognizers for a language L that could tell with very high accuracy that a stop (or fricative or vowel or...) had occurred, even if they could not recognize exactly which stop it was. The hope was that this would be enough to recognize most words. What was required was a measure of how well such a recognizer worked, or at least an estimate of how well it would work once it was made.

Such a recognizer could be represented by a partition θ of phonemes whose classes were the broad classes it recognized well. θ induces a partition W_θ of the set W of words in L. The elements of W_θ are word classes, or *cohorts* in the notation used by Shipman and Zue (1982). For example, if θ is the vowel-glide-other partition, the words 'yak', 'yap', 'wit', etc end up in one cohort, the words 'chopping', 'jotted', 'fatten', etc in another cohort, and so on.

Several measures were proposed for the effectiveness[2] e of a recognizer represented by a partition θ. Since larger cohorts are clearly worse, Shipman and Zue (1982) proposed that effectiveness be measured by the average cohort size: $e(\theta) = \frac{1}{|W_\theta|}\sum_{C\in W_\theta} n(C)$, where $n(C)$ is the number of words in cohort C. Huttenlocher (1985) pointed out that this did not account for word frequencies, and proposed that e be the expected cohort size: $e(\theta) = \sum_{C\in W_\theta} P(C)n(C)$. Note that $P(C) = \sum_{w\in C} P(w)$ is the probability that a random word is in cohort C, where $P(w)$ is the probability of word w. However, Carter (1987) noted that this did not adequately take into account word frequencies. He proposed that the expected cohort entropy be used instead: $e(\theta) = \sum_{C\in W_\theta} P(C)H(C)$. Note that

1 Wang (1967) generalized Hockett's definition to the opposition between elements of a set of phonemes.

2 The three proposed definitions summarized here all share the property that the higher they are, the worse the recognizer is. To be pedantic, they measure *in*effectiveness rather than effectiveness.

the entropy $H(C) = -\sum_{w \in C} \frac{p(w)}{p(C)} \log_2 \frac{p(w)}{p(C)}$ of cohort C is the uncertainty in trying to tell apart words in it; it is harder to do so when $H(C)$ is higher.

It turns out that Carter's definition of $e(\theta)$, the expected uncertainty given that one can tell which cohort a word is in, is the same as the conditional entropy given the same conditions, i.e. $H(W|W_\theta) = H(W) - H(W_\theta)$. As this is not obvious, his direct proof of it is reproduced below for completeness.

$$\begin{aligned}
\sum_{C \in W_\theta} P(C)H(C) &= -\sum_{C \in W_\theta} P(C) \sum_{w \in C} \frac{p(w)}{P(C)} \log \frac{p(w)}{P(C)} \\
&= -\sum_{C \in W_\theta} \sum_{w \in C} p(w) \log p(w) + \sum_{C \in W_\theta} \log P(C) \sum_{w \in C} p(w) \\
&= -\sum_{w \in W} p(w) \log p(w) + \sum_{C \in W_\theta} \log P(C) \cdot P(C) \\
&= H(W) - H(W_\theta)
\end{aligned}$$

Carter's final measure was the Percentage of Information Extracted by θ:

$$PIE(\theta) = \frac{H(W_\theta)}{H(W)} 100\% \tag{2}$$

$1 - PIE(\theta) = \frac{H(W) - H(W_\theta)}{H(W)}$ looks very similar to (1); our framework includes both as special cases. It is noteworthy that Carter does not cite Hockett's work, indicating that he was not aware of it.

3 Defining a framework

We assume that a language is a sequence of discrete units, and that the units can have a complicated structure.

3.1 Describing units

A language L is a sequence $L_{\mathtt{T}}$ of *objects* of *type* `T`, or `T`-objects. For example, phonemes are objects of type `phn`. Each `T`-object x has a *value* $v(x)$, which is one of a countable set $\Phi_{\mathtt{T}}$ of possible values. For convenience, we shall often make references to types implicitly, e.g. using L for $L_{\mathtt{T}}$ and 'object' instead of '`T`-object'.

Types can be atomic or non-atomic. Non-atomic types are made using atomic types and/or other non-atomic types. If `T` is non-atomic, then a `T`-object x is made of a positive number, say n, of components $x_1, \ldots, x_n$, which are objects of type $\mathtt{T}_1, \ldots, \mathtt{T}_n$. Its value $v(x)$ is the n-tuple $\langle v(x_1), \ldots, v(x_n) \rangle$ of the values of its components, and must be one of $\Pi_{j=1}^{n} \Phi_{\mathtt{T}_j}$. The set $\Phi_{\mathtt{T}}$ of all possible values a `T`-object can take is $\cup_{n=1}^{\infty} \Pi_{j=1}^{n} \Phi_{\mathtt{T}_j}$.

Two `T`-objects x and y are *equal* iff (if and only if) they have the same value, i.e. $v(x) = v(y)$. If `T` is atomic, it is clear what this means. If `T` is non-atomic, then $v(x) = v(y)$ iff they have the same number of components and $v(x_i) = v(y_i)\ \forall i = 1, \ldots, n$.

There are several ways in which non-atomic types can be formed; we make use of only two. In the first, and usual case, the number and types of components in a `T`-object depend only on its type. (Thus we can associate components with types, rather than with objects.) `T`-objects all have the same number, $n(\mathtt{T})$, of components, and have one of the values in $\Phi_{\mathtt{T}} = \Pi_{j=1}^{n(\mathtt{T})} \Phi_{\mathtt{T}_j}$. The second case is for type `string<T>`, where the number of components can be any positive integer, but all components are of the same type, `T`.

For example, we could use the following system to represent a human language as a sequence of words. A word is an object of type `wrd`, with two components, one of type `syl` and another of type `mea`. `mea` is an atomic type representing 'meaning'[3]. A syllable is an object of non-atomic type `syl`, and has two components, of type `string<phn>` and `str`. `phn` is an atomic type representing phonemes, while `str` is an atomic type representing stress. If the language was tonal, syllables could have a third component for tone.

More examples are given in Section 4.

3.2 Describing contrasts and their absence

It is not intuitively clear how to define a contrast in a language. One reason for this is that contrasts are better described by their absence than by their presence. Suppose c is some contrast in language $L_{\mathtt{T}}$. There are several ways to define the process by which c is removed from $L_{\mathtt{T}}$; we choose one that works object by object.

Consider the set $\Phi_{\mathtt{T}}$ of possible values of T-objects. In the absence of contrast c, some of the values will become indistinguishable from other values. "Equal in the absence of c" is an equivalence relation that induces a partition, call it θ_c, on the set $\Phi_{\mathtt{T}}$ of possible values of T-objects. For example, suppose English is represented as a sequence of phonemes ($\mathtt{T} = \mathtt{phn}$, $L_{\mathtt{T}} = English$) and c is the voicing contrast. Without voicing, phonemes like [t] and [d] sound identical, as would [s] and [z], or [f] and [v], etc. This is represented by the partition $\theta_{voicing}$ whose only equivalence classes with more than one element are {p,b}, {t,d}, {k,g}, {s,z}, {f,v}, {ʃ,ʒ}, {θ,ð} and {ʧ,ʤ}.

Just as c defines θ_c, so does any partition of $\Phi_{\mathtt{T}}$ define a contrast, i.e. $c \leftrightarrow \theta_c$. We thus define a contrast in a language $L_{\mathtt{T}}$ to be any partition of $\Phi_{\mathtt{T}}$. Notationally, this means we can drop c from our notation, and just use θ to represent a contrast. θ, being a partition of $\Phi_{\mathtt{T}}$, is implicitly parametrized by T. We will find it useful to identify θ with the function $g_{\mathtt{T},\theta} : \Phi_{\mathtt{T}} \to \theta$, where $g_{\mathtt{T},\theta}(v)$ is the equivalence class of v in θ.

Let us return to the question of what happens when a contrast θ disappears from $L_{\mathtt{T}}$. A new language $L_{\mathtt{T}_\theta}$ is created, which is a sequence of $\mathtt{T}_\theta$-objects. $\mathtt{T}_\theta$ is a new type that is defined to be just like T in its component structure, but its possible values are equivalence classes in θ. Therefore:

$$\Phi_{\mathtt{T}_\theta} = \theta \tag{3}$$

As already mentioned, the function converting $L_{\mathtt{T}}$ to $L_{\mathtt{T}_\theta}$ operates object by object. In other words, every T-object x in $L_{\mathtt{T}}$ is replaced by a $\mathtt{T}_\theta$-object with value $g_{\mathtt{T},\theta}(v(x))$. Note that because of (3), $g_{\mathtt{T},\theta}$ is a function from $\Phi_{\mathtt{T}}$ to $\Phi_{\mathtt{T}_\theta}$ as well.

Examples of contrasts are given in Section 5.

3.3 The functional load of a contrast

A language $L_{\mathtt{T}}$ is a sequence of T-objects. If we assume that $L_{\mathtt{T}}$ is generated by a stationary ergodic process, which we also call $L_{\mathtt{T}}$, then its entropy $H(L_{\mathtt{T}})$ is well-defined, being the entropy of its stationary distribution. The entropy of a distribution D over a countable set is $H(D) = -\sum_i p_i \log_2 p_i$, where p_i is the probability of the i-th member of D. Note that $p_i \log_2 p_i$ is taken to be zero if $p_i = 0$.

We define the functional load of a contrast θ in $L_{\mathtt{T}}$ as

$$FL_{\mathtt{T}}(\theta) = \frac{H(L_{\mathtt{T}}) - H(L_{\mathtt{T}_\theta})}{H(L_{\mathtt{T}})} \tag{4}$$

In practice, we assume that the stationary ergodic process is a very special process, namely a $(n-1)$-order Markov process, which we denote by $L_{\mathtt{T},n}$. This means that the probability distribution on the value of a T-object depends on the preceding $n-1$

3 This paper never goes beyond phonology, so we do not ever use such a type.

T-objects. The entropy of $L_{\mathtt{T},n}$, which is the entropy of the distribution of n-grams of T-objects, is an n-th order approximation to that of $L_{\mathtt{T}}$ that improves as n becomes larger; Shannon (1951) proved that $H(L_{\mathtt{T}}) = \lim_{n\to\infty} H(L_{\mathtt{T},n})$.

We may want to bear in mind a passing comment by Hockett (1967). He suggested that finite n might actually be more appropriate for languages, as articulatory constraints prevent the formation of infinitely long utterances. Perceptual mechanisms clump phonemes into cohesive units, such as syllables or words, when presented with long utterances. In principle, clumping never stops; sequences of words get clumped into sentences, and so on. How far the assumption of generation by a stationary, ergodic Markov process can be taken is not known.

We define the n-th order approximation to the functional load of contrast θ in $L_{\mathtt{T}}$ as

$$FL_{\mathtt{T},n}(\theta) = \frac{H(L_{\mathtt{T},n}) - H(L_{\mathtt{T}_\theta}, n)}{H(L_{\mathtt{T},n})} \tag{5}$$

Note that taking $\mathtt{T} = \mathtt{phn}$ gives Hockett's formula (1) while taking $\mathtt{T} = \mathtt{wrd}$, with $n = 1$ fixed, gives Carter's formula (2).

The parameters of $L_{\mathtt{T},n}$ must be estimated using a finite sample of its outputs, i.e. a finite sequence of T-objects. This finite sequence is called a *corpus*. We denote by $\hat{H}(L_{\mathtt{T},n}; S)$ the entropy of the process $L_{\mathtt{T},n}$ when its parameters are estimated using corpus S. N, the number of T-objects in S, and the structure of $\Phi_{\mathtt{T}}$, determine how large n can be made before sparse sampling problems become an issue.

There are several ways of finding the estimate $\hat{H}(L_{\mathtt{T},n}; S)$ from S. We used the classical method of normalized counts of n-grams in S. Suppose $c(u_1 \ldots u_n)$ is the number of times $u_1 \ldots u_n$ (each $u_i \in \Phi$) appears as a contiguous subsequence of S. Define a probability distribution D_n over n-grams by $p(u_1 \ldots u_n) = \frac{c(u_1 \ldots u_n)}{N-n+1}$. Then $\hat{H}(L_{\mathtt{T},n}; S) := \frac{1}{n} H(D_n)$.

To illustrate, consider a toy language L represented by a sequence of toy-objects with $\Phi_{\mathtt{toy}} = \{$a,b,c$\}$. The corpus to be used is $S =$'abaccaaccaabbacabab'. Say $n = 2$. The distribution D_2 of toy bigrams in S is (aa 2), (ab 4), (ac 3), (ba 3), (bb 1), (bc 0), (ca 3), (cb 0), (cc 2). $H(D_2) = -\frac{2}{18}\log_2\frac{2}{18} - \frac{4}{18}\log_2\frac{4}{18} - \ldots - \frac{2}{18}\log_2\frac{2}{18} = 2.7108$. So $\hat{H}(L_{\mathtt{toy},2}; S) = \frac{1}{2}2.7108 = 1.3554$.

This means that our estimate of the n-th order approximation to the functional load of a contrast θ in $L_{\mathtt{T}}$ is

$$\hat{FL}_{\mathtt{T},n}(\theta; S) = \frac{\hat{H}(L_{\mathtt{T},n}; S) - \hat{H}(L_{\mathtt{T}_\theta,n}; g_{\mathtt{T},\theta}(S))}{\hat{H}(L_{\mathtt{T},n}; S)} \tag{6}$$

For convenience, we will often write $FL_{\mathtt{T},n,S}(\theta)$ for $\hat{FL}_{\mathtt{T},n}(\theta; S)$.

Let us return to the toy example. If we do not make use of the b/c opposition, any occurrence of b or c in the corpus $S =$'abaccaaccaabbacabab' is taken to be an occurrence of the same symbol, which we call, say, d. The corresponding partition θ_{bc} of $\Phi_{\mathtt{toy}}$ is $\{\{a\},\{b,c\}\} \simeq \{a,d\} = \Phi_{\theta_{bc}}$. The converted corpus $g_{\mathtt{toy},\theta_{bc}}(S)$ reads 'adaddaaddaaddadadad'. The distribution of $\mathtt{toy}_{\theta_{bc}}$ bigrams is (aa 2), (ad 7), (da 6), (dd 3) and the resulting entropy 1.8016. Plugging these values in (6) gives $\hat{FL}_{\mathtt{toy},2}(\theta_{bc}; S) = \frac{2.7108-1.8016}{2.7108} = 0.335$, meaning that the b/c contrast carries over a third of the information in S — when n is 2.

Clearly, there are two nuisance parameters here, n and S. In section 6, we investigate how much difference the choices of n and S makes. We find they do not make as much difference as might be feared, possibly since the entropies in the numerator and denominator 'cancel out'. However, they are still certainly an issue to keep in mind, and a few remarks on them are in order.

Most linguists, when speaking of phonological rules, usually assume $n = 1$, going to $n = 2$ for a few rules involving word boundaries. This is both because many rules don't go beyond two word boundaries and because it is convenient to do so. In other words, the approximations we make here are no worse than those usually made by linguists.

That the choice of S makes a difference is clear; the entropy of a text can even be used to distinguish between authors (Kontoyannis, 1997) writing in the same language. We suspect, without proof, that FL is more robust than entropy to changes in S, since FL normalizes entropy both additively and multiplicatively.

4 What types to use for human languages

4.1 Non-tonal languages

In the calculations for Dutch, English, and German in Section 8, we used four types, for phonemes, stress, syllables and words. The first two types are atomic. All $\Phi_{\mathtt{T}}$ differ with language; the examples given here are for English.

- Objects of type `phn`, which we call phonemes for convenience, take values in $\Phi_{\mathtt{phn}} = \{$[p],[t],[k],...,[æ],[i],[ɪ]$\}$.
- `str`-objects take values in $\Phi_{\mathtt{str}} = \{$primary, secondary, unstressed$\}$.
- `syl`-objects (syllables) have two components; $n(\mathtt{syl}) = 2$. The first is of type `string<phn>` and the second of type `str`. Two syllables with values ⟨miŋ,unstressed⟩ and ⟨miŋ,primary⟩ are not equal, since although their phonemic components are equal, their stress components are not.
- `wrd`-objects (words) have a single component, of type `string<syl>`.

4.2 Tonal languages

In the calculations for Mandarin and Cantonese in this chapter, we used the same setup as for the non-tonal languages, bar two changes. First, of course, the sets of possible values ($\Phi_{\mathtt{phn}}$, $\Phi_{\mathtt{wrd}}$, etc) differ with language. Second, syllables have an additional component for tone, of atomic type `ton`. In Mandarin, for example, the set of possible tonal values is $\Phi_{\mathtt{ton}} = \{$high level, rising, low level, falling, no tone$\}$.

Of course, allocating tones to syllables is an idealization, since tone sandhi and coarticulation occur in continuous speech. An example of the former, due to Chao (1968), is with the words 'yi', 'qi', 'ba' and 'bu' which have high, high, high and falling tone in isolation[4]. In continuous speech they all have falling tone unless they are followed by a falling tone, in which case they have a rising tone. Such cases are predictable in that they could be corrected for with corpus pre-processing. However, we did not correct for them.

Regarding coarticulation, Xu (1993) found that "Mandarin speakers identify the tones presented in the original tonal contexts with high accuracy. Without the original context, however, correct identification drops below chance for tones that deviate much from the ideal contours due to coarticulation. When the original tonal context is altered, listeners compensate for the altered contexts as if they had been there originally. These results are

4 These words are written in Pinyin. They mean 'one', 'seven', 'eight' and 'no' respectively in English.

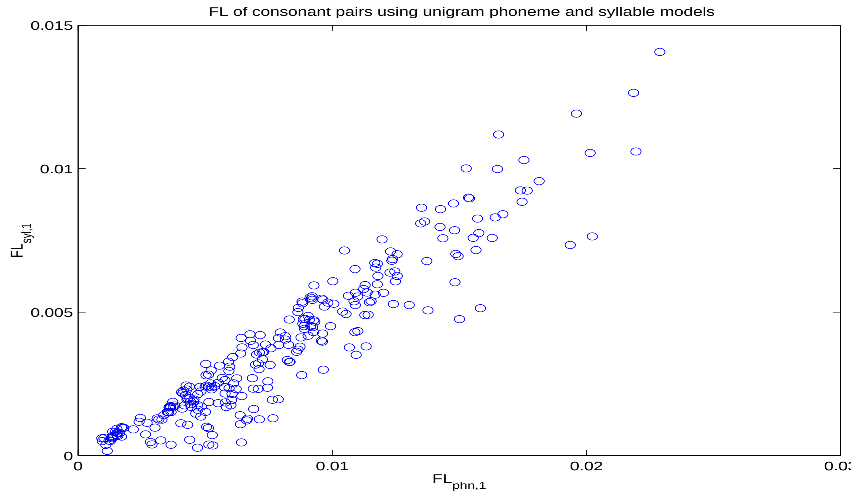

Figure 1
Comparing the FL of 276 consonant pairs using phoneme unigrams and syllable unigrams, in the Switchboard corpus. The correlation is 0.942.

interpreted as demonstrating listeners' ability to compensate for tonal coarticulation." While this justifies our idealization to a large extent, bear in mind that the compensation for coarticulation is by no means perfect, particularly where adjacent tones 'disagree' (Xu, 1993; Xu, 1994).

4.3 Extensions required

The model of phonology used in this paper is more general than classical structural phonology. However, one may well ask how we could make use of more sophisticated models such as autosegmental phonology (Goldsmith, 1976), especially since a computational framework for it already exists (Albro, 1993).

We are not sure how this can be done. However, we have some suggestions, which involve making components correspond to tiers. We need to assume that there is some overall (i.e. over all tiers) unit that no object in any tier ever straddles. For example, in a language where a tone can be associated with vowels in different words, such a unit would have to be strictly larger than a word. Even so, taking it to be a word still permits several phonological rules to be represented as contrasts (see Section 5.4). Among the details we have yet to sort out is how to represent association lines between tiers.

5 Examples of contrasts

In Section 3.2, any partition of $\Phi_{\mathtt{T}}$ defines a contrast in a language represented as a sequence of T-objects. This allows us to use the word 'contrast' in a more general sense than is standard, as the examples in this section show. These examples make use of the types defined in Section 4.

5.1 Phoneme oppositions

Nearly all previous work on FL, in both linguistics and speech recognition, has been on phoneme oppositions, especially binary.

Suppose a language is a sequence of phonemes. Almost any phoneme opposition can be represented by a partition θ of $\Phi_{\texttt{phn}}$ with the opposition being that between phonemes in the same equivalence class of θ. For example, the binary opposition of phonemes x and y is represented by θ being the partition with just one non-singleton equivalence class, $\{x, y\}$.

More generally, if the opposition is between phonemes in set $A \subseteq \Phi_{\texttt{phn}}$ then we can take θ to be the partition of $\Phi_{\texttt{phn}}$ with A as one equivalence class and all other classes with one phoneme each. $A = \{x, y\}$ is, of course, the binary opposition case of the previous paragraph. Note that the contrast here is 'distinguishing between phonemes within A', not 'distinguishing phonemes in A from phonemes not in A'. Table 5 has some examples.

Even more generally, if the opposition is between phonemes in several pairwise-disjoint sets of phonemes, take θ to be the partition defined by these sets. For example, if the opposition is between consonants and between vowels simultaneously, take θ to be the two-class partition of consonants and vowels. $FL(\theta)$ then represents the information lost when one can tell whether a consonant or vowel has occurred, though not which vowel or which consonant.

This is all very well if `T` is in fact `phn`. But what if the objects are syllables or words? In this case, we make use of inheritance across types. For example, if `T` = `syl`, since syllables have a `string<phn>` component, any partition of $\Phi_{\texttt{phn}}$ induces a partition of $\Phi_{\texttt{syl}}$. Similarly, if `T` = `wrd`, since words have a `string<syl>` component, any partition of $\Phi_{\texttt{phn}}$ induces a partition of $\Phi_{\texttt{syl}}$ which in turn induces one of $\Phi_{\texttt{wrd}}$. Thus partitions of $\Phi_{\texttt{phn}}$ are contrasts whether the objects are phonemes, syllables or words.

This is better explained if we use $g_{\texttt{T},\theta}$ instead of θ. Recall from Section 3 that $g_{\texttt{T},\theta}$ is the function converting the original language $L_{\texttt{T}}$ to the contrast-less language $L_{\texttt{T}_\theta}$ by sending all `T`-objects with values in the same equivalence class of θ to a $\texttt{T}_\theta$-object with the same value. For example, suppose, once again, that the contrast is between phonemes in some set A and that `T` = `phn`. For any phoneme $p \in \Phi_{\texttt{phn}}$,

$$g_{\texttt{phn},\theta}(p) = \begin{cases} A & \text{if} \quad p \in A \\ p & \text{if} \quad p \notin A \end{cases} \tag{7}$$

For convenience, we abuse notation by mapping p to itself, rather than to $\{p\}$, if $p \notin A$.

Now, suppose `T` = `syl` and θ is the same partition of $\Phi_{\texttt{phn}}$. Syllables have a component of type `string<phn>`; suppose for concreteness' sake they have only one other component, of type `str`. Thus, a typical syllable is an ordered 2-tuple $\langle p_1 \dots p_m, s \rangle$, where each $p_i \in \Phi_{\texttt{phn}}$ and $s \in \Phi_{\texttt{str}}$. Now we have

$$g_{\texttt{syl},\theta}(\langle p_1 \dots p_m, s \rangle) = \langle g_{\texttt{phn},\theta}(p_1) \dots g_{\texttt{phn},\theta}(p_m), s \rangle \tag{8}$$

Notice that until now, θ had to be a partition of $\Phi_{\texttt{T}}$. However, now `T` = `syl`, but θ is a partition of $\Phi_{\texttt{phn}}$. This is not a contradiction, but merely systematic abuse of notation, since any partition of $\Phi_{\texttt{phn}}$ naturally induces a partition of $\Phi_{\texttt{syl}}$.

If θ' is another partition of $\Phi_{\texttt{str}}$, represented by a function $h_{\texttt{str},\theta'}$, then θ and θ' applied simultaneously result in a contrast represented by a function taking $\langle p_1 \dots p_m, s \rangle$ to $\langle g_{\texttt{phn},\theta}(p_1) \dots g_{\texttt{phn},\theta}(p_m), h_{\texttt{str},\theta'}(s) \rangle$.

5.2 Distinctive Features

By distinctive feature, we refer to characteristics used to distinguish phonemes, such as aspiration, voicing, place, manner, etc. Distinctive features do not have to be binary.

Any distinctive feature can be represented by a partition θ of $\Phi_{\texttt{phn}}$ which has two or more phonemes in the same class iff they would be merged in the absence of the feature. For example, if voicing were lost in English, θ is $\theta_{voicing}$ in Section 3.2, where [t] and

[d] are in one equivalence class, [s] and [z] in another, [ʃ] and [ʒ] in another, etc, with all other phonemes in their own classes.

Most well-studied languages have several possible organizations of its phonemes and distinctive features[5] Any organization can be used, as long as one is specified. What we mean by organization is best explained by example; we used the organizations in Tables 2 and 3 for Mandarin, Dutch, English and German to get the FL of different features in each language in Table 4.

5.3 Suprasegmental contrasts

Suppose we model a language by a sequence of syllables, with each syllable having a stress component. Since any partition of $\Phi_{\texttt{str}}$ induces one of $\Phi_{\texttt{syl}}$ by inheritance, any partition θ of $\Phi_{\texttt{str}}$ is a contrast. This remains the case if we model a language by a sequence of words where words have a `string<syl>` component, since any partition of $\Phi_{\texttt{syl}}$ induces one of $\Phi_{\texttt{wrd}}$.

To find the FL of stress, use the partition of $\Phi_{\texttt{str}}$ with a single class containing all stress values. This is equivalent to not having any information about stress at all.

Suppose we were dealing with a language like English with different kinds of stress, and we wanted to find out how importance it was to be able to distinguish primary from secondary stress. Then we would use the partition {{primary,secondary},{absent}} of $\Phi_{\texttt{str}}$. If we wanted to find out how importance it was to distinguish secondary stress from no stress at all, we would use {{primary},{secondary,absent}} instead.

If we were modelling a tonal language, with syllables having a tonal component, then everything above said for stress would apply to tone, with tonal contrasts represented by partitions of $\Phi_{\texttt{ton}}$. For instance, to find the FL of tone, use the 1-class partition of $\Phi_{\texttt{ton}}$.

5.4 Phonological rules

In all the previously described contrasts, the conversion from `T`-object to `T`-object was absolute, i.e. it happened in every situation where it could happen. Sometimes, we would like the conversion to occur only in certain situations.

For example, if we wanted to find the functional load of vowels when `T` = `syl`, we would take θ to be the partition of $\Phi_{\texttt{phn}}$ whose only non-singleton equivalence class was V, the set of vowels. Defining $g_{\texttt{phn},\theta}$ as in (7), we would write, as in (8)

$$g_{\texttt{syl},\theta}(\langle p_1 \dots p_m, s\rangle) = \langle g_{\texttt{phn},\theta}(p_1) \dots g_{\texttt{phn},\theta}(p_m), s\rangle$$

Now, suppose we wanted to represent the contrast of vowel reduction, i.e. of not being able to distinguish between vowels in unstressed syllables. This means that every vowel is replaced by a single vowel placeholder, but only if the syllable containing it is unstressed. In other words, the mapping is now:

$$g_{\texttt{syl},\theta}(\langle p_1 \dots p_m, s\rangle) = \begin{cases} \langle g_{\texttt{phn},\theta}(p_1) \dots g_{\texttt{phn},\theta}(p_m), s\rangle & \text{if} \quad s \text{ is unstressed} \\ \langle p_1 \dots p_m, s\rangle & \text{if} \quad \text{not} \end{cases}$$

where

$$g_{\texttt{phn},\theta}(p) = \begin{cases} V & \text{if} \quad p \text{ is a vowel} \\ p & \text{if} \quad p \text{ is not a vowel} \end{cases} \tag{9}$$

5 The number of organizations is a monotonically increasing function of the number of studies of the language. The nature of this function requires, though not necessarily deserves, further study.

Some phonological rules in linguistics fit in this framework very nicely. For example, epenthesis of [t] in the consonant cluster [n_s] in English is represented by the function

$$g_{\texttt{syl},\theta}(\langle p_1 \dots p_m, s\rangle) = \begin{cases} \langle p_1 \dots p_i[\mathrm{t}]p_{i+1} \dots p_m, s\rangle & \text{if} \quad p_i = [\mathrm{n}] \;\&\; p_{i+1} = [\mathrm{s}] \\ \langle p_1 \dots p_m, s\rangle & \text{if} \quad \text{not} \end{cases}$$

In this case, θ corresponds to the partition of $\Phi_{\texttt{syl}}$ where two syllables are in the same equivalence class iff $g_{\texttt{syl},\theta}$ maps them to the same value. Thus the syllables [kænts] and [kæns] end up in one class, [lɪns] and [lɪnts] in another, and so on. If $\mathtt{T} = \mathtt{wrd}$ instead, then words like 'tense' and 'tents' would end up in the same class, 'mince' and 'mints' in another, and so on.

5.5 The contrast of a single phoneme

At first, it makes little sense to speak of the functional load of a single phoneme. After all, phonemic oppositions require at least two phonemes to be in opposition.

A clue to how to proceed is given by Ingram (1989), who states that the FL of [ð] in English must be low because "we could change all English /dh/ into [d]'s and still communicate". He was referring to the fact that /dh/, which is the most frequent consonant in English, does not intuitively seem to be most relied-upon consonant.

More generally, the question to be asked is 'how can a phoneme disappear from a language?' Some phonemes, like [h] in Cockney English, disappear. Others vanish by merging with other phonemes, e.g. [n] with [l] in Cantonese. The merger need not be absolute, i.e. with the same phoneme everywhere, of course.

We define the contrast of a single phoneme to be the phonological rule by which the phoneme disappears from the language. Therefore $FL(x)$ is the FL of the phonological rule for the disappearance of phoneme x.

Unfortunately, the process by which a phoneme disappears can rarely be predicted before it, if it ever does, disappears. What is needed is a comprehensive survey of how a given phoneme has disappeared from various languages in the past. Such a survey would be able to answer hypotheses like 'does /h/ ever disappear by a process other than deletion?', or 'do phonemes only merge with phonemes that share the same place (phonemes with secondary articulations being considered as having two places of articulation)?'

Our current working definition for $FL(x)$, in the case of disappearance-by-merger, is as follows. Suppose x can only potentially merge with phonemes in a set $S(x)$ of phonemes 'similar' to it, and that the probability that it merges with phoneme $y \in S(x)$ is $P(x, y)$. Then

$$FL(x) = \sum_{y \in S(x) - x} P(x, y) FL(x, y)$$

This can be interpreted as the expected FL of x, taken over possible absolute mergers. Alternatively, it can be interpreted as the FL of the process where x merges with phonemes in $S(x)$, merging with different phonemes in different environments such that $P(x, y)$ is the proportion of environments where x merges with y.

6 The robustness of the measure

6.1 Measuring robustness

We would like to speak of $FL_{\mathtt{T}}(\theta)$ without reference to the parameters n and S. This cannot be done if we expect any two possible measures to give the same *absolute* value for any contrast. For example, for most contrasts θ, $\hat{FL}_{\mathtt{T},n}\ (\theta; S)$ will be larger than

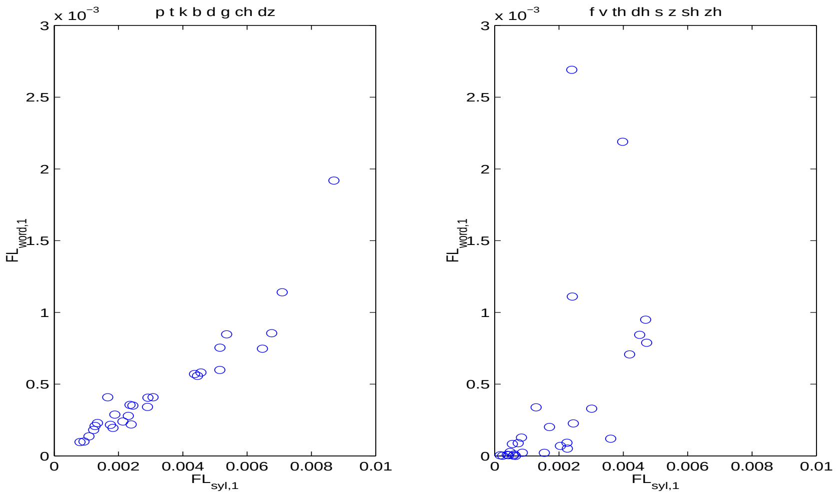

Figure 2
Comparing Functional Load values for 28 pairs of obstruent consonants using unigram syllable (horizontal axis) and word based computations. Both are based on the CELEX lexicon. The left plot is for pairs from {p,t,k,b,d,g,ʧ,ʤ}; the correlation is 0.927. The right plot is for pairs from {f,v,θ,ð,s,z,ʃ,ʒ}; the correlation is 0.611. Both plots are to the same scale; the horizontal axis is from 0 to 0.010 while the vertical is from 0 to 0.003.

$\hat{FL}_{\mathrm{T},n+1}(\theta; S)$ because larger n-grams capture more information. Instead we wish them to give the same 'relative' values, to be highly predictable from each other.

We define measures FL_1 and FL_2 to be *consistent* for a set $\mathbf{\Theta}$ of contrasts iff there is a constant γ_{12} such that $FL_1(\theta) = \gamma_{12} FL_2(\theta)\ \forall \theta \in \mathbf{\Theta}$.

In practice, we can only hope for $FL_1(\theta) \approx \gamma_{12} FL_2(\theta)$. Bearing in mind that what is important is not the value of γ_{12} but its existence, we define $\alpha_{\mathbf{\Theta}}(FL_1, FL_2)$ to be the linear (Pearson's) correlation between values $FL_1(\theta)$ and $FL_2(\theta)$, when θ is taken over all values in $\mathbf{\Theta}$. In other words,

$$\alpha_{\mathbf{\Theta}}(FL_1, FL_2) = \frac{1}{|\mathbf{\Theta}|} \sum_{\theta \in \mathbf{\Theta}} Z(FL_1(\theta)) Z(FL_2(\theta))$$

Note that $Z(FL_i(\theta)) = \frac{FL_i(\theta) - \mu_i}{\phi_i}$, where $\mu_i = \frac{1}{|\mathbf{\Theta}|} \sum_{\theta \in \mathbf{\Theta}} FL_i(\theta)$ and $\phi_i^2 = \frac{1}{|\mathbf{\Theta}|-1} \sum_{\theta \in \mathbf{\Theta}} (FL_i(\theta) - \mu_i)^2$.

The maximum, ideal, value of $\alpha_{\mathbf{\Theta}}$ is 1. We do not know how high it must be for FL_1 and FL_2 to be consistent in general, though we have rules of thumb for specific cases.

This section gives evidence for the consistency of $\hat{FL}_{\mathrm{T},n}(\theta; S)$ and $\hat{FL}_{\mathrm{T},n'}(\theta; S')$ for different $n, n' > 0$ and corpora S, S'. This restriction in the interpretation of FL still allows it to be useful, as described in Section 12.

6.2 Testing Procedure

Unless otherwise specified, we will restrict ourselves to a limited collection of contrasts, namely binary oppositions. These are very fine contrasts (i.e. the partition of $\Phi_{\mathtt{phn}}$ they rely on is almost the finest possible) and consistency for them is indicative of consistency for other contrasts.

Suppose that $\Phi_0 \subset \Phi_{\mathtt{phn}}$ is a subset of phonemes, and $\mathbf{\Theta}_{\Phi_0}$ is the set of all contrasts that are binary oppositions of pairs of phonemes in Φ_0. For example, if $\Phi_0 = \{w, x, y, z\}$, then $\mathbf{\Theta}_{\mathbf{\Phi_0}}$ is $\{\theta_{wx}, \theta_{wy}, \theta_{wz}, \theta_{xy}, \theta_{xz}, \theta_{yz}\}$. For convenience we define $\alpha_{\Phi_0}(FL_1, FL_2)$ to be $\alpha_{\mathbf{\Theta}_{\mathbf{\Phi_0}}}(FL_1, FL_2)$. All correlations reported here are extremely significant, having $p < 10^{-5}$ unless reported otherwise. Our rule of thumb is that FL_1 and FL_2 are consistent over Φ_0 if $\alpha_{\Phi_0}(FL_1, FL_2) > 0.9$.

While our testing was only done with English corpora, results should hold for other languages. The corpora used were CELEX (Baayen, Piepenbrock and Gulikers, 1995) and Switchboard (Godfrey, Holliman and McDaniel, 1992). CELEX is essentially a word-frequency list with each word having a citation form pronunciation and the frequency with which it appears in the 16 million word (24 million syllables) Birmingham/COBUILD corpus of British English. Switchboard (SWB) is a large 240-hour speech corpus, but we used the small but extra-carefully transcribed ISIP subset of it[6], which has 80 000 phonemes in 36 500 syllables in 2 hours of spontaneous telephone speech by American English speakers.

6.3 Consistency for different n

For any fixed $\mathtt{T}$, corpus S, $\Phi_0 \subseteq \Phi$, we want $\alpha_{\Phi_0}(FL_{\mathtt{T},m,S}, FL_{\mathtt{T},n,S})$ to be as close to 1 as possible for any positive integers m, n. Table 1 shows its value when $\mathtt{T} = \mathtt{phn}$, S is Switchboard, Φ_0 consists of all consonants (values for vowels are higher) and $1 \leq m, n \leq 5$. The correlation decreases with $|m - n|$ but remains high throughout.

	1	2	3	4
2	0.985			
3	0.956	0.988		
4	0.928	0.961	0.988	
5	0.878	0.906	0.947	0.978

Table 1
The correlation $\alpha_{consonants}$ between $FL_{phn,n,Switchboard}$ for different n.

Similar results are found when $\mathtt{T} = \mathtt{syl}$; $\alpha_{consonants}(FL_{\mathtt{syl},1,SWB}, FL_{\mathtt{syl},2,SWB}) = 0.945$. However, sparsity concerns about the small size of Switchboard made values of $FL_{\mathtt{syl},n,SWB}$ for $n > 2$ suspect and larger values of n were not tried.

For $\mathtt{T} = \mathtt{wrd}$, we used frequency and sequence information from the Brown corpus and pronunciation information from CELEX. We then computed $FL_{\mathtt{wrd},n,Brown-CELEX}$ values for 200 randomly generated partitions of $\Phi_{\mathtt{phn}}$, for $n = 1, 2, 3$, and found that the correlation was over 0.95 in each case.

We conclude from this that taking $n = 1$, i.e. estimating FL with unigrams, is adequate for many purposes. In the rest of this paper, n is 1 if not specified.

6.4 Consistency for different corpora

For any fixed type $\mathtt{T}$, $n > 0$, and $\Phi_0 \subseteq \Phi_{\mathtt{phn}}$, we want $\alpha_{\Phi_0}(FL_{\mathtt{T},n,S}, FL_{\mathtt{T},n,S'})$ to be as close to 1 as possible for different corpora S, S'. Taking advantage of the results of Section 6.3, we assume $n = 1$.

We deal with `syl` objects. The corpora in question are Switchboard and CELEX. Note that stress information was removed from CELEX for this comparison, since Switchboard syllables do not have stress information[7]. $\alpha_{consonants}$ $(FL_{\mathtt{syl},SWB}$, $FL_{\mathtt{syl},CELEX})$

6 Our thanks to the researchers at Mississippi State who have made this subset freely available at www.isip.msstate.edu/projects/switchboard
7 Gina Levow informed us that syllables are marked with stress in another subset of Switchboard. However, this was after the calculations in this paper were done.

$= 0.826$ while α_{vowels} ($FL_{\texttt{syl},SWB}$, $FL_{\texttt{syl},CELEX}$) $= 0.730$. Interestingly, some consonants fare better than others: $\alpha_{obstruents}$ ($FL_{\texttt{syl},SWB}$, $FL_{\texttt{syl},CELEX}$) $= 0.920$ while $\alpha_{non-obstr.\ cons.ts}$ ($FL_{\texttt{syl},SWB}$, $FL_{\texttt{syl},CELEX}$) is 0.762. More details of this experiment are in Section 7.

Although entropy is known to be very corpus dependent, it appears that the normalized differences in entropy are more well-behaved. This is certainly the case when obstruents are involved, in which case FL calculations are robust. Other contrasts require further work, though the computation of their FL is robust enough for many purposes.

6.5 Consistency for different objects

Object type is a necessary parameter when computing FL. Intuitively, we expect some consistency for different types, but not in the same way as for n and S, and therefore inconsistency across different types indicates interesting word structure patterns. In other words, comparisons of $FL_{\texttt{T,n,S}}$ and $FL_{\texttt{T}',\texttt{n,S}}$, for different `T` and `T'`, could prove to be a useful tool for linguistic analysis.

We compare `phn` and `syl`, for $n = 1$ and $S =$SWB. In this case, $\alpha_{consonants}$ ($FL_{\texttt{phn}}$, $FL_{\texttt{syl}}$) $= 0.942$, which is very high. The corresponding values for α_{vowels} is even higher. The surprise here is that $FL_{\texttt{phn}}$ is based on phoneme unigrams, i.e. how many times each phoneme appears, and thus makes no use of context.

We compare `syl` and `wrd` with $n = 1$ and $S =$CELEX. Here, context turns out to be more important; α_{vowels} ($FL_{\texttt{syl}}$, $FL_{\texttt{wrd}}$) is 0.752 and $\alpha_{obstruents}$ ($FL_{\texttt{syl}}$, $FL_{\texttt{wrd}}$) is 0.716. Interestingly, the latter figure really has two parts (see Figure 2) since $\alpha_{stops+affricates}$ is 0.927 while $\alpha_{fricatives}$ is 0.611 ($p = 0.001$). We do not know why this is so, nor why the latter figure (again) has two parts, with α higher for voiced fricatives than unvoiced.

7 Computing FL with non-ideal data

Robust FL computation means we can find usable FL values for languages for which inadequate data is available. For example, there are relatively few corpora that are manual phonetic transcriptions of 'the language as spoken'; this is particularly true for languages for which there are few or no native speakers. On the other hand, word-frequency pairs, with citation form pronunciations of words and frequencies based on *written texts*, are easier to find. To see if we can accurately estimate FL using word-frequency pairs, we look at the CELEX vs Switchboard calculations of Section 6 in more detail. These corpora represent opposite ends of several spectrums, which makes for a good test. The differences between them are summarized here:

- Switchboard and CELEX reflect different dialects, American and British respectively, of English.
- The frequencies in CELEX are mostly based on written sources.
- As CELEX gives word-frequency lists, all syllabifications in it are word-internal or at word boundaries, unlike Switchboard.
- CELEX reflects a much (>600 times) larger corpus than Switchboard.
- CELEX gives citation form pronunciations for each word. 30% of words also have other pronunciations, but there is (unsurprisingly) little information on how often each other pronunciation is used. The word-frequency list we extracted from CELEX assigned a single pronunciation to a word. This was the citation form except when other pronunciations were available, in which case we took the most common colloquial form.

- Each syllable in CELEX is marked as having one of three types of stress: primary, secondary and none. The syllable in monosyllabic words has primary stress. Syllables in our Switchboard data are not marked with stress. To make syllables comparable, the stress component was removed from the CELEX syllables.

At first sight, it would seem that we should compare $FL_{\texttt{wrd},CELEX}$ with $FL_{\texttt{wrd},SWB}$. But this requires making the sorts of assumptions (syllables don't cross word boundaries, same pronunciation each time) about Switchboard as for CELEX, the very assumptions we wish to test. To get an idea of what words look like in continuous speech, consider the ARPABET-transcribed SWB sentence below. Syllables are within square brackets and interphoneme silences have been removed.

```
    [l ay] [k ih n] [ao] [g ix] [s w eh] [n eh r] [iy] [b aa] [d iy]
[z aa n] [v ey] [k ey] [sh ih] [n er] [s ah m] [th ih ng k] [w iy]
[k ix n] [d r eh] [s el] [l el] [m ao r] [k ae] [zh w ax l]
```

The actual sentence is *"Like in August when everybody is on vacation or something we can dress a little more casual"*. Notice how often syllables cross word boundaries.

Even if we weaken the restriction so that words are pronounced in a limited set of ways, it is hard to draw the line on what 'limited' means. Therefore, we shall instead compare $FL_{\texttt{syl},CELEX}$ with $FL_{\texttt{syl},SWB}$. Then $\alpha_{\Phi_0}(FL_{\texttt{syl},SWB}, FL_{\texttt{syl},CELEX})$ is 0.730, 0.826 and 0.920 for vowels, consonants, and obstruents respectively.

We conclude that non-ideal corpora can give results consistent with ideal corpora that are very representative of speech for contrasts that involve consonants, particularly obstruent consonants.

8 An application in linguistic typology

	Labial	Alveolar	Alv-pal	Retroflex	Lateral	Velar
Stop	p {pʰ} [m]	t {tʰ} [n]				k {kʰ} [ŋ]
Affricate		ts {tsʰ}	tɕ {tɕʰ}	tʂ {tʂʰ}		
Fricative	f	s	ɕ	ʂ (ɹ)		x
Approximant					l	

Table 2
Feature values of consonants in Mandarin. Columns have different Place classes and rows different Manner classes. Aspirated consonants are in braces {}, voiced in parentheses () and nasalized in square brackets []. Note that ɹ is a voiced fricative in Mandarin, not an approximant. w and j are absent as they were treated as vowels.

When comparing different languages, one often finds claims such as "language X makes more use of such-and-such-a-contrast than language Y". Quantifying FL allows one to answer several questions harder than 'Does Xhosa make more use of clicks than French?' The most detailed questions, of course, require computations to be even more robust than they are at the moment.

This section has computations of FL for Dutch, English and German from CELEX (Baayen, Piepenbrock and Gulikers, 1995) and for Mandarin based on the TDT3 Multi-language Text Version 2.0 corpus of transcriptions of Voice of America Mandarin broadcasts. In all cases calculations were based on word-frequency pairs, with citation form pronunciations for the former and frequencies from mostly written corpora. The Mandarin word for VOA was excluded from the word-frequency pairs.

Each syllable in the three European languages has a stress component. $\Phi_{\texttt{str}} = $ {primary, secondary, unstressed} for English and {present, absent} for German and Dutch.

Syllable stress information was not available for Mandarin in our corpus, though of course tonal information was. Therefore Mandarin syllables had just two components, of type `string<phn>` and `ton`.

Some of our calculations will involve distinctive features for consonants. We use the distinctive features Place, Manner, Nasality, Voicing (for Dutch, English and German) and Aspiration (for Mandarin). All but the first two are binary features. We arrange the features in a hierarchical scheme that is a much simplified version of that proposed by Ladefoged (1997). Features do not have to be specified for each phoneme, e.g. Nasality is only specified for stops. Table 2 shows our arrangement of Mandarin features while Table 3 shows that for English, Dutch and German. Note the following in the latter :

	Labial	Den	Alveolar	P-A	Lat	Pal	Velar	Uvu	Glo
Approx.	*v*		r		l	j	w		
Fricative	f (v)	θ (ð)	s (z)	ʃ (ʒ)		ç	x (ɣ)	(ʁ)	h
Affricate	pf		ts	ʧ(ʤ)					
Stop	p (b) [m]		t (d) [n]				k (g) [ŋ]		

Table 3
Feature values of consonants in Dutch, German and English that are used in CELEX. Columns have different Place classes and rows different Manner classes. P-A stands for Post-Alveolar, Den for Dental, Lat for Laterals, Pal for Palatals, Uvu for uvular and Glo for Glottal.

- The exact place of several phonemes is dialect dependent, e.g. [r] and [x] in Dutch.
- The dentals [θ] and [ð] are present in English only.
- The rhotic [r] is in English and Dutch only, [ʁ] in German only.
- Dutch does not have the velar approximant [w], but instead the labial one [*v*].
- Only Dutch has phoneme [x].
- Only some borrowed words in Dutch have [g].
- The palatal [ç] occurs in only German and some borrowed English words.
- The affricates [pf] and [ts] are only found in German.
- The affricate [ʧ] is not found in Dutch.
- CELEX does not code for a voiceless uvular fricative in Dutch or German, though the IPA does (IPA Handbook, 1999).

Table 4 has FL values for the features defined above, while Table 5 has FL values for several sets of phonemes. The following conclusions can be drawn :

- Tones in Mandarin carry far more information than Stress in the non-tonal languages. When word information is added, the FL of Stress in the latter drops to almost nothing, while that for Mandarin remains very high, having a far larger FL than Manner or Place. In fact, as shown in Table 5, the FL of tone in Mandarin is comparable to that of vowels (see Surendran and Levow (2003) for more details). This emphasizes the lexical role Tone plays in Mandarin, a role clearly not played by Stress in the non-tonal languages.

Feature	Partition (non-singleton classes)	Syllables	Words
Aspiration			
Mandarin	pʰp tʰt tsʰ.ts tɕʰ.tɕ tʂʰ.tʂ kʰk	16.7	2.7
Voicing			
Dutch	pb fv td sz ʃʒ kg xɣ	30.2	3.1
English	pb fv θð td sz ʧʤ ʃʒ kg	23.3	4.5
German	pb fv td sz ʧʤ ʃʒ kg	20.7	1.1
Place			
Dutch	wlj fsʃhx ɣvzʒ ptk bdg mnŋ	67.1	11.4
English	rljw fθsʃçh vðzʒ ptk bdg mnŋ	72.5	20.1
German	ljw fsʃçh vzʒ ptk bdg mnŋ ʧ.pf.ts	60.5	12.6
Mandarin	ptk pʰtʰkʰ mnŋ ts.tɕ.tʂ tsʰ.tɕʰ.tʂʰ fsɕxʂ	65.0	14.2
Manner			
Dutch	wfp bv st dz sh ʒʤ xk gɣ	27.1	4.5
English	fp bv rst dz ʃʧʒʤ wk jç	39.2	11.4
German	fp.pf bv st.ts dz ʃʧ ʒʤ wk jç	27.4	8.0
Mandarin	fp t.ts.s tɕ.ɕ tʂ.ʂ kx	33.7	6.4
Nasality			
Dutch	bm dn gŋ	15.2	1.5
English	bm dn gŋ	11.6	3.3
German	bm dn gŋ	15.5	1.8
Mandarin	pm tn kŋ	8.0	3.1
Tone			
Mandarin	High.Rising.Low.Falling.Absent	107.5	21.3
Stress			
Dutch	Present.Absent	25.7	0.7
English	Primary.Secondary.Absent	26.9	0.1
German	Present.Absent	34.2	0.2

Table 4
Functional Load of several distinctive features in four languages. The second column describes the non-singleton classes in the partition used to obtain the FL value for a particular distinctive feature in a language. All values should be multiplied by 0.001. Phonemes represented by more than one character are separated from others using a period, e.g. the first Manner class for German has three phonemes : [p], [f] and [pf].

- Consonants have a higher FL than vowels.
- With respect to the way we have organized distinctive features, Place has a higher FL than Manner. However, consider also the more specific case of alveolars and fricatives. The former have a very high FL in English (as noticed in Pisoni et al (1985)), Dutch and German, over twice as high as that of fricatives despite the similar number of phonemes in the two sets. But distinguishing between alveolars involves working out Manner while distinguishing between fricatives involves Place.
- $FL_{\mathtt{wrd}}$ is always lower than $FL_{\mathtt{syl}}$. This is to be expected, since knowledge of words and word boundaries is additional information available to the listener that can be used to make up for deficiencies elsewhere.
- All four languages place comparable amounts of FL on Place, Manner and Nasality. Whether there is anything universal about this remains to be seen. There certainly does not appear to be any universal along the lines of stops having a higher/lower FL than fricatives. On a side note, the latter values may be useful tools when studying lenition in historical linguistics.
- Mandarin makes far more use of affricate oppositions than German or English.

9 An application in historical linguistics

Suppose we wish to investigate Martinet's hypothesis (Martinet, 1955) that FL plays some role in phoneme mergers. To do this properly, several examples of mergers are necessary, with appropriate corpora for each case. This is hard to get. However, we do have one example that we can use to illustrate the method of investigation.

As described by Zee (1999), [n] has merged with [l] in Cantonese in word-initial position in the last fifty years. We used a word-frequency list derived from CANCORP (Lee et al, 1996), a corpus of Cantonese child-adult speech which has conveniently coded [n] and [l] as they would have occurred before the merger. Merging only in word-initial position, we computed $FL_{\mathtt{wrd}}$(n,l), which is a completely meaningless value by itself. We therefore also computed $FL_{\mathtt{wrd}}$(x,y) for all consonants in Cantonese, and found that $FL_{\mathtt{wrd}}$(n,l) was larger than over 70% of them. That tells us that the [n]-[l] contrast did have a high FL before the merger.

Table 6 shows $FL_{\mathtt{wrd}}(\mathrm{n}, x)$ for all word-initial consonants x. The results are clear, and rather startling. Of all the consonants [n] could have merged with, it merged with the second 'worst' (in an optimal sense) choice! This result adds weight to those of King (1967), the only previous corpora-based test of Martinet's hypothesis.

10 An application in child language acquisition

As mentioned early in the paper, there has been a need in this field for a comprehensive FL measure for some time. A major question is what factors affect the age at which children acquire sounds in the language. This has been investigated recently by Stokes and Surendran (2003) for consonants in three languages.

The frequency of a sound is not a consistent (across languages) predictor of when a child start to use it. For example, they find that frequency correlates very significantly with age of acquisition in Cantonese children, but the corresponding correlation for English is not significant at all. In fact, the most common consonant in English speech is /ð/, which is among the last children acquire.

Phoneme set	Partition	Syllables	Words
Vowels			
Dutch		125.5	51.5
English		133.0	48.5
German		161.3	42.2
Mandarin		91.0	22.1
Consonants			
Dutch		335.8	192.5
English		309.8	176.4
German		335.6	153.8
Mandarin		234.7	80.5
Labials			
Dutch	pbmfvw	36.5	8.7
English	pbmfv	25.2	5.9
German	pbmfv.pf	23.0	3.6
Mandarin	p^{h}pfm	10.0	1.8
Alveolars			
Dutch	tdsznlr	101.5	37.5
English	tdsznrl	98.2	41.5
German	tdsznl.ts	89.3	22.7
Mandarin	t^{h}t.ts^{h}.ts.sn	24.7	7.5
Velars			
Dutch	kgŋxɣ	20.6	0.8
English	kgŋw	6.7	1.3
German	kgŋw	5.5	0.1
Mandarin	k^{h}kxŋ	8.8	1.4
Nasals			
Dutch	mnŋ	12.0	2.0
English	mnŋ	11.5	2.8
German	mnŋ	14.4	4.4
Mandarin	mnŋ	16.2	3.1
Fricatives			
Dutch	fvrszʃʒxh	39.1	7.8
English	fvθðszʃʒçh	39.6	17.8
German	fvrszʃʒçh	53.2	14.1
Mandarin	fsɕɹxʂ	20.7	5.1
Affricates			
English	ʧʤ	0.8	0.1
German	ʧʤ.pf.ts	0.7	0.0
Mandarin	ts^{h}.ts.tɕ.tɕ^{h}.tʂ.tʂ^{h}	25.1	5.1
Stops			
Dutch	ptkbdg	56.3	10.8
English	ptkbdg	43.3	10.6
German	ptkbdg	50.1	4.5
Mandarin	p^{h}t^{h}k^{h}ptk	29.3	6.2

Table 5
The FL of several sets of phonemes in four languages. The second column describes the non-singleton classes in the partition corresponding to each set and language. All values should be multiplied by 0.001.

x	l	$\mathrm{p^h}$	$\mathrm{t^h}$	$\mathrm{k^h}$	p	t	k	w	ts
$FL_{\mathtt{wrd}}(\mathrm{n}, x)$	9.0	2.8	0.7	3.4	0.1	1.4	7.0	0.4	0.3
x	$\mathrm{ts^h}$	m	h	f	s	ŋ	$\mathrm{k^{hw}}$	$\mathrm{k^w}$	j
$FL_{\mathtt{wrd}}(\mathrm{n}, x)$	4.8	9.1	2.5	2.3	2.2	1.1	0	0.0	3.7

Table 6
Functional load values of the opposition of [n] with other consonants in Cantonese before it merged with [l] in word-initial position. Values computed with the CANCORP corpus, $n = 1$ and $\mathtt{T} = \mathtt{wrd}$. Values should be multiplied by 10^{-4}.

On the other hand, the frequency of a phoneme is not the only measure of its importance to the language. One can estimate the FL of a phoneme as well, as described in Section 5.5. Recall that $FL(x) = \sum_{y \in S(x)-x} P(x,y)FL(x,y)$, where $S(x)$ is the set of 'similar' phonemes to x, and $P(x,y)$ is the probability that x merges with y.

Stokes and Surendran (2003) find that when x is a consonant, if $S(x)$ is taken to be the set of consonants with the same place and laryngeal setting, and $P(x,y)$ is proportional to the frequency of y, then the FL of a phoneme is significantly correlated ($p < 0.05$) to age of acquisition in the three languages they check, namely Cantonese, English and Mandarin. This makes a lot of sense if children find if easier to get place and laryngeal setting (voicing, aspiration) right than manner. Note that age of acquisition refers to initial appearance of a sound in the child's phonetic inventory, not how the child uses it in its phonemic system after that.

11 Applications in automatic speech recognition

FL has, of course, already been used in the ASR community by Carter (1987); the work of Shipman and Zue (1982), Huttenlocher (1985) and Kassel (1990) should also be mentioned.

That syllables in English can be represented as a sequence of phonemes plus a stress component, the cost of whose removal can be computed, is nothing new. Extending this to tonal languages in the natural way is a simple step, but it has not been, to our knowledge, been taken before, and has already produced (see Surendran and Levow (2003)) the important result that an ASR system for Mandarin that does not try to identify the underlying tone of a syllable can only work as well as one that does identify tone but does not identify vowels! Rephrasing PIE as FL might sound superficial; but even if rephrasing does not result in additional answering power, it can result in additional question-asking power.

In any case, our FL framework is an extension rather than a simple rephrasing. For example, detailed analyses of a phonetically-based ASR system can throw up problems that it would be useful to know the importance of — if they are not important, they can be ignored. Suppose an ASR system often errs in deciding whether there is or is not a [j] before a high vowel. A decision is taken to always ignore the presence of such a [j] (or alternatively, to impose its presence even when absent) — how much information will be lost by doing so? By finding the FL of such a contrast, which is represented by the rule below, researchers can make a better informed decision.

$$g_{\mathtt{syl},\theta}(\langle p_1 \dots p_m, s\rangle) = \begin{cases} \langle p_1 \dots p_{i-1} p_{i+1} \dots p_m, s\rangle & \text{if} \quad p_i = [\mathrm{y}] \ \& \ p_{i+1} \in \{\text{high vowels}\} \\ \langle p_1 \dots p_m, s\rangle & \text{if} \quad \text{not} \end{cases}$$

12 Interpreting FL values

A serious-looking limitation of FL values is that they are relative rather than absolute. However, this still allows them to be used in several applications. One example is correlation analysis, since $corr(X,Y) = corr(aX,Y)$ and $corr(\log(X), \log(Y)) = corr(\log(aX), \log(Y))$ for any $a > 0$. So if we want to see if there is any correlation between FL, or log FL, and some other parameter, we can do so with relative FL values.

Another way to interpret FL values is comparing them with other FL values computed the same way. For example, in Section 8 we wanted to see how important tones were in Mandarin, and got some number for FL(tones). Knowing the importance of identifying vowels, we compared FL(vowels) with FL(tones). The closeness of the values showed that tones were at least as important as vowels in Mandarin.

13 Conclusion

A language makes use of contrasts to convey information; we have proposed and empirically tested a framework for measuring the amount of use. Further statistical tests and improvements of the measure are required, but we believe several linguistic questions can already be moved from the realm of description and speculation to testable hypotheses.

14 Acknowledgement

We are very grateful to Gina-Anne Levow for help with the Mandarin data and several very useful discussions, Stephanie Stokes for the Cantonese data and introducing us to the child language literature, Bert Peeters for explaining to us how FL is viewed in the Martinet tradition, Yi Xu for details of the behaviour of tones in Mandarin, and John Goldsmith for several suggestions regarding the readability of this paper. Thanks also go to Sean Fulop, Derrick Higgins, Jinyun Ke, Caroline Lyon and Howard Nusbaum for their comments on earlier versions of this paper.

References

Albro, Daniel M. 1993. *"AMAR, a Computational Model of Autosegmental Phonology"* MIT Technical Report AITR-1450, Cambridge, MA.

Baayen, R. H., Piepenbrock, R., and Gulikers, L., 1995. *The Celex Lexical Database* (Release 2). Linguistic Data Consortium, Univ. of Pennsylvania (Distributor), Philadelphia, PA.

Carter, David M. 1987. An information-theoretical analysis of phonetic dictionary access. *Computer Speech and Language* 2:1–11.

Chao, Y. R. 1968. *A Grammar of Spoken Chinese* University of California Press, Berkeley.

Greenberg, H. H. 1959. A method of measuring functional yield as applied to tone in African languages. *Georgetown University Monograph Series on Language and Linguistics* 12:7–16.

Godfrey, J., Holliman E. and McDaniel, J. 1992. Telephone speech corpus for research and development. *Proc. IEEE ICASSP*, pp. 517–520.

Goldsmith, John. 1976. *Autosegmental Phonology.* PhD Thesis, Department of Linguistics, Massachusetts Institute of Technology.

Hockett, Charles F. 1955. *A Manual of Phonology.* International Journal of American Linguistics 21(4), Indiana University Publications.

Hockett, Charles F. 1967. The quantification of functional load. *Word* 23:320–339.

Huttenlocher, D. Exploiting sequential phonotactic constraints in recognizing spoken words. *MIT AI Lab Memo* 867.

Ingram, David. 1989. *First language acquisition: method, description and explanation.* Cambridge University Press, Cambridge, UK.

International Phonetic Association 1999. *Handbook of the International Phonetic Association* Cambridge University Press, Cambridge, UK.

Kassel, Robert. 1990. *"An informational-theoretical approach to studying phoneme collocational constraints"* MS Thesis, EECS Department, MIT.

King, Robert D. 1967. Functional load and sound change. *Language*, 43:831–852.

Kontoyannis, I. 1997. *"The complexity and entropy of literary styles"* NSF Technical Report No. 97, Department of Statistics, Stanford University.

Kučera, Henry. 1963. Entropy, redundancy and functional load. *American Contributions to the Fifth International Conference of Slavists (Sofia)*: 191–219.

Ladefoged, Peter. 1997. Linguistic phonetic descriptions. Chapter 19 in *The Handbook of Phonetic Sciences* Hardcastle and Laver (eds.), Blackwell Publishers.

Lass, Roger. 1980. *On Explaining Language Change.* Cambridge University Press.

Lass, Roger. 1997. *Historical Linguistics and Language Change.* Cambridge University Press.

Lee, T.H.T., Wong, C.H., Leung, C.S., Man, P., Cheung, A., Szeto, K. and Wong, C.S.P. 1996. *The development of grammatical competence in Cantonese-speaking children.* Report of a project funded by Research Grants Council, Chinese University of Hong Kong.

Martinet, André. 1955. *Économie des Changements Phonétiques.* Bern, Francke.

Mathesius, Vilém. 1929. La structure phonologique du lexique du tchèque moderne. *Travaux du Cercle Linguistique de Prague*, 1:67-84.

Meyerstein, R. S. 1970. Functional load: descriptive limitations, alternatives of assessment and extensions of application *Janua Linguarum, Series Minor* #99.

Peeters, Bert. 1992. *Diachronie, Phonologie et Linguistique Fonctionnelle.* Louvain-la-Neuve, Peeters.

Pisoni, D.B., Nusbaum, H.C., Luce, P.A. and Slowiaczek, L.M. 1985. Speech perception, word recognition and the structure of the lexicon *Speech Communication* 4: 75–95.

Pye, Clifton, Ingram, David and List, Helen. 1987. A comparison of initial and final consonant acquisition in English and Quiché. in K. E. Nelson and A. van Kleek (eds.), *Children's language* Vol. 6. Erlbaum, Hillsdale, NJ.

Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell Systems Technical Journal* 30:50-64.

Shipman, David W. and Zue, Victor W. 1982. Properties of large lexicons; implications for advanced isolated word recognition systems. *Proc. IEEE ICASSP* 546–549.

So, Lydia K. H., and Dodd, Barbara J. 1995. The acquisition of phonology by Cantonese-speaking children. *J. Child Lang.* 22: 473-495.

Stokes, Stephanie and Surendran, Dinoj. 2003. Articulatory complexity, ambient frequency and functional load as predictors of consonant development in children. *submitted.*

Surendran, Dinoj and Niyogi, Partha. 2003. Questioning the role of communicative efficiency in language evolution. *To be submitted.*

Surendran, Dinoj and Levow, Gina-Anne. 2003. Tones in Mandarin are as important as vowels. *To be submitted.*

Trubetzkoy, Nikolay. 1939. *Grundzüge der phonologie.* Travaux du Cercle Linguistique de Prague 7.

Wang, William S-Y. 1967. The measurement of functional load. *Phonetica* 16:36–54.

Xu, Yi. 1993. *Contextual tonal variation in Mandarin Chinese.* PhD Thesis, Department of Linguistics, The University of Connecticut.

Xu, Yi. 1994. Production and perception of coarticulated tones. *J. Acoust. Soc. Am.* 95: 2240-2253.

Zee, Eric. 1999. Change and variation in the syllable-initial and syllable-final consonants in Hong Kong Cantonese. *Journal of Chinese Linguistics* 27, 120–167.