Phonological Acquisition and Ambient Language :
A Corpus Based Cross-Linguistic Exploration

SANDRA WARREN

A thesis submitted in partial fulfilment of the
requirements of the University of Hertfordshire
for the degree of Doctor of Philosophy

Volume 1 of 2 Volumes

The programme of research was carried out in the
Department of Humanities
Faculty of Humanities, Languages and Education
University of Hertfordshire

December 2001

Phonological Acquisition and Ambient Language :
A Corpus Based Cross-Linguistic Exploration

Abstract

This study explores the relationship between phonological acquisition and ambient language. By utilising corpora of spoken adult languages (as an indicator of ambient language characteristics), an empirical rather than a theoretical approach is adopted for this research and through the development of a method of assessment (FUSE) it is proposed that a novel way of observing and illuminating this relationship can be made.

The concept of FUSE aims to combine previous measures of functional load, which are system based, with the recognition of the relative usage of particular phonemes by speakers of the language to differentiate between words. The phonemic systems of the two historically unrelated languages of Finnish and English have been assessed and phonemic usage data compiled. Child data samples for five children over three ages of development for each of the languages have been likewise assessed so that the similarity in usage between adult and child data samples could then be compared.

This study shows that the characteristics of the ambient language that the child needs to acquire varies for the two languages in terms of length of words (Finnish has longer words), the nature of syllable structures and phonological usage. Finnish was also found to have less WI minimal pairs and make less use of word initial contrastive phonemes.

The findings indicate that ambient language does have a role to play during the children's phonological development process. The children's usage closely reflected the adult language on both frequency and FUSE based assessments. The movement by the children towards their language specific goal was better indicated with the FUSE method of assessment than a purely frequency based assessment. When test cross-linguistically the FUSE results tend to suggest that the children, already at age 2, are showing few universal tendencies of usage.

Phonological Acquisition and Ambient Language :
A Corpus Based Cross-Linguistic Exploration

Table of Contents

Table of Contents (continued) **Page No.**

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

List of Tables

List of Tables (continued)

Chapter 1 : Introduction

1.1 Aims of the Research

The primary aim of this research is to explore the relationship between ambient language and phonological acquisition. By utilising corpora of spoken adult languages (as an indicator of ambient language characteristics), an empirical rather than a theoretical approach is adopted for this research (see Sampson 2001) and through the development of a method of assessment (FUSE) it is proposed that a novel way of observing and illuminating this relationship can be made.

The new method of assessment will be achieved by combining knowledge of the phonological and phonotactic systems of two historically unrelated languages, Finnish and English, with the demonstrable use that actual speakers of these languages make of these features in spoken language. By observing the application of the permissible characteristics of English and Finnish, as seen in the occurrence of specific phonemes and structures in language use, it is proposed that a new measure of the relative systemic importance of specific components can be ascertained. By observing the relationship between ambient and child language over several ages it is proposed that any development of this relationship can be seen. Also, by observing the languages cross-linguistically with this method any universality between children's developing phonology cross-linguistically can be explored.

The objective will be to develop a method of assessment which enables the component parts of the English and Finnish phonological systems to be ranked in terms of their relative importance to the speakers of the language. The assessment method must be designed in such a way that it can be applied to any language (to provide consistency in cross-linguistic studies) and to both adult and child languages (so that the effect of ambient language can be observed in terms of the child's developing phonemic system). The assessment method will be based on the principle of functional load, a phoneme's potential importance to the language (measured in terms of the minimal pairs or contrasts that it appears in), extended to additionally incorporate a measure of use as seen in the most frequently spoken words.

The FUSE method of assessment will enable a way of observing:

a) the extent to which there is a relationship between the phonological usage of children acquiring a particular phonological system and the adult language they are acquiring and how this relationship changes during phonological development.

b) whether there appears to be a universal cross-linguistic basis from which the children's phonology develops or whether even early phonological usage reflects a language specific base.

One result of analysing spoken language corpora at a detailed level will be new frequency statistics on the structures of words and syllables and phonemic usage. A third outcome will therefore be to:

c) provide frequency data for adult and child spoken English and Finnish.

The method of assessment will, in this study, be applied to phonemes in word initial position for the two languages of Finnish and English. Additionally, a purely frequency based assessment of the relationship will be made as a comparison with the FUSE findings.

1.2 Background to the Study

Linguists disagree on the role that ambient language, the language surrounding a child, plays in the child's developing phonological system (see Stoel-Gammon 1998, Ingram 1989, Jakobson 1968 and Chomsky 1965).

Historically linguists have argued both that:

- the specific language surrounding a child has an effect on the child's emerging language

- that the language surrounding a child has little or no input to the developmental process which is instead governed by some innate and universal faculty.

Language specific acquisition arguments were often supported by the observation that children learning different languages do in fact demonstrate different language patterns and acquisition paths. Also research into adult frequency and child production (Stoel-Gammon 1998, Ingram 1989, Dobrich & Scarborough 1992, Pye, Ingram & List 1987) has observed that children acquire earlier the phonemes and structures which are most frequently observed in adult languages surrounding them.

Alternatively the universalist argument based its premise on the observation that certain universal trends do seem to occur during development despite the fact that children are acquiring structurally different languages. Jakobson (1968) and Stampe (1969), for example, argued that phonological acquisition follows innate universal rules that are pre-determined by language universals or innate processes rather than language input from the learning environment itself.

Theories of phonological acquisition that have previously attempted to link phonological acquisition to features of ambient language have usually done so on the basis of predicted correlation between frequency of use within a particular language and acquisition order of phonemes (e.g. Skinner 1957, Olmsted 1966). Universalist theories have instead predicted correlation between frequency of use across the languages of the world and acquisition, suggesting a common approach to acquisition across languages.

As an example of the two different approaches, using frequency in ambient environment as predictor of early acquisition, would predict that the phoneme /ð/(as heard in the first sound of the English word 'the') and which is the second most frequently spoken English fricative (Wang & Crawford 1960) would be acquired earlier in the child's developing phonological system than a less frequent phoneme such as the English fricative /k/. The universalist approach, based upon predicted correlation between frequency in terms of a phoneme's universality (occurrence in languages of the world) and the acquisition order would instead predict that the phoneme /k/ which

is observed in more different languages than the phoneme /ð/ would be acquired earlier. Research to date, however, has not supported either of these theories alone as accurate predictors of phonological development (e.g. Sander 1972) and both language specific and universal factors appear to influence phonological acquisition. One problem has been how to measure and demonstrate the relative influences of universal and language specific factors during the acquisition process.

Functional load (Greenberg 1959, King 1967, Wang 1967, Meyerstein 1970) is a concept which maintains that a phoneme carries a certain 'weight' in terms of its necessity within a particular phonological system to distinguish between words of the language. It provides a way of recognising the importance of phonemes within the phonemic system and a way of measuring the relative importance of phonemes in terms of their systemic 'usefulness' as opposed to their 'frequency' of occurrence. If we accept that the function of a phonological system is lexical differentiation then functional load offers a way of measuring the phonemes that provide the most and the least differentiation, those that afford the greatest and least degree of contrast between linguistic units as measured in the number of minimal pairs (words which differ by only one phoneme) observed. It can thus be used to demonstrate the relative importance of phonemes in terms of their 'usefulness' to the language.

As Chapter 2 will explain, each language has its own unique phonological system, both in terms of a phonemic inventory and phonotactic rules. Each language makes use of the possibilities to use particular phonemes and structures in the lexicon (words of the language) to differing extents. Whilst languages might indeed share phonemes (as universalist theories observe) languages make different usage of these in their lexicons thus resulting in languages having their own functional load rank orders.

Pye et al (1987) also argued that one explanation for the cross-linguistic differences observed in child acquisition could be the linguistic experience of the child (i.e. the ambient language surrounding them). They argued, however, that it was not, in fact, the absolute frequency of a particular phoneme that determined its acquisition position but instead its functional load (i.e. its systemic usefulness). In the example above, the phoneme /ð/ which occurs as the contrast phoneme in relatively few different minimal

pairs, would be acquired later than /k/ which occurs in more words. Whilst not tested in any systematic fashion this hypothesis provides the basis for suggesting that children's acquisition of phonemes will be linked to the systemic usefulness of phonemes rather than purely their frequency of usage in adult language. Other research (e.g. Leinonen-Davis 1987) has also suggested that recognising the systemic importance of phonemes within a particular phonological system may enable a better observation of the processes of acquisition.

One benefit, therefore, of using the functional load approach is that the phonological 'system' is recognised as an influencing factor. One criticism of functional load has been that it treats all potential oppositions with the same weighting (Catford 1988) and does not also recognise the importance of frequency of usage. Therefore, whilst functional load recognises the relative systemic importance of phonemes, in terms of their potential 'usefulness' within words of a language's lexicon, it does not take into account the 'actual' use that the language makes of the phonemic system in terms of the words typically appearing in ambient language.

The method developed in this study aims to combine the systemic emphasis demonstrated in the functional load approach together with a recognition of the importance of language usage. The use of corpora of actual spoken language (rather than theoretically possible language use) will directly indicate the phonemic usage information required and also will enable the 'actual' (as opposed to theoretically possible) minimal pairs of both Finnish and English to be further assessed.

The ability to apply a constant method of assessment, one that uses the same input criteria of spoken language in a systematic process of measurement, one that can be applied to both adult and child language and one that can be applied cross-linguistically, could not only provide a new and potentially invaluable insight into the language specific vs. universalist argument but will provide a way of observing the differences between the phonological systems of languages and the relationship between adult language and children's developing phonological and lexical systems. The possibility of applying the method of assessment developed in this study to several stages of child development would additionally enable one to observe how the

relationship changes during the acquisition process. The ability to apply the assessment method to languages from different language families might also enable new previously unobserved tendencies to be found.

1.3 Scope of Research

As the vast majority of research into phonological acquisition has been carried out in English and other Indo-European languages the decision was made to select a language from a different language family alongside English. Finnish is a member of the Finno-Ugrian language family. It is spoken by approximately 4.6 million people in Finland as a first language and by the same number again in Sweden, North America and Australia (Iivonen 1998). Standard Finnish is phonetically based on Häme dialects and is mainly spoken in Central and Southern parts of Finland. It is unrelated historically to the English language apart from the use of many loan words.

The Finnish language was therefore selected as providing a useful comparison with the Indo-European language of English for the testing of the method and its usefulness in understanding phonological development. Any correlation found between the FUSE findings for the two unrelated adult languages or the developing phonemic systems of English and Finnish children will be difficult to explain other than with universal theories which would look at the similarities in the features of the two languages.

Before commencing any analysis of the possible relationship between a) the phonemic systems and word usage, and b) child phonological development, a standardised approach to the presentation of the various systems needed to be agreed so that they could later be compared and that correlation analysis could be undertaken. Providing such a basis for assessment required a thorough analysis of the phonemic systems of the two languages so that the resulting understanding would be suitable for both the adult and the child analyses. Various issues to do with the representation of a common framework presented themselves during the undertaking of this task and these are outlined in Chapter 2. Chapter 2 also provides a description of the English and Finnish phonemic systems both in terms of their inventories of phonemes and how these appear within the syllable and word structures. Furthermore, an overview of previous research

on the frequency of the various features as observed in spoken language of the two languages is explored.

In Chapter 3 phonological acquisition theories are described further and two opposing theories, one universalist and the other frequency based and language specific, are expanded to explore the predictions they make for the phonological acquisition of Finnish and English. The concept of functional load is also further explored here as a direct input to the new method developed for this research. The universalist approach will look at the concept that a phoneme's universality across languages of the world will determine its acquisition. This approach therefore neglects the role of ambient language and instead relies upon the phoneme's relative importance not within a language but across languages. The frequency data given in Chapter 2 will on the other hand be totally language specific and can be used to produce frequency based predictions for the two languages.

Once the framework for assessment had been developed there were several ways of approaching this research. A purely theoretical approach based on the phonemic systems of the two languages and the permissible sequences of phonemes (as predicted for syllable and word structures) was possible. It was felt however, that a measure of 'use' should at least observe the actual words of the languages in question rather than simply those sequences of phonemes that theoretically could form words (i.e. follow the phonotactic rules of the language) but do not exist as words within the language. Dictionaries provide information on 'real' words of the languages (i.e. what combinations of phonemes actually form words of that language). However, they do not provide usage information required in order to understand not what combinations might be possible nor simply what permissible combinations are words but which words are typically and most frequently used by speakers of that language. Whilst readily available and electronically accessible it was felt that dictionary collections of words do not truly reflect ambient language.

Without reference to the frequency of use of particular words both highly frequent and extremely rare words would be included in the analysis as having potentially the same chance to influence development. Problems related to representation of words in

dictionaries also existed. Dictionaries usually present words in their lemmatised forms, base forms without inflection, whilst language use is fully inflected. The streams of phonemes heard in language would therefore be expected to be very different from a dictionary representation of lemmatised words. It was decided therefore that corpora of spoken language would best suit the needs of this research. They directly present the rules of the language in the forms of the words that are both permitted as well as used in the language incorporating inflections and providing a direct measure of frequency of usage.

A corpus based approach whereby corpora of English and Finnish, adult and child data, could be taken as the direct representation of language usage was therefore adopted. Four corpora containing samples of spontaneous spoken language of Finnish and English children and adults were sought. That they should each enable an assessment of phonemic use was a pre-requisite to the selection of the corpora. The measure of functional use relies upon having available both the most frequent words of spoken language as indicators of 'use' and the phonemic and phonotactic sequences as indicators of the system in use. Corpora thus provide the evidence of those phonemes that are utilised within different syllable and word structures. An overview of the issues generally associated with Corpus Linguistics and an empirical approach to linguistic analysis is provided in Chapter 4, and the various corpora available for each of the languages used in this research are described.

Having observed the phonological system and the phonotactic rules particular to the two languages under observation, the permitted phonemes and sequences of phonemes, the processing rules for functional use calculation were devised. The original plan was to systematically delete the contrasts in the system so that the lexical reduction, in terms of the amount of resultant homophony (duplicate identical phonemic strings) produced, could be measured, much in line with traditional functional load approaches. The amount of homophony produced by each phoneme's deletion from the system could then be counted and the phonemes could be ranked in terms of most 'useful' to the lexical systems by word position. The rationale for this being that the loss of the more functionally useful phonemes would cause the most homophony. The resultant method for the calculation of FUSE and the development of the approach from a

phonotactic theoretical rule based method to one that eventually became more data driven, and the reasoning for this development, is provided in Chapter 5.

Even at an early stage of development of the FUSE method it was recognised that certain areas of the application of the method might be constrained by the contents of available corpora. For example, the various sizes of the adult corpora would need to be recognised as having a direct influence on the numbers of word types presented. It was felt that the over-riding important element would be to have as large a range of phonemically transcribed words as possible such that the contrastive abilities of phonemes would be more rigorously tested. The primary objective of analysing a true reflection of spoken language as represented in the strings of phonemes seen in words was felt to be of paramount importance. Whilst frequency information was also necessary in order to enable a selection of the most frequent words that best represent ambient language, as large a range as possible of different 'phonemic' representations of word forms was sought. In order, therefore, to achieve this aim of utilising already collected data sources and yet have as large a range of 'phonemic' representations as possible, certain restrictions to the method had to be allowed for due to the constraints in the available data. More specifically spoken language phonemic transcriptions (that were a minimum requirement) did not always reflect the level of frequency information required. In order to test the method on as large a range of phonemically transcribed word forms as possible for the English adult data, for example, meant a re-think of the 'usage' input to the FUSE method as frequency referred to orthographic rather than phonemic form. The application of the FUSE method to particular data sources, and their limitations are discussed in Chapter 6.

The first application of the method of measurement had to be developed to incorporate the known limitations of the readily available data at that time. The data limitations needed to be considered at this stage. Rather than applying the method for each of the four areas under assessment (i.e. adult and child English and Finnish) and coming up with differently created FUSE rankings for each area, a consistent approach was adopted across the four data sets such that both cross-linguistic and child to adult comparisons could be made.

Once the FUSE method of assessment had been developed, based around the available corpora and recognising the phonemic systems of the two languages under investigation, the method was applied to the four areas of analysis, adult English, child English, adult Finnish and child Finnish. The details of the actual data processing required for each of the four areas, including the corpus selection, is provided in Chapter 6.

For the initial application, where the usefulness of the method itself is being assessed it was decided to analyse in detail only the word initial phoneme position. The calculation of the FUSE totals and the rankings for each of the four areas under investigation was completed for only the phonemes seen at word initial position.

Chapter 7 provides the results of the FUSE calculations for each of the four data sets. In processing the data a large amount of general information about the actual frequencies and structures observed in the data was also ascertained. These general findings in some cases offer new knowledge, especially for Finnish, and are also, therefore, presented in Chapter 7.

Two ways of assessing the usefulness of the FUSE method were developed for this study. Firstly, the results of the adult FUSE rankings were compared with child phonological usage and acquisition order over the three ages under assessment . Thus, if we had found, for example, that the phoneme /s/ ranked the highest in word initial position according to the FUSE counts for Finnish, we might test the prediction that the Finnish child would use the phoneme /s/ more frequently or acquire this phoneme earlier. This study is not focusing particularly on early stages of development where phonemic acquisition order would be more relevant (see Chapter 3) but on the later age range of 2 to 5 years where many of the phonemes will have already been acquired. By looking at the frequencies of phonemes observed in the children's words at the three ages under analysis and ranking them accordingly, an assessment of how the children's frequencies match the adult FUSE rankings (representing ambient language) can be made. A strong correlation here would indicate that the FUSE calculations could be used as a direct prediction tool for phonological usage; the phonemes ranked the highest for FUSE in the adult data are utilised the most frequently by children acquiring

that language and phonological system. Observing similarities between the adult FUSE rankings and the frequency of use of phonemes by the children can be taken to suggest that ambient language does have a measurable effect on the child's acquisition route.

The child FUSE rankings, as seen at various stages of development, enabled an assessment of the internal lexical to phonemic processes operating within the system of the child as demonstrated by the words that the child chooses to use. A comparison of these child FUSE rankings with the adult FUSE rankings, representing the ambient language, provided a second way of assessing the FUSE method as a tool for exploring phonological acquisition.

Following the principles adopted for the adult language rankings whereby both the relative importance of phonemes, in terms of their abilities to signal contrasts within the words of the language, and the usage, in terms of the actual words used, are combined together to give a representation of functional use, the same procedure can be applied to the children's word lists at various stages of development. The children's words provided FUSE rankings which could be tested for correlation to the adult FUSE rankings. A move towards the adult FUSE rankings is expected here as after all the children are acquiring the adult phonological system and lexical base (i.e. words) but what FUSE provides will be a measure of how close the adult and child systems start and also how quickly the acquisition takes place. A close correlation at the outset of the child's development would be indicating that already, at the outset of lexical development, the child is using the phonemes that provide the most differentiation in that language.

As a direct comparison with the FUSE method of assessment, a purely frequency based assessment method was also applied. The adult word initial phoneme frequency rankings for the two languages were compared with the child word initial phoneme frequency rankings over the three ages to assess how closely they correlated. Both the starting relationship between adult and child at age 2 and whether any movement towards the adult language being acquired could be detected were assessed.

Comparing the results of the various English child to English adult correlation findings (i.e. frequency to FUSE, FUSE to FUSE and frequency to frequency) with the Finnish child to Finnish adult findings enabled a cross-linguistic view of the usefulness of the various assessment methods for the two languages.

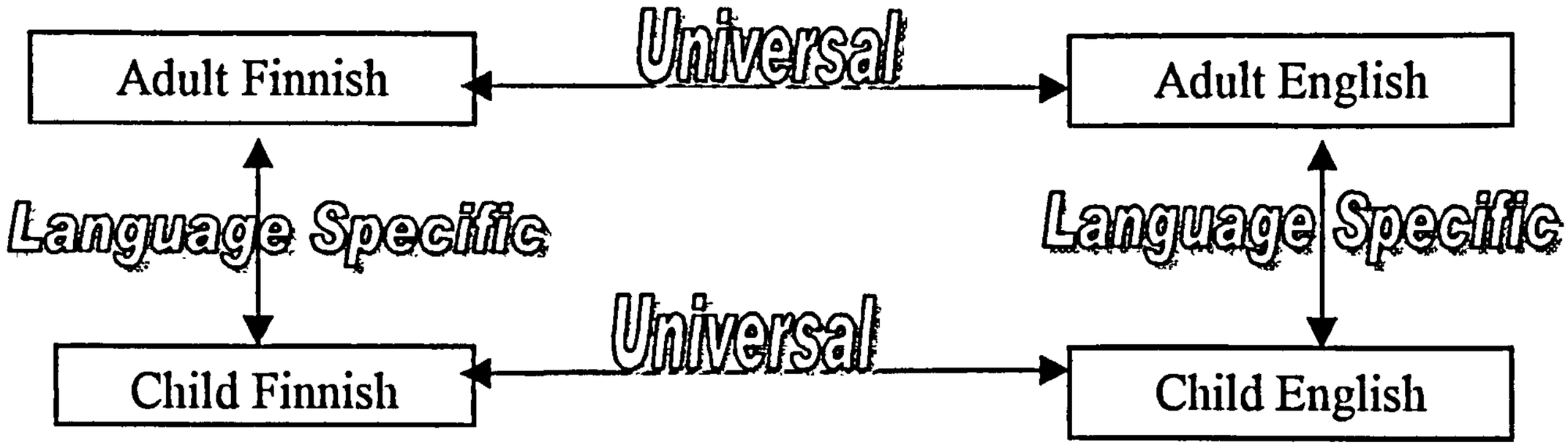
As a measure of the languages similarities the two sets of adult rankings, both word initial frequency rankings and FUSE rankings, were compared cross-linguistically. The rankings represent different phonemic systems and word usage and a lower correlation between the two adult languages rankings than between the adult and child of same language is expected. A close correlation between the two languages would be demonstrating a high degree of universal features that are in fact common to both languages.

Completing an assessment of adult FUSE for the two languages enabled an assessment of whether the same most functionally useful phonemes appear in the words of both languages despite the different phonological systems. Observing similarities or correlations between the adult language rankings can be considered to be indicating some universality inherent in the two systems, which may be reflected in the children's acquisition of the two languages.

As a final test of the usefulness of the FUSE method for exploring phonological acquisition, the FUSE rankings for the English and Finnish children were correlated. A close initial correlation would again be demonstrating a universal base from which phonological acquisition proceeds towards the language specific requirements of particular languages. The FUSE assessment provides a way of demonstrating this movement whether from a universal basis towards the ambient language surrounding them or alternatively, a closer correlation with the adult language FUSE initially indicating a language specific base.

These relationships, which are further explored in Chapter 8, can be demonstrated with diagram 1.1.

Diagram 1.1 – Language Specific and Cross-linguistic Relationships



In summary, for both the Finnish and English languages there are two main areas of analysis: adult spoken language and child spoken language. This gives four distinct processing phases and four sets of results. The findings for each of the four areas, adult English, adult Finnish, child English and child Finnish, and for word token frequency, word type structures and phoneme usage are provided in Chapter 7 together with the FUSE rankings for each of these four areas.

The application of the FUSE method to explore the interactions between the adult language and child language for the two languages is firstly tested by assessing the correlation between child phoneme frequency and adult FUSE. Secondly, child FUSE to adult FUSE is compared. Thirdly, the adult frequency to child frequency findings are compared. Finally the ranking results for the two languages are compared cross-linguistically for both frequency and FUSE. The various comparisons between the frequency and FUSE based assessment measures are provided using correlation statistics in Chapter 8.

As a final view on the process of acquisition the child data is presented for each child individually in Chapter 9 so that non-parametric linguistic analysis can be further explored.

Chapter 10 presents a discussion of the findings and identifies further areas of development for the new method and further possible areas of application.

Chapter 2 : Phonology

2.1 Introduction, Context and Definitions

The human vocal organs are capable of producing a wide range of sounds from the sounds that we might typically associate with language to others that we do not (e.g. whistling and clicking of teeth). Despite the fact that languages evolve, that there is always individual variation in the actual speech production and that the study of new languages may bring the discovery of new phonetic phenomena, linguists now feel that they have a basis for discriminating between linguistic and non-linguistic sounds (Ladefoged & Maddieson 1996).

The means of describing human speech sounds from a phonetic viewpoint, specifying the articulatory and/or acoustic properties of sound, is well established (e.g. Clark & Yallop 1995 and Ladefoged & Maddieson 1996). For example, the first sound of the English word 'mat' might be described using articulatory descriptions based upon the anatomy and physiology of the speech production organs, the position of the tongue, larynx and lips, as being a 'bilabial nasal' sound (Roach 1991). The sound might equally be described in acoustic terms as a sound wave of a particular resonance, amplitude, duration and frequency (Clark & Yallop 1995).

Another way of viewing these sounds is not simply as physical phenomena but as parts of a language system whose prime objective is to convey meaning (Ladefoged & Maddieson 1996). Observing sounds in terms of the function that they perform within a particular language enables the sounds that are used to differentiate between words of a particular language (and therefore serve to differentiate between meanings) to be identified. For example, in English the two words /kæt/ 'cat' and /mæt/ 'mat' have different meanings and they are separated by just one sound difference /k/ or /m/ as the first sound.

The two contrasting words are known as minimal pairs and they clearly demonstrate the sounds which act as contrastive segments within the language, serving to differentiate between meaning. Other examples of English minimal pairs include the

words /mit/ 'meet' and /fit/ 'feet', /dip/ 'deep' and /lip/ 'leap and /mit/ 'meet' and /min/ 'mean' etc. These contrastive sounds are known as phonemes and the phonemes can be said to be distinctive. It is this phonological approach to sound descriptions, observing sounds as part of a language system, that will be of interest in this study.

Each language has minimal pairs which serve to demonstrate which sounds of the language provide the contrasts and are phonemes of that language system. Whilst the actual acoustic properties of the pronunciation of particular phonemes will vary from speaker to speaker, a listener will observe the sound as part of a system that contains a specific number of phonemes that serve to differentiate meaning within that language.

Allophones, variations in the pronunciation of sounds based upon the influence of neighbouring sounds (Crystal 1987), are also observed to operate within languages. For example the phoneme /t/ may be realised as aspirated as at the beginning of the English word /ti/ 'tea' or unaspirated as in /it/ 'eat'. However, the pronunciation of /ti/ 'tea' or /it/ 'eat' with either an aspirated or unaspirated /t/ does not cause a change of meaning for English. The two forms or realisations are allophones of the phoneme /t/ (Roach 1991). In other languages, however, this difference in sound based upon aspiration/unaspiration might effect a change in meaning and each language must therefore be assessed not in terms of the actual realisation of sounds but in terms of their function within the system.

Other allophones of English include the various realisations of the phoneme /ʃ/. For example in the word /ʃu/ 'shoe' it is pronounced with the lips rounded but in /ʃi/ 'she' with the lips spread. English does not distinguish between the meanings of words using a contrast of lip rounded or spread for the phoneme /ʃ/, while the language of Lak (North Caucasian), for instance, does (Crystal 1987). Alternatively English might have phonemes (which serve to contrast within the English system) that are allophones in another language.

In addition to segmental phonemes being used as contrastive units some languages utilise supra-segmental features such as tempo, loudness and the prosodic features of tone and stress. For example, the Chinese language of Putonghua (Modern Standard Chinese) utilises tone as a phonemic feature. There are four such tones in Putonghua (high level, high rising, falling-rising and high falling) primarily characterized by voice pitch but also by length and intensity. Differences in tones can change the meaning of a word (Hua & Dodd 2000). For both Finnish and English these supra-segmental features serve to primarily provide emphasis rather than differentiate meaning. The current study does not therefore include supra-segmental feature descriptions.

Thus from the wide range of possible speech sounds that humans can make each language has its own particular inventory of sounds that it utilises to convey meaning in the form of words. As each language presents its own set of phonemes a thorough analysis of the systemic function of particular sounds for both Finnish and English must be undertaken before an inventory of phonemes can be drawn up. The first task of this study will therefore be to determine and describe all the segments that are known to distinguish lexical items within the English and Finnish languages. This is completed in Sections 2.2 and 2.5.

Whilst each language uses a specific number of contrasts in the form of phonemes to differentiate between words of that language the actual phonetic description of the pronunciation of particular phonemes might vary according to specific accent. A standard pronunciation of English, Received Pronunciation or RP, provides the basis of the phonetic description of English phonemes and the Häme dialect of Finnish provides the basis of description for Finnish.

As well as the variation in components of the phonemic system, the size of the segmental inventories of different languages varies. The number of segments which can be observed operating in languages of the world varies considerably from only 11 segments in the languages of Rotokas (Indo-Pacific) and Mura (Chibchan) to 141 in the language !Xu (Maddieson 1984). Segmental inventories are in the majority of cases between 20 and 37 segments in total (Crystal 1987) although it must also be recognised that segmental phonemes are not the only contrastive/distinctive feature of a language.

As discussed above, some languages also make use of supra-segmental features such as tone to provide the contrastive/distinctive feature but as these features are outside the remit of this study they are not presented further here. Whilst it might be expected that languages with comparatively small segmental inventories (under 20 segments) would suffer from problems at the morphemic level where a lack of contrastive possibilities results in high levels of homophony or particularly long morphemes, this does not appear to be the case. One example quoted by Maddieson (1984) is the language of Hawaiian which has an average of 3.5 phonemes per morpheme despite having an inventory of just 13 segments. RP English is generally recognised as having 44 phonemes (Roach 1991) however, such a straightforward assessment for Finnish is not possible (as discussed below).

Speech sounds, or phonemes, can be further divided into two main categories; vowels, where there is no obstruction to the air flow as it passes from the larynx to the lips and consonants, where there is obstruction to the air flow (Roach 1991). Each language again has its own inventory of vowel and consonant phonemes. Maddieson (1984) cites a mean ratio of vowel to consonant phoneme of .4 in languages of the world (i.e. there are usually almost twice as many consonant phonemes as vowel phonemes). Several intra-language dependencies have been observed with the sounds that are used in languages. The implication that if a particular phoneme X occurs then the phoneme Y will occur in the same language generally proves consistent for the majority of languages. For example, it is observed that nasals do not occur unless stops occur at the same place of articulations and there are only 5 known exceptions to this rule (Crystal 1987).

Whilst the initial basis of what constitutes a phoneme will be whether it can be seen to be operating as the contrast in minimal pairs, description based upon the articulatory characteristics of the various phonemes will enable the initial identification of the phonemes. Vowels are generally described phonetically in terms of their height and backness within the vocal cavity (Roach 1991). In general terms there are four or five vowel heights and three degrees of backness. Other parameters that are used to differentiate between vowel phonemes of some languages are lip attitudes, length, diphthongization and nasalisation. Vowel sounds are also described in terms of the

shape and position of the tongue. Open, mid or close is used to describe the vertical distance between the upper surface of the tongue and the palate. Front, central or back is used to describe the part of the tongue which is raised the highest. The length of the vowel sound relative to others can be described as being long or short. To enable some vowel phonemes to be further uniquely identified the laxness of the mouth and the roundness of the lips needs also to be given. These articulatory descriptors for each of the vowel phonemes of English and Finnish, whilst not of particular interest for this study, will be utilised during the phonemic inventory explanations such that each vowel phoneme can be uniquely identified from others.

Vowel phonemes are described as 'pure vowels' when they remain relatively constant, diphthongs when they glide from one vowel towards another and triphthongs when they glide from one vowel towards another and then towards a third (Roach 1991). Each language will make different usage of these various types of vowel phonemes. According to Lass (1984) the vowel inventories of different languages vary in size from the smallest vowel system which is comprised of only three pure vowel phonemes, as seen in the languages of Aleut and Moroccan Arabic, to the largest, Alsation German, which has twenty one. Maddieson (1984) using a different segmental approach (described below) finds a larger vowel system of forty six vowel phonemes and a mean vowel inventory of 8.7 vowel phonemes in world languages. Based upon the number of vowel phonemes that a language has, predictions can be made on the actual phonemes that these will be. For example, a minimal vowel system containing only 3 vowel phonemes will always have an [i] or [ɪ] or [e] plus a [u] or [ʊ] or [o] plus one low vowel either [æ] or [ɑ:] or [ɑ] (Lass 1984). Two other general findings are that mid vowels do not occur in a phonemic system unless high and low vowels occur and rounded front vowels do not occur unless unrounded front vowels of that same basic height occur (Maddieson 1984).

A phonemic description of vowels as components of a system acknowledges that the vowel phonemes are sounds that can occur on their own or at the centre of sequences of sounds (Crystal 1987). They therefore form the nucleus or basis of syllables, a concept discussed in more detail below in Sections 2.3 and 2.6.

Consonant phonemes can be described in articulatory terms in relation to place and manner of articulation and degree of voicing (voiced vs. voiceless).

The International Phonetic Alphabet provides for;

- eleven places of articulation: bilabial, labiodental, dental, alveolar, postalveolar, retroflex, palatal, velar, uvular, pharyngeal and glottal.

- ten types of articulation; plosive, nasal, trill, tap/flap, fricative, lateral fricative, approximant, lateral approximant, ejective stop and implosive.

- two degrees of voicing; voiced or unvoiced.

It is these descriptors that will be used as a basis for the English and Finnish consonant phoneme inventories such that each phoneme can be uniquely identified. From a purely phonological perspective consonants might be described as those sounds that cannot occur on their own in words or cannot form the nucleus of a syllable (Crystal 1987).

There is great variation in the number of consonant phonemes observed in languages ranging from 8 in Hawaiian to 80 in Ubykh according to Lass (1984) and from 6 to 95 (with a mean of 8.7) according to Maddieson (1984) . It is possible to make certain general predictions on the exact consonant phonemes that will be utilised within a phonemic system of a certain number of phonemes. For example, it is known that all languages have obstruents and the minimal obstruent system of three obstruent consonant phonemes is observed to be two pulmonic oral stops from the set /p, t, k/ and a glottal stop / ʔ/ or the three pulmonic oral stops /p, t, k/ (Lass 1984). The consonant phoneme /k/ does not occur without /t/, /p/ does not occur without /k/ and nasal consonants do not occur unless stops (including affricates) occur at the same place of articulation (Maddieson 1984).

2.1.1 Exploring the Concept of a Phonemic Inventory

Section 2.2 sets out to define the English phonemic system in terms of an inventory of vowel and consonant phonemes. Section 2.5 does the same for the Finnish phonemic system. However, in determining the phonemic inventories and phonotactic rules of Finnish and English there is one particularly interesting area which is common to both languages (the language specific issues will be presented in the relevant language sections). This discussion involves both the recognition and representation of 'sequences' of phones which tend to operate together as units in minimal pairs and as units to provide the contrasts between words. Both Finnish and English demonstrate sequences of consonant phonemes and vowel phonemes which can be seen to serve together to provide the contrast between words. For example, consonant clusters (sequences of consonants before or after a vowel nucleus in a syllable) occur in both languages. (e.g. /br/ at word initial position in the English words 'break' and 'bring', /bl/ in 'black' etc.). Consonant clusters as units seem to provide the contrast in minimal pairs. For example, in the minimal pair words 'black' /blæk/ and 'back' /bæk/ the consonant cluster /bl/ could be said to be providing the contrast with the consonant /b/. Whilst the phonotactic rules of the languages will determine which sequences might appear alongside which others, an assessment of those that carry a functional purpose, i.e. they serve to differentiate in the language, must be made. These sequences of phonemes, phonemic units which serve to differentiate in much the same way as singleton consonant or vowel phonemes, must be taken into account when building up the inventories of phonemes for the two languages.

The second issue, related to the treatment of 'units', has to do with the transcription standards utilised for the frequency measurements. Whilst it may be acceptable to recognise the diphthongs and triphthongs of English (Roach 1991) and the diphthongs, long vowels and geminates of Finnish (Vainio 1996) as 'units' for phonological analysis they are actually represented in transcripts as a series of phonemic transcription codes e.g. /eɪ/, /kk/. Some frequency analysis has treated these sequence as separate and dividable units for the purposes of reporting a phoneme's frequency. For example, the research on Finnish phoneme frequency completed by Pääkkönen

(1973) counts the frequency of phonemes based upon their orthographic representation in large spoken language corpora recognising sequences such as 'kk' (a Finnish geminate) or 'aa' (indication of a Finnish long vowel) as two phonemes. Similarly, in both English and Finnish frequency research consonant clusters which are comprised of more than one consonant phoneme (e.g. /br/, /sn/) have variably been treated as both 'units' and separate phonemes by different researchers. The different possible treatments of these 'units' will again result in very different frequency measures being found as well as different sets of minimal pairs being identified. For this reason not only will it be important during this initial definition of the phonemic systems of the two languages that units are searched for (with the same criteria for phonemes) but also that they are systematically treated through both the frequency analysis and the functional use measurement. A consistent approach must be applied to each language, for both adult and child word lists. The facility to not only recognise these sequences as phonemic units but to represent them as indivisible units in the processing must be explored through the development of functional use.

It is important to discuss sequences in this study, as their status has repercussions on both the representation of frequencies, the predictions for phonological acquisition and for the later processing of data. The two languages have therefore been individually examined in light of these sequences that might be regarded as units within the phonemic systems and the merits of each approach in relation to the two systems is discussed in Sections 2.3 and 2.6 below.

The research carried out in this study frequently involves close attention to details of English and Finnish pronunciation and the representation of sounds as phonemes needs to be clearly defined. The phonemic inventories provided in sections 2.2 (for English) and 2.5 (for Finnish) will utilise as many parameters of contrast as are necessary in order to uniquely describe each of the phonemes and identify the sound or phone that they represent and a unique symbol or code taken from the International Phonetic Alphabet will be assigned to it for future reference. Finnish word forms will be shown in italics (e.g. *talo),* English words forms (and the English translations of Finnish word forms) will be shown within single quotation marks (e.g. 'house') and phonemic representation will be within / /(e.g. /talo/).

The symbols of the International Phonetic Alphabet enable all the phonemes to be uniquely represented. However as only 44 symbols are needed to represent the phonemes of English and 21 to represent Finnish and as these symbols do not have to indicate precise phonetic quality there is in some cases a choice of symbol to represent a particular phoneme. For example, in order to utilise the symbol that is as close to the symbol that a phonetician might use for the vowel sound in the word 'cat' the symbol /æ/ should be utilised representing the sound of the vowel phoneme as being front and mid open. Some transcriptions use the symbol /a/, which phonetically indicates the cardinal open front vowel sound (a sound which is not heard in 'cat') simply because it is easier to present. Another example of differences in representation of two phonemes is seen with the phonemic representation of the vowel sounds in the words 'fit' and 'feat'. If the quality of the sound is taken to be more important than the length then the two vowels would be transcribed respectively as /ɪ/ and /i/. Alternatively, if length was seen as the main factor then the vowels could be represented simply as /i/ and /i:/. Other transcriptions use the two symbols /ɪ/ and /i:/. It is necessary therefore to adopt a standard for the representation of phonemes that will be utilised throughout this study such that phonemes of both the English and Finnish systems can be uniquely identified. In the main the standard is based on Gimson (1994). However for ease of presentation where one symbol suffices to uniquely represent a phoneme then this is adopted e.g. /ɪ/ for the short sound and /i/ for the long sound in English, /e/ and /E/ in Finnish. Diphthongs and triphthongs, affricates and other sequences which are accepted as 'units' operating together within the phonemic system also need to be identifiable as units when assigning symbols or codes for the inventory.

2.1.2 Phonotactic Structures

Phonemes can also be described by looking at their distribution in syllables and words (i.e. the word contexts in which they appear), both in terms of the sounds they can appear alongside, and the positions that they can appear in. For example, the vowel/consonant differentiation can be further demonstrated not, as before in terms of their production (obstruction/no-obstruction) but in terms of their pattern of distribution within words and syllables. Phonotactic description will be of particular interest for

this study as it will enable the sequences of phonemes, and therefore syllable and word structures permissible in the two languages, to be described. Thus a phoneme's description can be further enhanced by knowledge of its presence at a particular word or syllable position.

In order to present patterns of distribution a definition of what constitutes a word and syllable in phonology is necessary. Whereas a phonetic description might describe acoustic rhythm and peaks a phonology definition presents an inventory of positional constraints. Lass (1984) refers to the phonetic syllable as a 'performance unit' providing a purely phonetic description whilst the phonological syllable is a structural unit.

A purely phonetic description of the term 'syllable' would utilise the previous definitions of vowels (without obstruction to the airflow) and consonants (with obstruction to the air flow). A syllable with this definition would therefore be described as having a centre without obstruction to the airflow which sounds comparatively loud and the possibility of sounds with obstruction to the airflow and/or less volume surrounding it (Roach 1991).

According to Lass (1984) "the phonological syllable might be a kind of minimal phonotactic unit, say with a vowel as its nucleus, flanked by consonantal segments or legal clusters". Syllables with this description are seen to contain the minimum of a vowel phoneme which provides a 'nucleus' around which the rest of the syllable structure can be described. Syllable structure can then be described as a sequence of vowel phonemes (V) (including diphthongs and triphthongs) and consonant phonemes (C) or permissible consonant clusters. In English, for example, the word 'a' can be said to consist of a vowel syllable only (V) and the word 'cat' can be described as having the following syllable structure (CVC). A consonant or cluster of consonants before the vowel nucleus is known as the syllable onset (e.g. CV as in /fi/ 'fee' or CCV as in /fli/ 'flee/flea'). Consonant phonemes or consonant clusters following the vowel are known as the syllable coda (e.g. VC as in /ɪt/ 'it' or VCC as in /ɪts/ 'its'). A syllable with a vowel nucleus without preceding consonant phonemes is described as having a

zero onset. Syllables may, of course, demonstrate both of the elements of onset and coda (e.g. CVC as in /kæt/ 'cat', CCVC as in /plæt/ 'plait', CVCC as in /mæts/ 'mats' or CCVCC as in /plæts/ 'plaits', etc. A word can simply be defined as a series of one or more syllables and each language has specific rules determining the minimum and maximum number of syllables. The minimum component of an English word, for example, is one syllable. The phonotactic rules of the language thus determine how Cs and Vs can be combined in the language to form syllables and words.

There are also examples of syllables where the vowel nucleus is replaced by a syllabic consonant. For example, the second syllable of the English word /stju:dnts/ (CCCV, CCCC) 'students' appears to be comprised of a series of consonant phonemes /dnts/ without the vowel nucleus. Speakers of the language tend to accept that the word is comprised of two syllables rather than one and that the first syllable is more stressed than the second. The /n/ phoneme is here described as syllabic (Roach 1991). Other syllabic consonants are syllabic /l/ (e.g. /pe,dl/ 'pedal') and syllabic /r/ (e.g. /hɪs,tri/ 'history') (Roach 1991). Where sequences of syllables are joined together as words certain syllables will be more stressed than others and syllabic consonants never appear as the stressed syllable.

Often speakers in careful pronunciation would actually insert the vowel phoneme schwa /ə/ before the syllabic consonant (e.g. between the /d/ and the /nts/ in /stju:dnts/, between the /t/ and /r/ in /hɪstri/). The schwa is in fact the most frequently occurring vowel in English and is always associated with weak syllables (Roach 1991). It is recognised (e.g. Roach 1991) that certain processes of connected speech (e.g. elision, assimilation) produce additional syllabic consonants (e.g. /brəʊkŋ ki/ 'broken key' /ʌpmaust/ 'upper most'). However, as the representation of a word will be the basis of this study the processes of connected speech are not explored further. Each language has a range of permissible syllable and word structures. The permissible structures of Finnish and English are discussed in sections 2.3 and 2.6.

Each language has phonotactic rules which determine both the distribution of the phonemes within syllables and words (based upon position within syllable/word) and also neighbouring phonemes. Rules determine which consonants can co-occur as sequences, in which positions within words and syllables. English and Finnish phonotactic rules will further outline which singleton consonants and which consonant clusters can appear with which vowel nuclei and which particular sequences of phonemes are permissible at the various syllable positions. These are explored in Sections 2.3 and 2.6 respectively.

In theory, a language might make use of any combination of phonemes other than those that would be constrained because of human production limitations. Any of the vowel phonemes of a language could be used as a nucleus with any of the consonant, or permissible consonant clusters for that language. In reality, however, each language utilises only a sub-set from all the possible combinations. Research (e.g. Maddieson 1984) has found little relationship between segmental inventory size and syllable inventory size the possibility of there being correlations within the systems in terms of particular features co-occurring (as with phoneme inventories) has been suggested.

The phonotactic rules of each of the two languages will determine what particular phonemes might appear at which position within syllables and what types of syllables can appear at which position within words. Word final sequences of consonants are numerous as the affixation feature of English syntax means that words have phonemes added (e.g. to indicate plural or possessiveness and to indicate past tense). Finnish has a system of vowel harmony where the occurrence of a particular type of vowel will pre-empt the types of vowel phonemes that co-exist in the word. An overview analysis of the types of syllable and word formats will be undertaken for each of the two languages (Sections 2.3 and 2.6) in order to explore which phonemes occur at each word and syllable position.

The final sections of this chapter provide some statistical information regarding the frequency of the various phenomena so far highlighted for English and Finnish respectively. Whilst the phonotactic rules of a language determine the permissible sequences of phonemes not all the sequences are necessarily utilised to the same extent

in words of the language and words themselves are used with different frequencies. In reality despite a high number of permissible words, which would appear perfectly possible to an English speaker, the actual number of words seen in the total lexicon of English is much less and even less again when the words typically and most frequently used by speakers of the language is assessed. Languages also have homophony (two words that sound the same i.e. have the same sequence of phonemes) but have different meanings which further reduces the range of strings of phonemic usage observed.

For English the study of the frequency of use of specific phonemes, syllable structures, words and word structures is well established (e.g. Dewey 1923, Wang & Crawford 1960, Mines, Hanson and Shoup 1978, Knowles 1987). Whilst the availability of computerised corpora has meant that rapid and comprehensive analysis of very large amounts of data can now be undertaken relatively easily and consistently (Kennedy 1998) this sort of information is still more scarce for non Indo-European languages such as Finnish.

A number of similarities have been observed in the various English data sources in terms of the phonemes, words and word and syllable structures most frequently observed. Early work on word counting (e.g. Zipf 1935, Dewey 1923) pointed to the relative stability of the frequency distribution of words in discourses. For example, the 50 most commonly used word types in spoken language tend to make up about 60% of the word tokens (Miller 1963). The distribution of common patterns found when analysing linguistic performance, has led to the development of statistical models which aim by predicting the distributions of word types, type to token ratios etc. to support speech recognition and processing work. Whilst not of direct relevance for this research, which will to a large extent be determined by availability of already collected data, these models aim to reflect the accepted fact that in language words do not appear totally at random. Baayen (1993) completed a review of some of these but concluded that at this stage 'no model can lay claim to exclusive validity'. Whilst mentioned for the purposes of this introduction to frequency based performance analysis no models are presented further than a brief overview during the statistical presentations of frequencies where they support this study itself. Many of these findings still need testing cross-linguistically anyway before their suitability could be assessed.

The frequency statistics, based upon the words most likely to be used by speakers of a language, give a clear indication of the likely structures that this study needs to cater for in the data processing. The frequencies of words in spoken language, the frequencies of phones within spoken words and the types of syllable structures and their frequencies, are to be explored such that the relationship between frequency in adult and child languages can be assessed.

2.2 The Phonemic System of English

This section describes the inventory of the phonemes of English.

The English language utilises a total of 44 or 49 different phonemes depending upon which method of classification is chosen. Roach (1991), for example, includes the triphthongs in his classification and defines the system as containing 24 consonant phonemes and 25 vowel phonemes. Gimson (1994), on the other hand, does not include triphthongs and thus recognises the 24 consonant phonemes but only 20 vowels. From a functional point of view each of the 49 phonemes defined by Roach (1991) can be observed in minimal pairs serving to distinguish between lexical items in English and so it is this classification that will be used for this study.

2.2.1 English Consonant Phonemes

The English language utilises 22 'singleton' consonant phonemes (phonemes which are represented by one transcription symbol).

These are;

/ b, p, d, t, g, k, v, f, ð, θ, z, s, ʒ, ʃ, h, m, n, ŋ, l, w, r and j /.

In addition to these there are two affricate phonemes (represented by two transcription symbols);

/ tʃ and dʒ /

Whilst affricates are often described as complex consonants, beginning as plosives and ending as fricatives, they are accepted as operating as one phoneme within the system (Roach 1991).

That each of these 24 consonant phonemes carries a contrastive ability within the language is demonstrated in the following examples of English minimal pairs. The following examples demonstrate the singleton consonant phonemes in use at word initial position and group the English words by minimal pairs such that the contrastive abilities of the phonemes can be recognised;

- /bɪt/, /pɪt/, /tɪt/, /kɪt/, /fɪt/, /sɪt/, /hɪt/, /nɪt/, /lɪt/, /wɪt/, /rɪt/
  ('bit', 'pit', 'tit', 'kit', 'fit', 'sit', 'hit', 'knit', 'lit', 'wit', 'writ').
- /tɪp/, /dɪp/, /pɪp/, /sɪp/, /zɪp/, /ʃɪp/, /hɪp/, /nɪp/, /lɪp/, /wɪp/, /rɪp/, /tʃɪp/.
  ('tip', 'dip', 'pip', 'sip', 'zip', 'ship', 'hip', 'nip', 'lip', 'whip', 'rip', 'chip').
- /cæp/, /gæp/, /tæp/, /bæp/, /sæp/, /zæp/, /mæp/, /næp/, /læp/, /ræp/
  ('cap', 'gap', 'tap', 'bap', 'sap', 'zap', 'map', 'nap', 'lap', 'rap'_
- /bɒb/, /fɒb/, /sɒb/, /hɒb/, /mɒb/, /nɒb/, /lɒb/, /rɒb/, /jɒb/, /dʒɒb/.
  ('bob', 'fob', 'sob', 'hob', 'mob', 'knob', 'lob', 'rob', 'yob', 'job').

The following examples demonstrate some word final position phoneme contrasts;

- /etʃ/, /edʒ/, /eg/, /et/, /eb/
  ('etch', 'edge', 'egg', 'ate', 'ebb').
- /rɪp/ and /rɪb/, /rɪt/, /rɪk/, /rɪm/, /rɪŋ/.
  ('rip', 'rib', 'writ', 'rick', 'rim', 'ring').

The following examples demonstrate some within word position phoneme contrasts;

- /regim/ and /redim/ ('regime' and 'redeem').
- /meʒə/, /metə/, /mekka/ ('measure' and 'meta', 'mekka').

As has been discussed above consonants can be described phonetically in terms of their manner and place of articulation and whether they are voiced or voiceless. Appendix 2.1 provides the IPA representation of the English consonant phonemes together with the phonetic descriptions of place, manner of articulation and degree of voicing. Appendix 2.2 provides further examples of the phonemes in use within English words and more fully expands the phonotactic possibilities of each of the phonemes.

2.2.2 English Vowel Phonemes

The English phonemic system utilises a range of different types of vowel phonemes including pure vowels, diphthongs and triphthongs (Roach 1991).

Twenty five different vowel phonemes can be said to differentiate between words of the English language. Twelve of these are pure vowel phonemes, eight are diphthongs and five are triphthongs (Roach 1991).

2.2.2.1 English Pure Vowel Phonemes

The twelve pure vowel phonemes are described as being of relatively long or short length.

- the 6 short vowels are / ɪ, e, æ, ʌ, ɒ, ʊ /.

As demonstrated by the vowel sounds in the centre of the following English words;

/bɪt/ 'bit', /bet/ 'bet', /bæt/ 'bat', /bʌt/'but', /kɒt/ 'cot', /kʊd/ 'could'.

- the 5 long vowels are / i, 3, ɑ, ɔ, u /

(which alternatively could be written as / i:, 3:, ɑ :, ɔ :, u: /).

As demonstrated in the English words;

/hit/ 'heat', /h3t/'hurt', /hɑt/ 'heart', /mɔ/ 'more', /hu/ 'who'.

- the schwa vowel /ə/ as in the first vowel sound of the English word /əbaʊt/ 'about' and the last vowel sound of the word /betə/ 'better'.

Appendix 2.3 provides the descriptive parameters for the vowel phonemes of English using the common vowel descriptors of front, central or back, open, mid or close as already defined above in Section 2.1. Appendix 2.3 also provides the phonemic transcription codes that will be used from here on when presenting the English vowel phonemes.

2.2.2.2 English Diphthongs & Triphthongs

A further 13 vowel sounds are seen in the form of diphthongs and triphthongs.

RP English is generally accepted to have 8 diphthongs; vowel sounds which move from one vowel towards another (Roach 1991). Three of these are described as centring diphthongs as they glide towards the central schwa /ə/ vowel. These diphthongs can be seen as the vowel phonemes in the English words /hɪə/'here', /heə/ 'hair', and /ʃʊə/ 'sure'.

The remaining five are regarded as closing diphthongs (i.e. they glide towards a vowel with a closer distance between the upper surface of the tongue and the palate). In English three closing diphthongs glide towards the /ɪ/ vowel these are /eɪ, aɪ, ɔɪ/ as seen in the English words /meɪ/ 'may', /maɪ/ 'my' and /bɔɪ/ 'boy', and two glide towards the /ʊ/ vowel, these are /əʊ, aʊ/ as in the English words /nəʊ/ 'no' and /naʊ/ 'now'.

Triphthongs are vowel sounds which move from one vowel towards another and then on again towards a third vowel sound. Roach (1991) identifies a total of five

triphthong sounds as seen in RP English, these being the five closing diphthongs described above each with the addition of /ə/; /eɪə, aɪə, ɔɪə, eʊə, and aʊə/. These triphthongs are demonstrated in the English words;

/pleɪə/ 'player', /waɪə/ 'wire', /rɔɪəl/ 'royal', /sleʊə/ 'slower' and /aʊə/'our'.

Gimson (1994) does not include these 5 triphthongs in his original inventory of 44 phonemes (see above) as he feels that in actual pronunciation part of the triphthong, usually the second element, is omitted. He therefore classifies these as diphthongs + [ə]. For the purposes of this research it will be necessary to differentiate between the diphthongs and triphthongs such that the contrasts that they serve to make within the language are maintained. For example, the words 'why' and 'wire', written phonemically / waɪ/ and /waɪə/ differ only in respect of their use of a diphthong or a triphthong.

Further examples of words utilising these diphthong and triphthong vowel phonemes are given in Appendix 2.4 together with the full phonemic representation of the phonemic units of diphthongs and triphthongs.

2.3 English Phonotactics

2.3.1 Word and Syllable Structure in English

As has already been described above, English makes use of a variety of different phoneme types within its inventory; consonant phonemes of both singleton consonants and affricates (represented 'C') and vowel phonemes of pure vowels, diphthongs and triphthongs (all represented as 'V').

The English syllable can be said to consist of four structures (Lass 1984);

V - the vowel nucleus alone, the minimum component of a syllable.
(e.g. /ɑ/ 'are' and /ɔ/ 'or')

CV - the nucleus with a syllable onset of a consonant phoneme.

(e.g. /kɑ/ 'car' and /tʃɔ/ 'chore')

VC - the nucleus with a coda of a consonant phoneme.

(e.g. /ɑt/ 'art' and /ɔt/ 'ought')

CVC - the nucleus with both an onset and coda.

(e.g. /kɑt/ 'cart' and '/kɔt/ 'caught'.

However, in addition to these syllable structures, the English language permits sequences of consonants (known as consonant clusters) of up to three consonant phonemes in syllable initial position and up to four consonant phonemes in syllable final position (Roach 1991). For example, a vowel nucleus with a two consonant cluster in syllable initial position (represented CCV) is seen in the English word /slaʊ/ 'slow' and a vowel nucleus with a three consonant cluster in syllable initial position (represented CCCV) is seen in the English word /spreɪ/ 'spray'. An example of a vowel followed by a two consonant cluster (VCC) is the English word /ɑsk/ 'ask', a vowel followed by a three consonant cluster (VCCC) is /ɑsks/ 'asks'. An example of a four consonant cluster syllable ending is /teksts/ 'texts', representing CVCCCC.

The permissible syllable structures in English are 19; the nucleus (V) alone i.e. without an onset or coda, the nucleus (V) with a syllable onset which may be one singleton consonant (C), a two consonant cluster (CC) or a three consonant cluster (CCC) and the nucleus with a coda of one singleton consonant (C), a two consonant cluster (CC), a three consonant cluster (CCC) or a four consonant cluster (CCCC).

The structure of English syllables can thus be presented as ;

V	as in 'a' /ə/ and 'I' /aɪ/
CV	as in 'my' /maɪ/
CCV	as in 'try' /traɪ/
CCCV	as in 'spray' /spreɪ/

VC	as in ‘at’ /æt/
VCC	as in ‘apt’ /æpt/
VCCC	as in ‘axed’ /ækst/
VCCCC	as in ‘elves’ /elfθs/
CVC	as in ‘cat’ /kæt/
CCVC	as in ‘flat’ /flæt/
CCCVC	as in ‘sprat’ /spræt/
CVCC	as in ‘bank’ /bæŋk/
CVCCC	as in ‘banks’ /bæŋks/
CVCCCC	as in ‘texts’ /teksts/
CCVCC	as in ‘flits’ /flɪts/
CCVCCC	as in ‘fifths’ /fɪfθs/
CCVCCCC	as in ‘twelfths’ /twelfθs/
CCCVCCC	as in ‘strands’ /strænds/
CCCVCCCC	

Consonant clusters are sometimes represented simply as C2, C3, C4. Giving, for example, C3VC4 for the structure shown above as CCCVCCCC.

English word structures are combinations of the syllable structures shown above. Words theoretically could be any combinations of these syllables, thus words of two syllables may be any of the 19 syllable structures followed once again by any of the syllable structures, 19 x 19, that is 361 different word structure types. This figure increases exponentially with the number of syllables in a word; a word of five syllables could theoretically have 19x19x19x19x19 different structures. Whilst it is possible for English words and syllables to have this range of structures English words may not actually use all the possibilities. Section 2.4, English Frequency Statistics, presents the typical structures seen in spoken language.

The list above provides a representation of the possible syllable structures. It will now be necessary to assess which actual phonemes from the English phonemic system are acceptable at the various syllable positions and word positions.

2.3.2 Phonemic Content of The English Syllable

As described above English makes use of a total 49 phonemes. English does not, however, fully exploit all the possible combinations of its phonemes as will be explored in this section. For example, the phonemes /e, æ, ɑ, ʌ/ do not occur finally and the phoneme /ŋ/ does not occur initially. Long vowels and diphthongs do not precede final /ŋ/ and the types of consonant cluster permitted are subject to constraints (Gimson 1994).

2.3.2.1 English Consonant Clusters

The total of 49 phonemes can be further supplemented by the consonant clusters, sequences of 2, 3 or 4 consonants appearing together at the beginnings or ends of syllables.

As can be seen from the above, the English phonemic system consists of a total of 24 consonant phonemes. If there were no restrictions on the permissible sequences of consonant phonemes seen together within words and if it was physically possible to produce each of them in combination, then a range of 576 (24 x 24) two consonant clusters (CC) and 13,829 (24x24x24) three consonant clusters (CCC) would be observed in the words of English. Each language, however makes use of only a subset of those that it is possible for humans to articulate. For example, in English no clusters are possible that include the phonemes /tʃ, dʒ, ð, z/ and /r, j, w/ can occur in clusters only as the non-initial element (Gimson 1994). Just as each language has its own permissible sequences of consonants, certain clusters can only occur in syllable initial or syllable final positions (Roach 1991). For example, the consonant clusters /fs, stl, mh, spw/ are unknown as initial sequences (Gimson 1994). The consonant

clusters that appear at syllable initial and syllable final positions follow strict rules which govern the presence and positioning within the cluster of certain phonemes.

Clusters of consonant phonemes at syllable initial or syllable final position serve in much the same way as singleton consonant phonemes to differentiate between words. From a functional point of view singleton consonants and consonant clusters can be observed in minimal pairs providing contrasts with both singleton phonemes and also with other consonant clusters (e.g. /pæt/ 'pat' and /plæt/ 'plait', /sɪp/ 'sip' and /skɪp/ 'skip', /slɪp/ 'slip' or /snɪp/ 'snip', /sent/ 'sent' and /spent/ 'spent', '/klæp/ 'clap' and /slæp/ 'slap' etc.). The aim of this research is to assess the functional use that English makes of the phonemes available to it. The word initial consonant clusters explored here will be added to the inventory of phonemes that will be used as the basis for the current study, particularly in the calculation of functional use.

Other sequences of two or more consonant phonemes might also be seen adjacent to each other within words at a syllable boundary e.g. 'clockwork' /klɒk,wɜk/, 'likewise' /laɪk,waɪz/, however, these are not considered as consonant clusters as they are divided by a syllable boundary. The knowledge of the permissible syllable initial and syllable final consonant clusters will often therefore provide a guide to syllable boundaries (Gimson 1994). Consonant sequences that are found within words that are not one of the consonant clusters listed below can be said to be sequences and not clusters. The implications for the various ways of handling consonant clusters for this study are provided in Chapter 6.

The syllable initial two consonant clusters (CC) permissible in English (Roach 1991) are;

/sp, st, sk, sf, sm, sn, sl, sw, sj, pl, bl, kl, gl, fl, pr, br, tr, dr, kr, gr, fr, θr, ʃr, tw, dw, kw, θw, gw, pj, bj, tj, dj, kj, fj, vj, hj, mj, nj, lj/.

As in /spik/ ‘speak’, /steɪbl/ ‘stable’, /skɔlə/ ‘scholar’, /sfɪə/ ‘sphere’, /smɑt/ ’smart’, /snəʊ/ ‘snow’, /slaɪt/ ‘slight’, /swit/ ‘sweet’, /ensjʊ/ ‘ensue’ , /plɑɪ/’play’, /blɒk/ ‘block’, /klɪə/ ’clear’, /glɑs/ ’glass’, /flɔ/ ‘floor’, /prɪsaɪs/ ’precise’, /brænd/ ’brand’, /trʌst/ ‘trust’, /drɑft/ ‘draught, /kraʊd/ ‘crowd’, /grum/ groom, /frɒm/ ‘from’, /θri/ ‘three’, /ʃrɪŋk/ ‘shrink’, /twentɪ/ ‘twenty’, /dwel/ ‘dwell’, /θwɔt/ ‘thwart’, /gwen/ ‘Gwen’, /pjʊə/ ‘pure’, /bjutɪ/ ‘beauty’, /tjun/ ‘tune’, /djutɪ/ ‘duty’, /qwaɪət/ ‘quiet’, /fju/ ‘few’, /vju/ ‘view’, /hjumə/ ‘humour’, /mjuzɪk/ ‘music’, /njuz/ ‘news’ and /lʊə/ ‘lure’.

The syllable initial three consonant clusters (CCC) permissible in English (Roach 1991) are;

/spl, spr, spj, str, stj, skl, skr, skw, skj/

As seen for example in the words /splendɪd/ ‘splendid’, /sprɪŋ/ ‘spring’, /spjum/ ‘spume’, /streŋθ/ ‘strength’, /stu/ ‘stew’, /sklərəʊsɪs/ ‘sclerosis’, /skrɪpt/ ‘script’, /skweə/ ‘square’, and /skewer/‘skewer’.

Whilst this range of word initial structures is permissible the actual usage of the various word initial consonant cluster possibilities varies. Some sequences are found in only a very few words (e.g. /smj/ in ‘smew’, /gj/ as in ‘gules’), some are exemplified only by their use in certain proper names (e.g. /gw/ as in ‘gwen) and some only occur in recently imported foreign words (e.g. /ʃn/ as in ‘schnapps’ or /ʃw/ in ‘schweppes’) (Gimson 1994). Some sequences only occur in words rarely used by speakers of English as will be discussed in Section 2.4(below).

The syllable final two consonant clusters permissible in English are;

/ŋk, ŋs, sk, lt, lz, nt, nz, mp, mz, ps, bz, ts, dz, ks, gz,
fs, vz, θs, pt, bd, kt, gd, ft, st, ʃt, nt, lt, ld, wd, rt and rd/

As seen, for example, in the words;

/θæŋk/ 'thank', /rɪŋs/ 'rings', /ɑsk/ 'ask', /hɪlt/ 'hilt', /belz/ 'bells', /ænt/ 'ant', /mænz/'mans', /hʌmp/'hump', /emz/'aims', /tɒps/'tops', /ləʊbz/'lobes', /kæts/'cats', , /rɒdz/'rods', /æks/'axe', /tægz/'tags', /laɪfs/'life's', /ləʊvz/ 'loaves', /slɒθs/ 'sloths', /æpt/'apt', /ləʊbd/ 'lobed', /ækt/'act', /bægd/'bagged', /lɑft/'laughed', /lɑst/'last', /læʃt/ 'lashed', /kɑnt/ 'can't', /tɪlt/'tilt', /hɜld/'hurled', /əʊwd/ 'owed', /ɑrt/'art', /bɔd/ 'board',

The word final three consonant clusters permissible in English are;

/ lpt, ŋks, ndz, lfθ, fθs, kst, pst, lfθs, mpts, ksθs, ksts /

As seen, for example, in the words;

/helpt/ 'helped', /θæŋks/ 'thanks', /endz/ 'ends', /twelfθ/'twelfth', /fɪfθs/ 'fifths', /nekst/ 'next', /læpst/ 'lapsed', /twelfθs/ 'twelfths', /prɒmpts/ 'prompts', /sɪksθs/ 'sixths' and /teksts/ 'texts'.

Gimson (1994) identifies several problems in defining a list of all permissible final consonant clusters. Sometimes a word has more than once accepted pronunciation (e.g. /tθ/ or /dθ/ at the end of 'hundredth', /ntʃ/ or /nʃ/ at the end of 'French') or has a rare sequence (e.g. /ns/ 'prince', /nts/ 'prints'). With the standard pronunciation of RP these possible variations of sequences at word final position should be standardly represented. Also, the syllabic consonants discussed above might appear as part of clusters. However again for the purposes of this study they are assumed to be divided by the schwa /ə/ phoneme (see above).

Gimson (1994) accounts for the greater complexity of final consonant clusters by the fact that final /t, d, s, z/ frequently represent a suffixed morpheme (e.g. past-tense, possessive). The syntactic rules of English mean that certain sequences of consonants are therefore frequently seen at word final position. English utilises suffixation to indicate plural, adding the /s/ phoneme after a voiceless consonant and a /z/ after voiced consonants e.g. 'cats' and 'dogs', /kæts/ but /dɒgz/. English forms the past tense by adding the /t/ phoneme after a voiceless consonant, or the voiced /d/ phoneme after a voiced consonant (e.g. /wɔlkt and bægd/ 'walked' and 'bagged').

2.3.3 Phonemic Content of English Words

The minimum component of an English word is a single syllable by itself. As has been discussed above the minimum component of a syllable is the nucleus (vowel phoneme) in isolation. The minimum component of an English syllable would therefore be one of the twenty five vowel phonemes identified in 2.2 above.

Gimson (1994) identifies ten vowels which are recognised as monosyllabic words;

/ə/ 'a', /ɑ/ 'are', /ɔ/ 'or', /ɜ/ 'err', /eɪ/ 'A', /aɪ/ 'I/eye',
/əʊ/ 'owe/O', /ɪə/ 'ear' and /eə/ 'air', /i/ 'E'.

A further two are identified as shortened forms of words generally used by speakers /e/ for 'he' and /u/ for 'who' and one colloquial form /ɔɪ/ for the exclamation 'oy!'. As Gimson does not include triphthongs in his inventory of English phonemes (see above) the word /aʊə/ 'hour/our' is missing from the list.

All vowels occur initially however word initial /ʊ/ and /ʊə/ only occur in foreign proper names such as 'Uppsala' and 'Urdu' (Gimson 1994).

The syllable onset may be any of the word initial consonant clusters described above and any of the singleton consonants described in 2.2 except / ŋ /(Roach 1991). Whilst

the onset consonant /ʒ/ is permissible in word initial position it only occurs in loan words initially before /ɪ/, /i/, /æ/ or /ɑ/ (Gimson 1994).

A total of seventy one possible onsets are therefore permissible at the word initial position in English (i.e. 23 singleton consonants, 39 word initial 2 consonant clusters, nine word initial 3 consonant clusters).

Any consonant may be a syllable final consonant except /h, r, w and j/ (Roach 1991). Using the phonemic inventory provided in section 2.2 and the word final consonant clusters listed above a total of sixty eight different syllable codas may be present at word final position for English (20 singleton consonants, 37 word final 2 consonant clusters, seven word final three consonant clusters and 4 word final 4 consonant clusters).

2.3.4 The English Lexicon

This study takes as its starting point the permissible sequences of phonemes of English words.

Simply taking into account the permissible syllable and word structures described above a figure of 113,000 different one syllable word possibilities alone might be expected to be found in the English lexicon (see Appendix 2.5). It is recognised however that not all permissible structures will actually be used by English speakers. In reality despite this high number of possible words, words which would be permissible and would appear possible to a native English speaker, the actual number of words typically used by English speakers is much less than this figure. In order to build the processing for FUSE based around permitted possibilities of sequences of phonemes it firstly will be necessary to abstract from the possible words firstly those sequences of phonemes that are actually recognised as words of the language. In order to identify those words that are most frequently spoken by speakers a further assessment based upon usage will need to be completed.

Estimates on the average lexicon of an educated adult English speaker mention word type ranges of typically 50,000 words rising to 60,000 or even possibly 80,000 if all proper names and idiomatic expressions are included (Aitchison 1994). This figure includes words that speakers understand and could potentially produce. However, how many of these words (and word structures) actually occur in spoken language can be estimated to be much less and even less again when only the words frequently and consistently used in the language of many people, are taken into account. Also, despite this huge amount of possible phoneme sequences to provide unique word forms English, as discussed above, still demonstrates a certain amount of naturally occurring homophony e.g. /dɪə/ for 'dear and deer', /daɪ/ for 'dye, di and die', /peə/ for 'pair and pear' and /pɔ/ for 'pour, pore and poor'.

Sections 2.2 and 2.3 above describe the components of the English phonological system and the phonotactic rules of English word and syllable formation in order to define the permissible components of English words. Section 2.4 discusses the actual use that the English language makes of these possibilities that is how frequently the spoken language makes use of these various possibilities in the words of English. Summarising statistically the most common words, syllable structures and phonemes of spoken English will enable a picture of the nature of ambient language to be completed.

2.4 English Frequency Statistics

A summary of the main findings for English phoneme frequencies, word types and syllable and word formats are now presented together with an outline of some the issues in ascertaining frequency data due to the variations in approach that have been adopted.

2.4.1 Phoneme Frequency

It is a well cited fact that certain phonemes are observed more frequently in spoken English than others. As early as 1963 Miller stated that 'nine phonemes make up more than half of our vocal behaviour and the most frequently used sound occurs more than

100 times as often as the least used sound'. Knowledge of the estimates of a particular phoneme's relative frequency in spoken English enables us to assess the likely frequency of a phoneme in the ambient language surrounding children acquiring English. A comparison between these findings and the actual frequency of the phoneme in the data selected for processing in this study provides a base line measure for assessing the suitability of the corpora that has been selected.

There are several problems, however, when comparing phonemic frequency figures and care needs to be taken that similar types of data are being compared. Corpora of written language, have been shown to have different distributions (e.g. more past tense use, less use of the pronoun 'I' etc.) than spoken language corpora (Kennedy 1998) all of which would affect frequency. Some frequency analysis is based upon word tokens and others on word types (see below). With the frequency of a phoneme based upon running text and word tokens, a phoneme would be counted each time a word containing that phoneme occurs and the word frequency would have a direct impact upon the phoneme's frequency count. In addition to this, in order for the various phoneme frequencies to be directly compared the same phonemic inventory must be adopted. For example, this study recognises word initial consonant clusters as carrying phonemic status and operating as units together within the phonemic system. In order to compare the findings of this study with existing frequency findings cluster frequencies must be counted such that each cluster has been counted in its entirety and not with each individual consonant component being counted separately.

The variety of spoken English is also relevant. For example, Northern British English might be expected to display a higher occurrence of the phoneme /ʊ/ as this phoneme is often used instead of /ʌ/ and American English has a different vowel system before /r/ etc. (Wells 1982). As has been already discussed the RP pronunciation of English words will be assumed for this study and forms the basis of the phonemic inventory described above. Utilising phoneme frequency reports for RP English, as opposed to another accent, will ensure that like phonemic systems are being compared.

One early study completed by Dewey in 1923 assessed the overall frequencies of phonemes in standard English prose. Recognising a total of forty two phonemes it was found that the most frequent phonemes, in order frequency, were /ɪ, n, t, r, ə, s, d, æ, l, 3/. Consonant phonemes accounted for 62% of the total frequency and vowel phonemes the remaining 38%.

Knowles (1987) completed an assessment of phonemic frequency based on a sample of 10,000 phonemes from both spoken and written texts. A phonemic notation system was used that recognised the 22 singleton consonants, 2 diphthongs, 12 pure vowels and 8 diphthongs (discussed above in Section 2.2). The research did not recognise or count triphthongs and consonant clusters as units instead counting each component of these phonemic units as a separate entity.

The following overall phonemic frequency ranking was found (most frequent first);

/ə, ɪ, n, t, s, d, l, r, ð, z, k, e, w, m, aɪ, b, p, g, v, æ, i, ɒ, əʊ,
eɪ, u, ʌ, ɔ, j, h, ŋ, g, ʃ, aʊ, dʒ, 3, θ, ɑ, tʃ, ʊ, ɪə, eə, ɔɪ, ʒ, ʊə/

Mines, Hanson and Shoup (1978) studied a corpus of 103,887 phoneme occurrences, taken from casual conversational American English and ranked the order of both consonant and vowel phonemes according to frequency of occurrence. Whilst utilising an inventory of 48 phonemes to represent the American phonemic system some British English diphthongs and no triphthongs were recognised. Consonants that were part of clusters were counted individually, as were the vowel phoneme components of triphthongs and some of the diphthongs. In addition, the Arpabet symbols utilised for their transcription were in some cases different than the standards used for this study (e.g. for the diphthong /əʊ/ the phoneme /o/ was given, /ɛ/ for /e/, and /e/ for /eɪ/). More importantly, certain phonemes were recognised that were not used for the system outlined above in Section 2.2 (e.g. the glottal stop /ʔ/ and the flap /ɾ/) and the syllabic consonants of /l, m and n/ were given separate phonemic status to the phonemes /l, m and n/ (rather than having the schwa vowel /ə/ inserted as a weak syllable).

Despite all of these differences in approach and pronunciation, this research provides one of the rare phoneme frequency reports that includes both consonants and vowel phonemes.

The following order of frequency was obtained for the phonemes (most frequent first);

/ə, n, t, ɪ, s, r, i, l, d, e, ð, k, m, ɑɪ, w, z, æ, b, o,
p, v, e, f, ʌ, ɑ, h, g, u, y, ŋ, ɔ, ʊ, θ, ɑʊ/.

In this study the most frequently produced ten phonemes /ə, n, t, ɪ, s, r, i, l, d, ɛ/ accounted for 47% of all the data. It was noted that most phonemes were spread throughout many different words although the high frequency of some of the phonemes was directly related to the high frequency of the words in which they appeared. For example, the schwa phoneme /ə/ which occurred most frequently in the data overall had 54% of its occurrences in realisations of only six different words. These six words came from the ten most frequent lexical items recorded ('the', 'uh', 'a', 'to', 'of' and 'was'). The phoneme /ð/ had 50% of its occurrences in repetitions of only two words, /ðɛ/ 'the' and /ðæt/ 'that', both of which are among the top ten most frequently found words. The phoneme /ɑɪ/ had 40% of its occurrences in the pronunciations of the pronoun 'I' which is the third most frequent word. This study provides a direct example of the impact of using running text for the phoneme frequency analysis, whereby each phoneme is counted when it appears in a word and the word's frequency thus has a direct impact on the phoneme frequency found.

Wang & Crawford (1960) recognised that there are problems with assessing 'unlike' corpora for frequency statistics and therefore completed research to normalise data on consonant phoneme frequency from many earlier studies. Their assessment included frequency findings produced by Trnka (1935), Fowler (1957), Carroll (1952), Hayden (1950), Whitney (1874), Dewey (1923), Voelker (1937), French (1930), Fry (1947) and Tobias (1959). These studies provided a variety of dialects from British and American English and various sample sources (i.e. dictionary, prose, newspapers, radio

announcements, telephone conversations). It was concluded that despite the various styles of data and dialects observed, these studies came up with a high degree of correlation of the most frequently produced consonant phonemes, these being

/t, n, r, s, d, l, ð, k, m, w, j, z, v, h, f, b, g, p, ŋ, ʃ, θ, ʒ/.

This study also showed that dictionary sampling based upon word types (where a word and the phonemes it contains are counted only once) in English yields statistical results which are very different from those obtained by sampling continuous speech where the word and the phonemes it contains are counted each time the word appears.

As it is the range of words that will be important for this study, it will provide phoneme frequency counts that are different to those outlined above in two respects. Firstly the phoneme frequency counts found in this study will be based upon word types rather than word tokens. Secondly, the words included for this study are those most frequently observed in spoken language rather than those that might appear in a dictionary but would only be rarely used by English speakers. It will be interesting to observe whether there are significant differences for British English between the phoneme frequencies found above based upon word tokens and those found in this study for most frequent word types.

For this study the function that a particular phoneme performs in the language, i.e. its systemic value, is of prime importance and position of the phoneme within the context of word is a more important frequency measure than overall frequency. As has been shown above the English system uses a wider variety of consonants at the beginning than at the end of words. Only 5 different phonemes are found to make up 50% of the final consonants whilst 8 make up 50% of the initial consonants (Miller 1963). Some consonant phonemes are used mainly in certain positions within words e.g. the phonemes /w, h, j, h, b, g, f, p, ð, θ/ are used mainly in word initial position whilst the phonemes /ŋ, z, v, r/ are used principally in word final position (Gimson 1994). This study will collect frequency information relating to word initial phoneme usage, the

definition of phoneme being both singleton phoneme and phonemic units such as consonant clusters.

2.4.2 Word Type Frequency

Research into the frequency of English word types in running text has found that certain words tend to be consistently utilised more frequently than others. Kennedy (1998) demonstrated a high level of similarity between the 50 most frequent words observed in different corpora by completing a comparison of the rank ordering of the most frequent words in various general corpora including the Birmingham, Brown, LOB and London-Lund Corpora (discussed in Chapter 4). When comparing the most frequent words, he found a "striking consistency" in the list of most frequent words despite differences in;

- the size of the corpora
- the fact that British, American and New Zealand English were represented
- the variation in the period of corpora production (from the 1960's to the 1980's)
- the fact that both spoken and written texts were included
- words of both children and adults.

For example, the Birmingham corpus which includes more than 1.5 million words of spoken English, compared closely with some of the other much smaller corpora and even with corpora of written language, corpora collected some 20 years earlier and text written for children and adults. The differences in the actual ranks were attributed to particular words reflecting the nature of the different corpora (e.g. a higher frequency of 'I' and a lower ranking of 'had and were' in the spoken corpora). He found, however, a degree of consistency between the lists of the most frequently produced words.

All of the corpora under assessment included the following 32 words within their 50 most frequent words;

'the, of, and, to, a, in, that, I, it, was, is, he, for, you, on, with, as, be, but, they, at, have, not, this, are, or, we, she, one, all, there, which'.

It is interesting to observe that all the words within this list are closed class function words of one syllable. With the individual word frequencies observed in the different corpora totalled together the fifty most frequent words overall, apart from the word form 'said,' are in fact function words. In fact the Birmingham Corpus displays only 12 non function words in the top 100 most frequently spoken words. Kennedy (1998) notes that after about the 50th frequency ranked word content words begin to predominate in all general corpora.

Another interesting aspect of this list of the most frequent English words is that many of them are homophones with other frequent words e.g. 'to' with 'two', 'I' with 'eye', 'for' with 'four', 'one' with 'won' and 'there' with 'their' and 'they're'.

Specialist corpora often contain higher frequencies of words within a particular genre or subject field (for example technical terms or names) and content words feature much earlier (i.e. with higher frequencies) in more narrowly focused corpora. Deviation from these words is therefore observed much earlier with specialist corpora.

When all but the smallest corpora of running text are analysed it is usual to find that a relatively small number of different words (word types) make up most of the actual word occurrences (word tokens). A corpora containing a large number of tokens, as represented by running text, may, when analysed for types, be found therefore to be much smaller in terms of the range of words presented (Gibbon, Moore & Winski 1997). The focus of this research will be the range of words available for analysis of minimal pairs and as such a large number of different word types will be sought.

Hapax legomena words (words appearing only once in a given corpus) usually account for many of the word types seen in corpora (Kennedy 1998). Despite differences in size of corpora being assessed Kennedy (1998) found that "almost 40% of the words in a corpus of over five million words occur only once". The statistical regularities of type to token ratios proposed by Zipf and cited by Shannon as early as 1949 (Crystal

1987) have more recently been demonstrated using larger modern corpora. This law predicts that a small number of words will be very common but a significant number will be used infrequently.

In order that an assessment of the relative frequencies of the words as found in different corpora can be made the rankings for words in four well established English language corpora (London Lund, LOB, Birmingham and Brown) are provided in Appendix 2.6 The main differences between the corpora and their suitability for use in this study is discussed in Chapter 4.

2.4.3 Word and Syllable Frequency

The studies above support earlier studies of word structure (e.g. Zipf 1935) in finding that the majority of the commonly used words in English are in fact monosyllables. It has been predicted that the word structures seen the most frequently in large volumes of spoken data will follow the general rule of maximal communicative intent with minimal effort (Tobin 1997). This rule predicts that the most frequently observed words will be short and made up of simple structures (e.g. one syllable words of V, CVC etc. structures).

One study which assessed the most frequently observed syllable structures based on the telephone speech of businessmen, showed that consonant-vowel-consonant(CVC) syllables comprised 33.5% of all the syllable tokens spoken, CV syllables made up 21.8%, VC syllables 20.3% and V 9.7% (French, Carter & Koenig 1930).

According to Miller (1963) the most frequently used English syllables are;

/ðə, ʌv, m, ænd, ɪ, ə, tu, ɪŋ, rɪ, ɪt, ðæt/

The most frequent syllable /ðə/ makes up about 7% of the syllables of spoken English. Half of our speech uses only 70 different syllables, but 1370 syllable types are required before 93.4% of English syllable tokens are included.

2.5 The Phonemic System of Finnish

This section provides an inventory of the phonemes of Finnish. The phonetic characteristics of the Finnish phonemes are described such that individual sounds can be observed and readily identified and also so that the range of phonetic attributes can be seen. Sequences of phones that are seen operating together as units within the system (as discussed above in 2.1) are also explored. For Finnish, this can be taken to include the long/double vowels and geminates.

Providing an inventory of the phonemes of a language might seem a straightforward task: The inventory is a list of those phonemes that can be seen to carry functional load (i.e. differentiate between words / appear in minimal pairs). The Finnish language, however, provides some interesting phenomena that make it more difficult to provide a finite and undisputed list. Two of these, the geminate/single consonant issue and the long/short vowel contrast are presented below together with the phoneme inventories. These issues need to be considered when compiling the inventory of phonemes that will be assessed for this research such that a decision can be made on how they are dealt with.

The Finnish system utilises a total of 21 different phonemes (Iivonen 1998): 13 are classed as consonants and 8 are classed as vowel phonemes. The eight vowel phonemes and eight of the consonant phonemes are observed as both long (double) and short (single) forms which distinguish between words of the language. With length included as a feature the total phonemic inventory of Finnish could be argued to increase to 29. In this section the various arguments for the handling of the long and short vowels, the single and double consonants and other sequences of phonemes such as diphthongs are explored.

In addition to the problems encountered in classifying certain phonemes, Karlsson (1983) states that the number of phonemes considered to be part of the Finnish inventory can vary from 19 to 25 depending on which of 5 classification systems or paradigms you choose.

System 1, called 'minimaalinen järjestelmä', the minimal system, contains 19 phonemes that are seen in all dialects of the Finnish language. According to this there are the vowel phonemes /i, y, e, œ, æ, u, o and ɑ/ and 11 consonant phonemes / p, t, k, s, h, m, n, l, r, j and v/.

System 2 has the phoneme /ŋ/ added to the minimal system given above. Whilst there are some dialects which only have this phoneme before /k/ (i.e. in sequences of /ŋk/) or in sequences resulting in assimilation this system of 20 phonemes is still the most common system seen in Finnish regardless of education and communicative situation etc. (Karlsson, 1983).

System 3 includes the loan phoneme /d/ in addition to the system of 20 phonemes given above. This system of 21 phonemes is described as the 'ydinjäjestelmäksi', core or basic phonemic system of Finnish. Although the phoneme /d/ still presents a lot of variation in its pronunciation depending upon dialect, education level and exposure to loan words, it is commonly utilised in most dialects.

System 4, also called the maximal basic system, contains the 21 phonemes already presented plus the phoneme /f/ which is common though it is only present in loan words.

System 5, 'maksimaalinen järjestelmä', the maximal system, adds the marginal phonemes /b, g and ʃ/ which occur in loan words to the 22 phonemes already presented.

All phonemes that occur in a spoken language corpora are to be included in this study. The phonemic system that will be described therefore relates to the maximal system, System 5, above.

2.5.1 Finnish Consonant Phonemes

2.5.1.1 Finnish Singleton Consonants

It is generally accepted that the Finnish system utilises 13 singleton consonant (Iivonen 1998, Sulkala & Karjalainen 1992).

These are;

/ p, t, d, k, s, h, m, n, ŋ, l, r, j and v/.

Each of these singleton phonemes carries a contrastive ability within the language as can be seen in the following examples of Finnish minimal pairs;

Suu 'mouth', *luu* 'bone', *muu* 'other', *kuu* 'the moon' and *juu* 'yes',
Talo 'house', *palo* 'fire', *salo* 'wilds', *valo* 'light' and *jalo* 'noble',
Sauma 'seam' and *lauma* 'herd',
Maton 'rugs' and *madon* 'worm's'
Kannas 'isthmus' and *kangas* 'fabric'.

This list can be seen to be the 13 consonant phonemes seen in system 3 above. However, it does not include the phonemes that have more recently entered the language in loan words. The maximal system of a total 17 consonant phonemes therefore includes the following four phonemes in the Finnish phonemic inventory;

/f, b, g, ʃ/

Loan words are frequently observed in Finnish spoken language, particularly loan words from Indo-European languages such as Swedish and English (Sulkala & Karjalainen 1992). Some of the loan words have introduced new sounds which are pronounced correctly by Finnish speakers as in the case of /f/ e.g. *filmi* 'film, *toffee*

'toffee', *sifoni* 'siphon' and *mikrofoni* 'microphone' (see above). Other sounds have had their pronunciation adapted to suit the Finnish phonological system or are simply substituted by Finnish phonemes; the phonemes /b, g and ʃ/ as seen in the words *bussi* 'bus', *golf* 'golf' and *shakki* 'chess' are more difficult for native Finnish speakers to pronounce and are often substituted by [p, k or s] respectively Karlsson (1983).

Appendix 2.7 provides a phonetic description with place and manner of articulation and voicing information for the Finnish consonant phonemes. It is noted that the degree of voicing by Finnish speakers is much less than the degree of voicing of the same target sound by English speakers.

2.5.1.2 Finnish Geminates

In addition to the single consonants described above Finnish uses a system of double (long) consonants or geminates (Karlsson 1983 and Iivonen 1998). Eight of the singleton consonants described above (/p, t, k, s, m, n, l and r/) can also appear as double consonants (phonetically long).

Geminates are written orthographically with two identical consonants. Finnish words that demonstrate the use of geminates include; *sil<u>l</u>oin* 'at that time', *La<u>pp</u>i* 'Lapland', *ko<u>kk</u>i* 'cook', *ka<u>ss</u>a* 'cash register', *ka<u>tt</u>o* 'roof', *tu<u>kk</u>a* 'hair' and *se<u>mm</u>oinen* 'that kind'.

One way of viewing these double consonants is as sequences of two identical consonant phonemes (Iivonen 1998 and Karlsson 1983). Thus, using the four examples from above, /kokki/ *ko<u>kk</u>i* 'cook', /kasa /*ka<u>ss</u>a* 'cash register', /katto/ *ka<u>tt</u>o* 'roof', /tukka/ *tu<u>kk</u>a* 'hair' each word would be regarded as having five phonemes. In phoneme frequency calculations the consonant phoneme would be counted twice in each word. This treatment of geminates, as sequences of two consonants split by a syllable boundary, is supported by the fact that the geminate consonant pairs can be seen to contrast with sequences of single consonants e.g. Lappi 'Lapland' vs. lapsi 'child', kokki 'cook' vs. koski 'rapids', kassa 'cash' vs. kansa 'people', katto 'roof' vs. katso 'look' and tukka 'hair' vs. tuhka 'ashes'.

Each of these geminate phonemes does, however, carry a contrastive ability not only with other geminates and two sequence consonants but also with the singleton consonants described above, as can be demonstrated in the following examples of Finnish minimal pairs;

/sepæ/ *sepä* 'it' (plus enclitic suffix) and /seppæ/ *seppä* 'smith',

/kɑto/ *kaṯo* 'to look' and /kɑtto/ *kaṯṯo* 'roof',

/kuka/ kuka 'who' and /kukka/ kukka 'flower',

/kɑsɑ/ *kaṣa* 'heap' and /kɑssɑ/ *kaṣṣa* 'cash register',

/kumi/ *kum̲i* 'rubber' and /kummi/ *kum̲m̲i* 'godparent',

/kɑna/ *kan̲a* 'chicken' and /kɑnna/ *kan̲n̲a* 'carry'(imperative),

/tili/ *tiḷi* 'salary' and /tilli/ *tiḷḷi* 'dill',

/vɑrɑs/ *vaṛas* 'thief' and /vɑrrɑs/ 'skewer'.

Arguments for the treatment of geminates as one phonemic unit (comprised of two identical consonant phones) which is a separate and different phoneme than the single consonant phoneme occurrence of the same phone are thus based around the evidence that single and geminate versions of the phonemes are contrastive. The Finnish minimal word pair kuka 'who' and kukka 'flower' differ only in respect of them having a single consonant or a geminate to provide the contrast. The argument for regarding geminates as one phonemic unit is also supported by the fact that geminates act like other single phoneme consonants and follow other singleton consonants to produce consonant sequences e.g. kan̲sa 'nation' and kan̲ṣ̲a 'with'. Treating the geminates as single phonemes, long consonants, rather than sequences of two identical phonemes, does however pose the problem of syllable boundary identification. With the example above kuk:a 'kukka' one phoneme, the geminate phoneme /k:/, appears to be ambisyllabic (i.e. to simultaneously exist within 2 syllables CV<u>C</u>V).

For this study a decision needs to be made on the handling of geminates either as phonemic units (comprised of two elements e.g. /kk/, /tt/ etc.) or alternatively as sequences of two identical consonant phonemes. With the first approach the geminate

/kk/, for example, would be treated as a phonemic unit which contrasts in minimal pairs with /k/ (e.g. /kʊkɑ/ vs /kʊkkɑ/)or any other singleton consonant phoneme as well as with all other geminates. With the second approach the sequence of two phonemes , /k/ and /k/ for example, would only contrast with other two consonant sequences where one consonant was identical (e.g. /ss/ to /sk/, /ts/ etc.) Each approach would provide not only different phonemic system inventories (e.g. including or excluding the phonemic unit /kk/) but different frequency assessments (one frequency count for /k/ and one for the phonemic unit /kk/ or one for all occurrences of /k/).

In order to fully test the contrastive abilities of the phonemes of the language it would be important to recognise that geminates may provide the contrast with both singleton consonants and consonant clusters as well as other geminates. As the focus of this initial study will be on the word initial phonemic usage of Finnish and geminates do not in fact appear in word initial position (see Section 2.6) they would not be found as word initial contrasts in any assessment. The processing and frequency assessments of geminates for this study is considered further in Chapter 6.

Unlike the English consonant phoneme to orthographic representation relationship where there is not a one to one relationship between phoneme and orthographic representation the majority of the Finnish consonant phonemes are expressed by a constant orthographic letter (e.g. /s/ is represented as 's', /t:/ as 'tt' etc.). The exception to the rule is the phoneme /ŋ/ which is represented in Finnish orthography as 'nk' and the long /ŋ/ as 'ng' (Sulkala & Karjalainen 1992).

2.5.2 Finnish Vowel Phonemes

2.5.2.1 Finnish Pure Vowel Phonemes

The Finnish language utilises 8 different vowel phonemes (Sulkala & Karjalainen 1992). These are;

/i, y, e, œ, æ, u, o and ɑ/.

These vowel phonemes can be demonstrated in the following Finnish words:

/sitte/ *sitte* 'so'

/kylla/ *kyllä* 'yes'

/ettæ/ *että* 'that'

/tyœ/ *työ* '*work* '

/mutta/ *mutta* 'but'

/oka/ *oka* '*thorn*'

Each of these vowel phonemes carries a contrastive ability within the language as can be seen in the following examples of Finnish minimal pairs:

/aseita/ *aseita* 'weapons' and /useita/ *useita* 'several'

/aletti:n/ *alettii*n 'we started' and /eletti:n/ *elettiin* 'we lived'

/osti/ *osti* 'bought' and /asti/ *asti* 'up until her/now'

/ylen/ *ylen* 'really' and /olen/ *olen* 'I am'

/ollat/ *ollat* 'to be' and /illat/ *illat* 'evenings'

/kæsi/ *käsi* 'hand' and /kosi/ *kosi* 'proposed'

/lœssi/ *lössi* 'a group' and /lossi/ *lossi* 'a ferry'

Five of these, /i, y, e, œ and æ/ are classed as front vowels and the remaining three /u, o and ɑ/ are classed as back vowels. The articulatory descriptions of these vowel phonemes are provided in Appendix 2.8 together with the phonemic transcription that will be used for this study. It is recognised that there is sometimes variation in the representation of these phonemes (e.g. Iivonen 1998 represents /ɑ/ as /a/ and /œ/ as /ø/). However, the representation given in the list of vowels above and in appendix 2.8 will be followed throughout this study.

Finnish phonology is mainly phonematic (i.e. there is a one to one relationship between the orthographic representation of graphemes and phonemes). The vowel phoneme /i/ is represented as the grapheme 'i', the phoneme /e/ as 'e', the phoneme /ɑ/ for 'a' etc. The phonemes /œ/ and /æ/ are represented orthographically as 'ö' and 'ä' respectively.

Each of the 8 Finnish vowel phonemes given above also has a phonetically longer version sometimes known as a double or long vowels (Iivonen 1998). Thus the original single vowel phonemes described above also have a long vowel equivalent; /i:, y:, e:, œ:, æ:, u:, o: and ɑ:/ (orthographically represented as 'ii, yy, ee, öö, ää, uu, oo, and aa' respectively).

These long vowel phonemes are demonstrated in the following Finnish words;

/hi:ri/ *hiiri* 'mouse',

/ɑlennusmy:nti/ *alennusmyynti* 'sales',

/ɑnte:ksi/ *anteeksi* 'sorry',

/insinœ:ri/ *insinööri* 'engineer',

/elæ:/ *elää* 'to live',

/tu:li/ *tuuli* 'wind',

/*jo*:/ *joo* 'yes', and

/ɑ:mʊ/ *aamu* 'morning'.

The durational ratio between single and double vowels depends to a large extent upon syllable position and phonotactic factors but tends to be in the region of approximately 1:2 and 1:2.6 for first syllable occurrences and 1:1.4 for later syllable occurrences (Iivonen 1998).

There has been much discussion (Harrikari 1998, Karlsson 1983, Iivonen 1999, Maddieson 1984) about the status of long vowels and whether the long and short versions of the same vowel sounds should in fact be treated as separate phonemic

entities within the phonemic system. Some descriptions regard long vowels as sequences of two identical short vowel elements in Finnish (e.g. Karlsson 1983). Some studies have therefore counted long vowels as two vowel phoneme entries, sequences of two identical vowels, for frequency analysis purposes making no distinction between long vowels, and short vowels (e.g. Hakulinen & Leino 1983, Pääkkönen 1973, Häkkinen 1983).

The two versions of the vowels, short and long, are, however, contrastive within the Finnish language as can be demonstrated in the following examples of Finnish minimal pairs;

/sinæ/ *sinä* 'you' and /si:næ/ *siinä* 'there',

/ryppy/ *ryppy* 'wrinkle' and /ry:ppy/ *ryyppy* 'drink/swig'

/te/ *te* 'you' and /te:/ *tee* 'tea'

/tællæ/ *tällä tavalla* 'in this way' and /tæ:llæ/ *täällä* 'over here'.

/tuli/ *tuli* 'fire' and /tu:li/ *tuuli* 'wind'

/jo/ *jo* 'already' and /jo:/ *joo* 'yes'

/varat/ *varat* 'means/funds' and /va:rat/ *vaarat* 'dangers'.

Linguists today (e.g. Maddieson 1984, Iivonen 1999) generally cite the two vowel types, long and short, as two different vowel phonemes thereby utilising length as a separate phonological feature and it is this approach that will be taken for this research.

Following this approach Finnish can be shown to contain a total of 16 contrasting vowel phonemes. Appendix 2.8 gives a phonetic description of the 16 Finnish vowel phonemes, recognising and utilising length as a phonemic feature. The orthographic representation of the vowel phonemes, that is how they are seen in written texts, is also provided.

2.5.2.2 Finnish Diphthongs

Sulkala & Karjalainen (1992) list 17 genuine diphthongs in Finnish. Fifteen diphthongs that glide towards a closer vowel;

/ɑi, ei, oi, ui, yi, æi, œi, ɑu, ou, eu, iu, æy, œy and ey/.

These can be seen in the following examples of Finnish words;

/ɑita/ *aita* 'rail'

/ei/ *ei* 'no'

/oikeɑ/ *oikea* 'right'

/uidɑ/ *uida* 'swim'

/hyi/ *hyi* 'yuk'

/æi/ *äiti* 'mother'

/œinen/ *öinen* 'nocturnal'

/ɑuto/ *auto* 'car'

/oulu/ *Oulu* 'Oulu'

/euro:ppɑ/ *Eurooppa* 'Europe'

/tiukkɑ/ *tiukka* 'tight'

/æyskiæ/ *äyskiä* 'bark'

/œylætti/ *öylätti* 'wafer'

/leyhytellæ/ *leyhytella* 'to fan'

It can be noted that Karlsson, 1983, recognises 18 diphthongs and includes the diphthong /iy/ to the list given above.

The three remaining diphthongs glide towards a more open vowel;

/yœ, ie, and uo/.

These can be seen in the Finnish words;

/yœ/ *yö* 'night',

/kiero/ *kie̲ro* 'cunning'

/uoma/ *uoma* 'riverbed'.

2.6 Finnish Phonotactics

In Section 2.5 an inventory of Finnish phonemes was explored. This study, however, needs to explore not only the phonemes utilised by the Finnish language but also the rules for combing the phonemes into words (i.e. phonotactic rules). The permissible combinations and the sequences of phonemes together with distributional restrictions based upon position within syllable and word are again different for each language. The permitted structures of Finnish words and syllables together with the permitted permutations and patterns of phonemes are explored in this section.

2.6.1 Word and Syllable Structure in Finnish

The minimum component of a Finnish word is one syllable however it is generally accepted that there are only a few monosyllabic words in the Finnish lexicon and these are always lexical morphemes (Sulkala & Karjalainen 1992) that tend to be of the type CV e.g. *me* 'we', *se* 'you' and *ja* 'and' (Pajunen & Palomäki 1984).

Finnish permits 24 different syllable structures (Sulkala & Karjalainen 1992). This includes 16 basic structures; the nucleus (V) alone (i.e. without an onset or coda), the nucleus (V) with a syllable onset which may be one singleton consonant (CV), a two consonant cluster (CCV) or a three consonant cluster (CCCV) and the nucleus with a coda of one singleton consonant (VC), of a two consonant cluster (VCC) or of a three consonant cluster (VCCC).

Finnish also accepts sequences of vowels that are not considered as diphthongs/triphthongs but together as candidates for one syllable nucleus (see 2.6.2

below). The inventory of syllable structures is therefore extended to include these vowel sequences (VV) in each of the possible nucleus. If there are two vowels in the syllable then there cannot be a consonant cluster at the end (Sulkala & Karjalainen 1992).

The possible structures of Finnish syllable structures according to Sulkala & Karjalainen (1992) can thus be presented as follows:

V
CV
CCV
CCCV
VC
VCC
VCCC
CVC
CCVC
CCCVC
CVCC
CVCCC
CCVCC
CCVCCC
CCCVCC
CCCVCCC
VV
CVV
CCVV
CCCVV
VVC
CVVC
CCVVC
CCCVVC

Häkkinen (1983) lists the following structures as the only syllable structures actually observed in Finnish words;

V	e.g. *i-so*
VV	e.g. *ei*
VVC	e.g. *ois* 'to be' (a form of)
CV	e.g. *ja* 'and'
CVV	e.g. *hei* 'hello'
VC	e.g. *on* 'is'
VCC	e.g. *ilk-ku-a* 'to taunt'
CVC	e.g. *nyt* 'now'
CVCC	e.g. *kals-ke* 'lapping/ breaking wave'
CVVC	e.g. *näin* 'like this'

As has already been discussed above, the minimum component of a syllable is a single vowel or diphthong (V), as demonstrated in the first syllable of the Finnish words e.g. *apu* (V,CV) 'help' and, *opet-taa (V,CVC,CV)* 'to teach'. By adding a syllable onset (CV) of the consonant phoneme /p/ to the first syllable of the previous example *apu* the resulting first syllable structure *pa-pu (CV,CV)* 'bean' is produced.

An example of a syllable coda can be found in the first syllable of *kukka* (CVC,CV) 'flower'. Syllables may, of course, demonstrate both of the elements of onset and coda (CVC) as in the words *haas-ka* 'carcass', *mies* 'man', *kais-ta* 'lane', *tuul-ta* 'wind'.

.

If there were no restrictions on word structure, Finnish words could be seen to consist of combinations of any of the syllable structures seen above. However, phonotactic rules restrict the possibilities for word structure.

According to Häkkinen (1983) consonant clusters appearing as syllable onsets (i.e. syllable initial position (SI)) can only appear in word initial syllables (WISI). Consonant clusters also appear as codas (i.e. in syllable final position (SF)).

2.6.2 Phonemic Content of The Finnish Syllable

2.6.2.1 Finnish Consonant Sequences and Clusters

As can be seen above the Finnish phonemic system consists of a total of 13 consonants. If there were no restrictions on the permissible sequences of consonants then a total of 169 (13 x 13) different two consonant sequences and 2197 (13 x13 x 13) three consonant sequences would be observed in the words of the language. All languages, however, make use of only a subset of these possibilities with only certain sequences being seen. Häkkinen (1983) lists the following 69 two consonant sequences and 25 three consonant sequences as being permissible in Finnish words;

Permissible Sequences of Two Consonants Seen In Finnish Words:

/p/ can be followed by /l, p, r, s or t/

/t/ can be followed by /j, k, l, r, s, t or v/.

/k/ can be followed by /k, l, r, s or t/.

/s/ can be followed by /k, l, m, n, p, s, t, v/

/h/ can be followed by /d, h, j, k, l, m, n, r, t and v/.

/m/ can be followed by /m, p, r, s or t/.

/n/ can be followed by /h, j, n, r, s, t or v/

/ŋ/ can be followed by /k or ŋ/

/l/ can be followed by /h, j, k, l, m, n, p, s, t or v/

/r/ can be followed by /h, j, k, l, m, n, p, s, r, t or v/

/j and v/ cannot be followed by another phoneme.

N.B. Although not classed as geminates sequences of identical consonants /hh/, /vv/ and /ŋŋ/ are also observed marginally in the language as in 'livvi', 'huhho' and 'Helsingissa'.

Permissible Sequences of Three Consonants Seen In Finnish Words:

/l/ can be followed by /kk, pp, ss, tt, sk, st or ts/

/m/ can be followed by /pp, ss, st or ps/

/n/ŋ/ can be followed by /kk, ss, tt, sk, st, ks or ts/

/r/ can be followed by /kk, pp, ss, tt, sk, st or ts/

Research in Finland (e.g. Karlsson 1983, Iivonen 1998) has used the term *konsonanttiyhtymä* 'consonant cluster' when referring to within word sequences of consonants. The previous definition of 'consonant cluster' for this research depended upon the consonant sequence existing within the same syllable (i.e. as a syllable onset or coda). Whilst *konsonanttiyhtymä* are consecutive consonant phonemes (and therefore may be one of the list of valid Finnish consonant sequences given above), whether they actually appear within the same or different syllables depends upon the positioning of the syllable boundary (e.g. /hv/ in *kah/va* or *kahv/a* 'handle', /lv/ in *kal/vo* or *kalv/o* 'film' or /rv/ in *kar/va karv/a* 'hair'). For the purposes of this research consonant clusters are defined as consecutive consonants that could only be argued to exist within the same syllable (e.g. /ŋks/ in the spoken Finnish word form *onks* 'is it?' (colloquial form) and /ns/ in *ens* 'first' (colloquial form)). Within word consonant sequences could be argued to exist in different syllables and are hereafter referred to as consonant sequences and not *konsonanttiyhtymä*.

Whereas a word initial consonant phoneme can always be described as WISI later consonant phonemes are dependent upon their relationship with other phonemes for their description. For example, a consonant appearing in position three of a word might be the second consonant of a CV<u>C</u> one syllable word and would therefor be described as WISF, the third consonant of a 3 consonant cluster described as WISI CCC, the consonant of a second syllable of a word starting CVC described as WFSI or WWSI (depending upon the remainder of the word) or even the second consonant of a 2 consonant cluster ending the first syllable of a word and here described as WWSI CC or WFSI CC.

The non-existence of a syllable boundary between consonants in a sequence is thus the determiner of which consonant sequences are to be regarded as true consonant clusters. Only one syllable words can be seen to contain sequences which are unequivocally true clusters as there is no possible syllable division (e.g. the Finnish word form *onks* 'are you' with only one syllable contains the consonant cluster /ŋks /). Words of more than one syllable can be defined using the permissible syllable structures given above such that consecutive consonants fit into one of the previously defined syllable structures. However even following the rules given above for syllable structures and phonemic content of structures the great number of permissible sequences for Finnish still means there is a certain amount of discussion. For example *kalvo* 'transparency/film' is clearly seen to have two (V) nuclei indicating 2 syllables. However even when regarded as a two syllable word it could be argued to have either the format CVC/CV or the format CVCC/V, or CV/CCV. With the first definition the consecutive consonants can be seen to exist across a syllable boundary and the word form does not therefore contain a true consonant cluster but a consonant sequence. The second and third proposed formats do contain consonant clusters. Word medial position sequences of more than two consonants are therefore problematic for Finnish when trying to define clusters as the syllable boundary could be argued to appear in several places. For example, a word of the format CVCCCVC which contains the a sequence of three consonants could be divided as CVCC/CVC or as CVC/CCVC or even as CVCCC/VC or CV/CCCVC.

Word initial consonant sequences can therefore always be regarded as consonant clusters; they appear at the beginning of a syllable in syllable initial position. They do not appear between two vowels and so their precise positioning within the syllable is clear and easily definable. This research will be focusing on the word initial position where sequences of consonants can always be regarded as consonant clusters. Consonant sequences appearing within words, whether defined as clusters or sequences will only occur in word initial minimal pairs (where only the first phoneme is different) if the identical sequence is seen in the two words. In terms of the processing for this study the word initial consonant clusters will be shown and described as phonemic units, different to, and contrasting with, singleton consonants and other clusters.

2.6.2.2 Finnish Vowel Sequences

In addition to the 18 diphthongs provided above which are sequences of two vowels operating as the nucleus of one syllable there are other combinations of vowels which are seen in sequence in Finnish words (Sulkala & Karjalainen 1992, Karlsson 1983). According to Sulkala & Karjalainen (1992) these have originated due to the disappearance of an intermediate consonant or sometimes due to loan words. They may also represent consonant gradation.

Sequences of two vowels which are not considered diphthongs are demonstrated in the following Finnish words;

/haen/ *ha-en* 'I fetch',

/tɑot/ *ta-ot* 'you forge',

/seassa/ *se-assa* 'among',

/teos/ *te-os* 'work, product',

/lian/ *li-an* 'dirt' gen.,

/hioa/ *hi-o-a* 'grind, polish',

/iæn/ *i-än* 'age' gen.,

/koe/ *ko-e* 'test',

/suan/ *su-an* 'brush' gen.,

/pue/ *pu-e* 'put (your clothes) on',

/kyetæ/ *ky-e-tä* 'be able',

/pystyæ/ *pys-ty-ä* 'be able',

/sæe/ *sä-e* 'line, verse',

/næœn/ *nä-ön* '(eye) sight, vision' gen.,

/sœælœæ/ *säi-lö-ä* 'preserve',

/vihreæ/ *vih-re-ä* 'green',

/ɣksiœ/ *yk-si-ö* 'one room apartment',

/lipeœidæ/ *li-pe-öi-dä* 'soak in',

/sæilœessæ/ *säi-lö-es-sä* 'while preserving',

Three vowel sequences are seen in;

/maailma/ *maa-il-ma* 'world',

/aie/ *ai-e* 'intention',

/kauan/ *kau-an* 'a long time',

/ampiainen/ *am-pi-ai-nen* 'wasp',

/liian/ *lii-an* 'too(much)',

/munuainen/ *mu-nu-ai-nen* 'kidney',

/muualla/ *muu-al-la* 'elsewhere'.

Four vowel sequences are seen in;

/raaoissa/ *raa-ois-sa* 'unripe',

/aioin/ *ai-oin* 'I intend',

/kaiuissa/ *kai-uis-sa* 'in the echoes',

/tauoissa/ *tau-ois-sa* 'in the pauses',

There are many different vowel sequences but since only two vowels (with this definition this includes the long vowels and diphthongs) can occur in the same syllable a syllable boundary for sequences of more than two vowels must exist.

Vowel sequences of two, three or four vowels e.g. *h<u>ae</u>n* 'I fetch', *m<u>uua</u>l/la* 'elsewhere', *k<u>aiui</u>*ssa 'in the echoes' are treated differently by different Finnish researchers with regard to frequency calculations. Whilst they are not considered to be diphthongs or triphthongs, occurring together as the nuclei of one syllable, whether they are considered to be several nuclei of different but consecutive syllables or consecutive vowels forming a vowel sequences which is the nuclei of one syllable has varied.

Further complications arise when definitions are considered. Häkkinen(1983) states that of the diphthongs shown above those ending in /u/ or /y/ are considered to be sequences, and not diphthongs, when they appear in a closed second syllable e.g. *rak/ka/us, sol/miu/tu/a* but *siis/tiy/ty/ä.*

This study will be focusing on word initial phonemic usage whether that be a word initial consonant or consonant cluster or a vowel without onset as the first syllable of a word. Within word sequences of vowels should not therefore present a problem. The implications for processing of sequences of vowels in word initial position is considered further in Section 6.

2.6.3 Phonemic Content of Finnish Words

If the eight vowels of Finnish and the thirteen consonants of Finnish had no distributional restrictions, the total number of different syllables that could be produced from the ten syllable type structures outlined above would be about 33,000 (Wiik 1977). However the rules of Finnish phonotactics that determine which sequences of actual phonemes can exist in certain positions within syllables restrict this to approximately 5,700. Of these only about 3000 are used in Finnish.

As Finnish has a system of vowel harmony, the combinatory possibilities for vowel phonemes in the same word are restricted. The vowels /a, o, u/ can never occur in the same word as the vowels /ä, ö, y/. The vowels /e/ and /i/ are neutral and can therefore mix with any vowel. Affixation abides by these rules e.g. *ssa* with *auto* as in *autossa* 'in the car' but ssä with *keittiö* 'kitchen' as in *keittiössä* 'in the kitchen', *ko* with *puhutko* 'do you speak?' but *kö* with *kysyt* as in *kysytkö* 'do you ask?'. In compound words the vowels can mix freely and the second word determines the vowel used in affixation e.g. *kirjahyllyssä* 'in the bookcase' but *kirjakaupassa* 'in the bookshop'.

The rules for the positioning of combinations of phonemes is well documented for standard spoken and written Finnish language. The phonotactic rules as provided in Sulkala & Karjalainen (1992) are now summarised.

2.6.3.1 Finnish Consonants and Consonant Clusters

2.6.3.1.1 Word Initial Consonants and Consonant Clusters

Of the 13 singleton consonants listed above only 11 may occur alone in word initial position in native Finnish words: /p, t, k, s, h, m, n, l, r, j and v/ (Sulkala & Karjalanainen 1992). Word initial /d/ is seen only in loan words.

Whilst originally word initial consonant clusters did not exist in Finnish words, present day speakers do accept combinations of a stop and a liquid as representative of present day Finnish (Sulkala & Karjalainen 1992). The consonant clusters /pl, pr, tr, kr and kl/ are widely seen in loan words (e.g. *'planeetta, presidentti, traktori, krassi, klinikka')*. Karlsson, 1983, lists many consonant clusters seen in loan words such as *blokki, bravo, draama, flanelli, fraasi, glögi, gneissi, knoppi, kvartsi, pneumaattinen, slaavilainen, smokki, snobismi, Sveitsi, ksylofoni, psykologia, sfinksi, skitso, spagetti, staattinen, tsaari* etc.

Word initial consonant clusters are, however, difficult for some speakers of Finnish to pronounce, particularly those with an eastern Finnish dialect where often only the final consonant of the cluster is actually pronounced e.g. *residentti* for *presidentti* (Sulkala & Karjalainen 1992). Häkkinen (1983) states that syllable final consonant clusters occur only in word initial or within word syllable position.

2.6.3.1.2 Word Medial Consonants and Consonant Clusters

All 13 consonants listed above can occur in the middle of native Finnish words. In native Finnish words initial /d/ only appears between vowels (e.g. *sydän* 'heart' or between the phoneme /h/ and a vowel e.g. *tehdas, mahdoton, kehdanmut)*.

The phoneme /ŋ/ is represented in Finnish orthography as 'nk' and the long /ŋ/ as 'ng' (Sulkala & Karjalainen 1992). When the phoneme /n/ is directly followed by a /k/ it is pronounced /ŋ/ thus giving the consonant cluster /ŋk/ e.g. *Helsinki* pronounced

/helsiŋki /.The letter 'g' appears in loan words and following the letter 'n' gives a sequence of /ŋŋ/. Thus /ŋ/ only appears together with another /ŋ/ or preceding a /k/ in the Finnish phonemic system.

There are more than 60 possible sequences of two consonants at the boundary between the first and second syllable in native Finnish words (e.g. *kahva* 'handle', *kalvo* 'film', *karva* 'hair'). There are only 20 that are acceptable at the boundary between the second and third syllable these are /hk, nk, sk, kk, pp, lt, nt, rt, st, tt, ts, lm, rm, rv, hm/ as in *karahka, katiska, porhaltaa, karitsa, kanerva/*.

The rules that govern which consonant phonemes can appear in sequence with other consonants as given by Sulkala & Karjalainen (1992) are;

1. Nasal consonant phonemes (/m, n, ŋ/ do not follow stops.
2. The labial stop /m/ does not combine with any other and stop.
3. A non-dental stop (/p, d ,k/ is never followed by a semi-vowel (/j and v/).
4. Only the non-homorganic stop /h/ follows /n/.
5. A nasal /m, n, ŋ/ is not followed by a non-identical sonorant.
6. No two liquids follow each other.
7. A non-dental consonant and a dental obstruent are not followed by /h/.
8. Two different consonants with the same manner of production cannot be combined. The one exception to this rule being /t/ and /k/ e.g. *kaktus*, 'cactus' and *kotka*, 'eagle'.
9. Only the consonants /p, t, k, m, n, ŋ, s, l and r/ can appear as geminates i.e. long consonants e.g. *kauppa* 'shop' , *sitten* 'then', *kukka* 'flower', *mummi* 'granny', *ennen* 'before', *vessa* 'toilet', *olla* 'to be' and *ymmarrän* 'I understand'.

There are, however, certain standard word forms which do appear to contradict the rules provided by Sulkala & Karjalainen (1992). For example rule 6 states that no two liquids follow each other. However, two liquids are seen in sequence in the verb *kurluttaa* and in the proper name *Kaarlo* (Iivonen 1999).

Derivations of the basic word forms produces many more sequences of consonants e.g./nt, kk, hk, ln, nk, nn, sk, mm, st, ht/ as in *kukinto* 'flowering', *nurmikko* 'lawn', *harvahko* 'sparse', *suudelma* 'kiss' and *ahdinko* 'difficulty'.

Three consonant clusters (CCC) are common in native Finnish words particularly at the boundary between the first and second syllables and in loan words.

Rule 1. Nasal or liquid (/m, n, ŋ, l, r/) + /kk/, /pp/, /tt/ or /ss/.

Rule 2. Liquid (/l, r/) + /sk/, /st/, or /ts/.

Rule 3. Nasal (/m, n, ŋ/) + homorganic stop + /s/, /nsk/ or /nst/.

Only loan words, particularly those of Latin origin, can have word medial clusters of four consonants e.g. *demonstratio* 'demonstration' (Sulkala & Karjalainen 1992).

2.6.3.1.3 Word Final Consonants and Consonant Clusters

Only the consonant phonemes /t, s, n, r, l/ can occur in word final position in native Finnish words (Sulkala & Karjalainen 1992). For example the Finnish words *nyt* 'now', *kaunis* 'beautiful', *avain* 'key', *askar* 'work' and *ol* 'am' (a form of the word 'to be').

No consonant clusters are used word finally in native Finnish words. They do however appear in loan words e.g. *teleks, kliimaks*, and in original words often as onomatopoeic interjections ending in a cluster of a stop plus /s/ (e.g. *niks* 'snap', *rits* 'crackle' and *naks* 'click').

Consonant clusters also appear in spoken Finnish when the omission of final vowels occurs leaving a consonant cluster in the final position e.g. *talost* from *talosta*, *onks* from *onkos* 'is it' and *'yks'* from *yksi* 'one'.

Processes of connected speech account for some systematic changes in the way that words are pronounced. For example, in some dialects word final /n/ changes to /m/

before a word initial /p/ (e.g. /menem pihale/ *menen pihalle* 'I'm going into the garden') and word final /n/ to /ŋ/ / before /k/ (e.g. /menen koti:n/ *menen kotiin* 'I'm going home'*)*. Word final /t/ is often lost e.g. *nyt* to *ny.* Processes of connected speech, whilst not of concern for this study, are also seen to combine and then shorten words (e.g. *otats* from *otatko sinä)* resulting in unusual word final endings.

2.6.3.2 Vowel Vowel (VV) Sequences

Sulkala & Karjalainen (1992) state that 'a maximum of two vowels can belong to the same syllable' and if there are two vowels in a syllable there cannot be a CC at the end. As Sulkala & Karjalainen (1992) take the approach whereby long vowels are regarded as two short vowels then it can be concluded that this statement describes the situation whereby a syllable comprises either one long vowel (e.g. ii, ee, oo etc.) or a sequence of two short vowels including diphthongs (e.g. ai, ei, oi etc.) but not sequences of a long vowel with any other singleton vowel as this would require three vowel sequences.

Sulkala & Karjalainen (1992) list 20 combinations of two vowel sequences that can be seen to occur in Finnish words. Sixteen are said to only exist at the boundary between the 1st and 2nd syllable these are the long vowels /ii, yy, ää, uu, ee, oo and aa/ and the diphthongs ending in /i, e or o/ (/ai, ei, oi, ui, yi, æi, œi, /ie, uo/.

Twenty are seen in later syllables. Permitted VV sequences in the second syllable with no coda are /aa, oo, uu/ e.g. *vajaa* 'short', *takoo* 'forge', *pilttuu* 'stall'. In the second syllable of loan words and inflected Finnish words the VV sequences of /ee, yy, öö and ai/ are seen e.g. *tekee* 'he/she does', *näkyy* 'is seen'.

Karlsson (1983) provides the following list of permissible CVV sequences;

/h/ can be followed by /ää, yy, ai/,

/j/ can be followed by /ää, uu, aa, oi, äi, uo/,

/k/ can be followed by /yy, uu, ai, oi/

/l/ can be followed by /uu, oi, öi, ie, uo, yö/,

/m/ can be followed by /yy, uu, aa, öi, yi/

/n/ can be followed by /ai, yi, uo/

/p/ can be followed by /ii, ää, yy, uu, ui/

/s/ can be followed by /ää, yy, uu, aa, ai, oi, ui, ei, öi, uo, yö/

/t/ can be followed by /ee, aa, ai, oi, äi, ie, uo, yö/

/v/ can be followed by /ai, oi, ei, ie, uo, yö/

N.B. /r/ cannot be followed by any VV sequences.

Spontaneous spoken language will demonstrate processes of connected speech such as assimilation (where one sound moves towards another) and elision (where a sound is lost). These processes are likely to affect the sequences of phonemes observed in Finnish words. Spoken Finnish has not, however, to date been researched sufficiently to provide these different phonotactic rules and the rules for written language are insufficient to provide the necessary guide for homophony testing where phonemes are to be replaced by only permissible replacement at each word position. For example, a look at the standard Finnish phonotactic rules outlined above one might conclude that a /j/ cannot appear in word initial syllable final (WISF) position. However, in connected speech /j/ is found in WISF position e.g. *kaij joskus* (from *kai joskus* 'maybe sometimes'). A fuller discussion of the implications of using spontaneous speech for this research is given in Chapter 6.

Section 2.7 presents what is known to date of the actual use that the Finnish language makes of these possibilities in terms of the range of syllables and word structures and the frequencies of phonemes seen in the words of the spoken language. As there is little previous research on some of these frequencies one aim of this research will be to present frequency data on phoneme occurrence.

2.7 Finnish Frequency Statistics

Sections 2.5 and 2.6 above describe the components of the Finnish phonological system and the rules of Finnish word and syllable formation (i.e. the phonotactics). This section summarises what is known statistically about these structures. That is, how

frequently the spoken language makes use of these various possibilities in the words of Finnish.

This research focuses on spoken language. There is, however, very little published work to date that outlines in any detail the statistics of spoken Finnish in a standard way. Most research deals only with written language (e.g. Häkkinen 1983) whilst those that have looked at spoken language have restricted their study to a form of standard dialect heard on radio and TV broadcasts (Pääkönen 1973). One study (Pajunen & Palomäki 1984) observed a range of differences in word lengths, word structures and phoneme frequencies between spoken and written Finnish. Vainio (1996) concludes that spoken and written samples of the language may differ in several important ways.

Firstly, the frequency analyses of Finnish phonemes/graphemes and words varies according to the form in which they are represented i.e. written or spoken. Vainio (1996) finds that different words occupy a higher status in spoken Finnish than in written Finnish. Newspaper reports and spoken language such as broadcasts often contain high frequencies of words specific to current topics of news items and names thus context plays an important role in the analysis of data. Vainio concludes that the context of the written material that he analyses e.g. magazine interview situations where people often reminisce about the past, may well cause a tendency to use the words *oli* 'was', *sitte* 'then' and hence the high frequencies of these words and the phonemes they contain.

In addition to this, the non-systematic way of treating geminates and long vowels (discussed above) and the various ways of treating phonemic units such as diphthongs and consonant clusters has sometimes resulted in contradictory results. A count of the distribution of graphemes as completed by Pääkonen would be expected to give different results to one of phonemes as completed by Vainio. With a count of graphemes the long vowels and geminates, which are represented by two graphemes e.g. 'ii', 'kk', would have each grapheme individually counted thus ignoring the phonemic status of the long vowels and geminates.

One of the only studies to look specifically at spontaneous colloquial spoken Finnish is that completed by Vainio (1996). Vainio compared a large spoken language corpus (600,000 running words) with written texts to produce statistics on the different frequencies of phonemes, diphones and triphones. This study did not go on to consider the particular syllable and word structures and their frequencies and the frequencies of phonemes by word position however it did recognise the necessity for long vowel phonemes and geminates having phonemic status. The findings from this and some of the other research completed on Finnish frequency statistics to-date is now summarised under the headings of word length, phoneme frequency, word type frequency and syllable type frequency.

2.7.1 Phoneme Frequency

Table 2.1 shows Vainio's (1996) findings on the ten most frequently observed phonemes in spoken Finnish.

Table 2.1 – Vainio's Frequency of Phonemes

PHONEME	FREQUENCY as %
/i/	10.58
/a/	8.93
/e/	7.75
/s/	7.33
/t/	7.19
/n/	7.16
/o/	5.98
/k/	5.38
	4.50
	4.05

The most frequent phonemes, in order of frequency were found to be;

/i/, /ɑ/, /e/, /s/, /t/, /n/ and /o/.

The first long vowel /i:/ occurred at rank 16 and the first geminate /ll/ at position 17.

In their study Pajunen & Palomäki (1984) concluded that whereas 'a' is the most common Finnish grapheme of written discourse data, the phoneme /i/ occurs the most frequently overall in speech.

Pääkkönen (1973) found little difference in rank ordering of the phonemes in terms of frequency between written and spoken material with the most frequent phonemes in both cases being (in order of rank) /ɑ, ɪ, t, n, e, s and l/. A straight orthographic transcription was used with letters being counted to represent phonemes and those phonemes without a specific orthographic representation such as /ŋ/ were not counted.

The vowel phonemes /i/ and /e/ which are indifferent to vowel harmony were found to have particularly high frequencies compared to the other vowels. Of the consonants the most frequent group were identified as the dentals /t, d, s, r, l, n/ (65%), followed by the labials /p, b, m, f, v/ (14%). The voiceless plosives /k, p and t/ were found to account for almost a third of the consonants presence. Pääkkönen concluded that the phonemes that are utilised most in the suffixes and particles of Finnish will be seen most frequently in both standard spoken and written Finnish, particularly in the more formal contexts.

Häkkinen (1983) looked specifically at the language of Finnish nursery stories. Whilst the stories were those that would typically be read to Finnish children by parents, the actual readings were not phonetically transcribed. Instead, the written texts were analysed for grapheme frequency.

Häkkinen's findings on the most frequent phonemes are provided in table 2.2.

Table 2.2 – The Ten Most Frequent Finnish Graphemes (Häkkinen 1983)

	All Positions	Word Initial	Word Final
1.	/ɑ/	/k/	/n/
2.	/i/	/s/	/a/
3.	/n/	/j/	/i/
4.	/t/	/p/	/æ/
5.	/e/	/t/	/t/
6.	/s/	/m/	/e/
7.	/k/	/v/	/o/
8.	/l/	/n/	/s/
9.	/æ/	/h/	/u/
10.	/u/	/o/	/r/.

The ratios of singleton phonemes, both consonants and vowels, to consonant sequences and vowel sequences are provided in table 2.3.

Table 2.3 –Phoneme Ratios (Häkkinen (1983))

Singleton Consonants (C)	-	40.3%
Singleton Vowels (Short) (V)	-	36.6%
Consonant Sequences (CC)	-	6.3%
Long Vowels	-	5.7%
Diphthongs	-	5.7%
Geminates	-	5.1%
Consonant Sequences (CCC)	-	0.3%

The diphthongs ordered by frequency were found to be;

1. /oi/
2. /ai/
3. /ei/
4. /uo/
5. /ie/
6. /ui/
7. /au/
8. /æi/

9. /yœ/

10. /yi/

The long vowels were ranked in order of frequency giving; /a:, i:, ä:, e:, u:, o:, y: and ö:/.

The most frequent two consonant sequences ranked in order of frequency are given as;

/ll, tt, st, ss, nn, kk, ks, ŋk, lt, nt, sk, jt, lk, ts, ns, mm, tk, lm, rt, rm/.

The most frequent three consonant sequences are given as;

/nss, rkk, ntt, rtt, lkk, rss, lsk, ŋkk, ltt, rst, mpp, rpp, rsk, ntr/.

2.7.2 Word Type Frequency

Research by Pääkkönen (1973) looked specifically at the effect of context on the statistics seen in written and standard spoken Finnish. He analysed a corpus of current standard Finnish created from newspapers and magazines of 1967, Finnish literature and spoken radio programmes of 1968 and compared these to parliamentary speeches from 1968-1969. Two thirds of his data was for written Finnish. He found that the average number of phonemes per word was similar for both spoken and written samples; 7.09 phonemes for spoken and 7.36 for written.

Pajunen & Palomäki (1984) concluded that the mean length of a word token in spoken Finnish is approximately 4.6 phones with the most commonly occurring word token being 2-3 phones in length and the majority of word tokens consisting of between 2 and 6 phones. This compares to word tokens of 7.5 graphemes in mean length for written texts, a most commonly occurring word token of 6 graphemes and with the majority being 5 or 6 graphemes. As well as the presence of inflectional morphs lengthening the word tokens (with more of these seen in written text) compounds also account for the higher mean length of work tokens as their findings show that compounds constitute approximately 3% of the spoken word tokens compared to 15% of written text word tokens. Their report also commented on a higher ratio of inflectional markers in written

text (four out of five word tokens containing an inflectional marker in written text, only one out of two in speech) and concluded that case is the determining factor in the difference in length between the word tokens of spoken and written discourse data; a case marker occurs in 30% if spoken data words and in 50% of written data words.

Karlsson (1983) notes that one syllable words make up only about 0.1% of the 70,000 words in a standard dictionary of Finnish. Long words are therefore expected to be more common in Finnish than in English.

2.7.3 Word and Syllable Structure Frequency

The most frequent syllables types of CV, CVC and CVV accounted for about 80% of all occurrences (40 %, 28% and 12.% respectively) in a corpora analysed by Iivonen (1998).

Häkkinen (1983) analysed the words for the frequency of syllable types at specific word positions. It was found that the most frequent syllable type structure (CV) over all word positions structure, CV, also occurred most frequently in word initial, word medial and word final positions as table 2.4, below, demonstrates.

Table 2.4 – Syllable Type Frequency (Häkkinen 1983)

Syllable Type	General	Word Initial	Word Medial	Word Final
CV	1	1	1	1
CVC	2	2	2	2
CVV	3	3	3	4
CVVC	4	4	4	3
VC	5	5	5	6
V	6	6	6	5
VV	7	7	8	8
CVCC	9	8	-	-
VVC	8	9	7	7
VCC	10	10	-	-

n.b. – indicates not present

She also looked at the frequency percentage of actual phonemes found by different syllable types out of those theoretically possible. For example, there are 8 short vowels and each was tested to see its presence in the syllable structure type V.

Her findings concluded that;

All the vowel phonemes were found in V syllable structures.

95.5% of the possible CV structures found overall (100% in word initial position, 96.2% in word final position).

72% of the possible VV syllable structures were found (52.2% in word initial position, 26.7% in word final position).

76.7% of the possible CVV syllable structures were found (81.9% in word initial position 66.5% in word final position)

81.9% of the possible VC syllable structures were found (75% in word initial position, 62.5% in word final position).

74.1% of the possible CVC syllable structures were found (77.2% in word initial position, 61% in word final position).

16% of the possible VVC syllable structures were found (45.9% in word initial position, 30.7% in word final position)

32.7% of the possible CVVC syllable structures were found (54.2% in word initial position, 36.2% in word final position).

16.7% of the possible VCC syllable structures were found.

32.1% of the possible CVCC syllable structures were found.

This chapter set out to explore the phonemic systems of English and Finnish with the objective of defining an inventory of phonemes and acceptable phonemic units for the two languages. In addition previous findings on the frequency of usage of these phonemes and phonemic units and typical structures of words and syllables were presented such that the subsequent development of the processing required for the FUSE assessment (presented in Chapter 5) could be based upon representative structures and frequencies of usage.

Chapter 3 will now present in outline various theories of phonological acquisition and the predictions that these theories make about the relationship between ambient language and phonological development. Previous findings on child phonological development for the two languages will also be presented so that new findings obtained with this study can be identified and presented in Chapters 7 and 9.

Chapter 3 : Theories of Phonological Acquisition

This chapter sets out to explore two main theories of phonological acquisition and the predictions that these theories make about the role of ambient language during the acquisition process.

The role of ambient language in phonological acquisition can, in simplistic terms, be viewed in two extreme and opposing ways:

a) that phonological acquisition is totally dependent upon ambient language (i.e. is completely language specific).

b) that phonological acquisition is totally independent of ambient language and follows universal rules regardless of the specific language being acquired.

With a language specific approach the language surrounding a child is said to have a direct impact upon the phonological acquisition of a child acquiring that language. Often this relationship is observed in the frequency of occurrence of particular phonemes in the adult ambient language in relation to the earliest acquisition in the child language (i.e. the child will acquire the earliest those phonemes that occur most frequently in adult language). Cross-linguistic research has provided evidence of the effects of ambient language on very early vocalisations (e.g. Boysson-Bardies et al., 1989, 1992).

The alternative viewpoint is that the actual language surrounding a child does not have such an important role to play during language acquisition as certain innate processes which predispose the child to learn language in a specific way, however poor the input. Chomsky's "poverty of stimulus" argument (1965) claims as adult language is full of retracings, ellipses and slips of the tongue no learner could possibly build an adult language out of such degenerate input. Also, that children are seen to produce novel meaningful strings of words that have not been previously heard in ambient language is seen to indicate that they must, therefore, have something else to assist with the acquisition process, perhaps a universal grammar or a natural innate tendency to form

language, something that Chomsky calls the Language Acquisition Device (LAD). This theory also helps to explain the cross-linguistic similarities that are observed (see below) as children acquiring different languages might actually go about the process of acquisition following innate rules rather than external factors such as the frequency of input etc. At the level of phonological, rather than syntactic development, the child does hear direct evidence of its surrounding phonological system, in terms of the words it is surrounded by. It hears not only the phonemes themselves but the rules of the phonemic system it is acquiring (i.e. how the phonemes can be combined together to form words). With the external input for phonological acquisition being words made up of phonemes from a particular phonological system (rather than infinite combinations of words) it could be argued that there is less opportunity for speakers to display those 'slips' as a substitution of a phoneme for another or mis-pronunciations of phones themselves.

Fletcher & MacWhinney (1995) can accept the role of certain innate processes of development e.g. anatomical maturation, developing abilities of fine motor control, auditory sensitivities etc. but question the role of innateness in theories of phonological development where the phoneme is not definable in phonetic terms (i.e. as an absolute) but only in terms of contrast. However, despite this language specific evidence of phonemes and phonological system in input, cross-linguistically children are observed acquiring phonemes along predictable routes regardless of the language specific input they have been exposed to.

Jakobson (1968) argues that phonological acquisition follows innate universal rules that are pre-determined by language universals and Stampe (1969) refers to innate processes. These theories link phonological acquisition to a phoneme's usefulness, not in terms of its usefulness within a particular phonemic system, but in terms of its usefulness across languages (i.e. how many languages make use of this feature). The 'marked/unmarked' theory espoused by Jakobson (1941/1968) and Eckman (1977) hypothesises that there are sounds or features which are phonetically more natural (i.e. unmarked) and therefore acquired earlier by the children. There are many other models, within this general framework of non-language specific processes, directly affecting the acquisition process. For example, the biological model discussed by

Locke (1993) and Kent (1992) emphasises the role of articulatory and perceptual constraints on children's acquisition of phonology and Schwartz (1988) mentions the possibility of phonological selectivity. Section 3.1 follows some of these arguments further.

Cross-linguistic research has added a new dimension to the argument highlighting not only the cross-linguistic differences that might be expected with a language specific approach when observing children's phonological acquisition for different languages but also some similarities which cannot be fully explained by a language specific approach. The inter-action between language specific components observed in child language and universal trends observed across children acquiring different languages is therefore more complex than the simplistic view outlined above. Theories of phonological acquisition must be able to account for not only the similarities observed among children acquiring the same language (and phonological system) but the differences as well. They must also be able to account for the similarities found when viewing phonological acquisition cross-linguistically (i.e. that certain phonemes do tend to appear earlier in the child language of several different languages). A summary of the general findings on phonological acquisition are presented in Section 3.2 and specific cross-linguistic findings in Section 3.3.

Ingram (1989:210) states that "The course of early phonological development is heavily influenced by the properties of individual words" promoting the idea that phonological acquisition follows the child's exposure to a particular lexical base. Grunwell (1985), in fact identified the word for prioritising therapy goals by considering the phonemes usefulness in language. Leinonen (1990) suggests that one theoretical possibility is that the order of phonological acquisition may be motivated by homophony reduction in a child's lexical system. Research into these areas is ongoing, however, they do appear to offer a viable theoretical explanation for individual diversity. In recognising the fundamental universal principles which do, as Jakobson suggested, appear to underlie phonological acquisition but at the same time allowing for variation, based upon experience/exposure to the language or some other factor, they encompass a multi-dimensional model of language acquisition and allow for complexity.

Phonological acquisition can therefore be said to follow certain routes partly influenced by language specific features and partly by universal features. The concept of functional load as a measure which recognises both the inherent language specific features surrounding a child and the systemic usefulness of particular phonemes acting within the overall phonemic system has been cited as a potentially useful method for assessing child phonological development (e.g. Pye, Ingram & List 1987 and Ingram 1989, see below). Very little research has been conducted in this area, however, with particular attention to child language, hence the focus of this study. Functional load as a concept will therefore be explored in this chapter in Section 3.4.

Section 3.5 will provide a summary of the predictions that both a totally language specific frequency based approach and Jakobson's Universalist theory make for Finnish and English phonological acquisition. This will enable a further assessment of the usefulness of these methods to be outlined in light of the findings of this study (Chapters 7 and 8). This study will observe the frequencies found in the adult language corpora of English and Finnish and then go on to assess whether there does appear to be correlation between these and child language frequencies. The new FUSE method of assessment developed for this study aims, in addition to the purely frequency based method above, to include an element of the systemic usefulness of phonemes based upon their functional load. The second proposal will therefore assess the correlation between child language and FUSE.

The segmental approach to viewing phonology as has already been presented in Chapter 2 will be adopted and the focus on the segmental elements of language remains. The basis of measurement will be the words utilised by children as a direct representation of production ability and acquisition of the phonemic system (rather than simply the acquisition of phones). It is recognised too that a child's production of words may not be a full reflection of their perceptive abilities, however, the possible sequence of events necessary for the processing of words from input (ambient language) to output (production) is beyond the remit of this study. The production of word forms will signify acquisition for this study and will be used for the basis of assessment of the acquisition process.

3.1 Theories of Phonological Acquisition and the Role of Ambient Language

The 'nature' or 'nurture' debate is at the centre of many acquisition debates. At one end of the spectrum is the view that the environment directly surrounding a child will determine the child's acquisition route (e.g. behaviourist theories) whilst at the other is the view that children are somehow predisposed to acquire certain innately human abilities, of which language is one. There are therefore opposing views on the role that ambient language plays during language acquisition, including phonological acquisition.

Following the behaviourist view, phonological acquisition is predicted to be language specific so that children acquiring different languages follow different acquisition routes. In line with this children exposed to the same adult language would be expected to demonstrate acquisition processes that correlate to the adult language surrounding them. Various attempts to explain this correlation and provide a measure of this closeness have been proposed mainly based upon the frequencies observed in adult data (e.g. Stoel-Gammon 1998 and Mines, Hanson & Shoup 1978). Languages differ with respect to the relative frequencies of usage of various phonemes, word lengths and consonant clusters. With this viewpoint children acquiring different languages would be expected to have different and unique acquisition routes based upon the frequencies observed in the particular language environment that they find themselves in.

It is recognised that frequency of usage may be only one factor that is influencing this process. Olmsted (1966) believed these influencing factors to be based around the ease of perception and the frequency of occurrence of phonemes in adult language (i.e. the child would acquire those phoneme contrasts that are easier to perceive and are heard in the language around it). Empirical studies, however, have not supported this hypothesis (Olmsted 1971). Ambient languages have other common features too which may equally be influencing the acquisition process. For example, some sounds in a the phonemic inventory of a particular language might be inherently more difficult to articulate than others and stress patterns are also likely to effect individual phoneme production.

Jakobson(1941/1968) suggested that whether a sound would be acquired early by children learning a particular language could be explained by the relationship between the phonemic inventory of the language and the distribution of the phoneme amongst the world's languages. According to his 'Laws of irreversible solidarity' nasals, front consonants and stops (found in virtually all languages) would be acquired earlier than their oppositions, orals, back consonants and fricatives. He proposed that there are certain sounds which are more basic and central to all human languages and these sounds would therefore be acquired earlier than other sounds. Jakobson's view of phonological acquisition in terms of oppositions or contrasts set the agenda for subsequent studies of child phonology.

According to Jakobson (1968), the order of emergence of phonological contrasts is instead regulated by structural laws which are universal and innate. However, just as each language has a different phonological system with a different set of feature contrasts, the theory has different implications for different languages. Each child would therefore be expected to follow a particular, pre-set, pre-definable phonological acquisition sequence for the particular language they are acquiring. Jakobson's theory does, within the framework of contrasts, allow for individual variation, be it through selection or governed innately. For example, stating that a bilabial vs. oral contrast will emerge simply implies that the child may select any of the three bilabials that are possible in English (/p/ or /b/ and /m/), it does not provide an order for the three. An outline of the development of this theory is presented in more details below.

The notion of 'markedness' has been used to interpret the similarities and differences in the order of sound acquisition (e.g. Eckman 1977). It was hypothesised that those sounds which appear early in a child's inventory are maximally unmarked, while those occurring late are marked. Therefor, children would use unmarked sounds as substitutions for marked sounds. Edwards (1974) study of English speaking children aged 1;8 to 3;11 found that children usually substituted the unmarked member for those marked contrasts (e.g. [s] for /ʃ/) but details varied from one child to another and from one developmental stage to another.

Some researchers found that the traditional labels in the taxonomy of oppositions such as voice, place and manner of articulation were not adequate when explaining the order of acquisition of phonemes. A more detailed description unit was adopted; the feature. The feature system focused on the articulatory differences between phonemes (see Chomsky & Halle 1968). Among the most important features are those distinguishing between vowels and consonants (sonorant, vocalic, consonantal), those distinguishing the sound in terms of place of articulation (anterior, coronal, high, low, back and rounded) and those distinguishing the sounds in terms of manner of articulation (nasal, lateral, continuant, delayed release and strident). Each phoneme is described in terms of the combination of these features that it displays. Unmarked features are predicted to be acquired first because unmarked features are considered more phonetically natural. Children, according to this hypothesis, would tend to replace marked features with unmarked features (e.g. Menyuk 1968).

One of the main problems with both of the approaches outlined above, language specific or universal, is that even children learning the same language do tend to display individual differences. In a way therefore despite the two approaches outlined above being seen to be at opposite ends of the argument they do both accept that there is a natural path of acquisition for a specific language.

Many linguists, having observed the variation in acquisition order of actual phonemes between children acquiring not different but the same language and phonemic system, have reached the conclusion that phonemic acquisition cannot be either a totally innate and universal process nor totally language specific. Whilst it may follow certain universal principles a degree of variability based around the child's unique experiences of language must be allowed for. Cognitively based theories (e.g. Macken and Ferguson 1983) attribute similarity and universal features of acquisition to the nature of articulatory and auditory systems of children and variation to the child being an active learner who participates in the acquisition process.

Whilst there are general trends, such as frequency of particular phonemes, that can be observed within the adult/ambient language surrounding a child each child would actually have a unique exposure. The specific speech directed to the child by a

caregiver (child directed speech) will be unique and therefore a child's path to acquisition might be expected to be idiosyncratic and based upon the environment directly surrounding him. Macken and Ferguson (1983), for example, argue that it is precisely this variability of input that explains the many differences observed in children's language development.

Within the general framework of theories allowing a direct role for individual ambient language influence are the connectionist accounts of language acquisition with the view that the child learns or builds his/her own language system directly from the input it receives (i.e. from the surrounding environment). Learning, with this approach, is the result of relative strengthening and weakening of neural pathways due to varying input. Parallel Distributed Processing (PDP) theories and the neural network models of Rumelhart and McClelland (1986) set out to demonstrate how this might be achieved and there is currently much ongoing research into this area. With this idea there would be little role for a theory which helps to predict or pre-empt the path of phonological acquisition as each child's path of acquisition would be totally unique. These theories therefore offer little support for clinical language work which would use a framework of normal acquisition to both help identify abnormal acquisition and to treat children who were not developing along this normal path.

3.2 Features of Child Phonological and Lexical Development

Children generally acquire most of the segmental and supra-segmental elements of their language between the ages of 1;6 and 8;0 (Stoel-Gammon & Dunn 1985) and the main advances of phonological development between the ages of 2;0 to 4;0 (Ingram 1989).

Research into phonological development and acquisition of specific phones has in the main focused on consonants. This may be because vowels are more difficult to transcribe reliably (Vihman 1996) or because in the main they are mastered earlier than consonant phonemes and tend to show fewer errors (Stoel-Gammon & Herrington 1990). According to Stoel-Gammon (1998), the average 2 year old has a phonetic inventory containing voiced and voiceless labial, alveolar and velar stops, labial and alveolar nasals, the glides [h] and [w] some fricatives. By the age of 3 all the basic

elements of the adult system are likely to be present with the inventory including examples of all place, manner and voice classes.

Although it is recognised that there is considerable variation in both the age at which children acquire particular segments and also the precise order of acquisition of the segments, some general developmental patterns, in terms of the types of consonants and vowels, have been observed. Jakobson's claims on the order of phonemic acquisition of sound classes for Indo-European languages were confirmed by several early studies (e.g. Leopold 1947 and Templin 1957). Most children in these studies of Indo-European language acquisition were observed to acquire the classes of stops and nasal before fricatives, affricates and liquids and most acquired the front consonants before the back ones, as predicted by Jakobson.

Creaghead (1989:53) summarises research into English phonological acquisition order giving actual phoneme order within manner of articulation class as follows;

Nasals - /m/ and /n/ before /ŋ/

Stops - /p/ and /b/ before /t/, /d/, /k/ and /g/

Fricatives and Affricates - /f/ before all others, /θ/ and /ð/ last of all

Glides and Liquids - /w/ and /j/ before /r/ and /l/.

Several studies of Finnish phonological acquisition, including those of Iivonen (1986) and Toivainen (1990), have however highlighted a major difference in the acquisition of the Finnish phonemes. Finnish children appear to acquire dental consonants before labial consonants. Toivainen (1990:63) gives the example of his own son who "produced the dental plosive /t/ and velar /k/ before the age of one year, whereas the bilabial /p/ did not appear until six months later". Another very interesting feature noted by Toivainen is the late introduction of the voiced dental plosive /d/ which was the last consonant phoneme to be acquired by all the children he studied. He gives the

explanation of the /d/ being only a marginal consonant for this late acquisition, "it may be said to belong to only careful, correct spoken Finnish .... scarcely used in the rural dialects" (Toivanen 1990:65).

One of the few studies to look specifically at Finnish phoneme acquisition by word position was that completed by Kunnari (2000). When looking specifically at word initial phoneme inventories of Finnish children she found that the inventories were extremely restricted. At the four word stage only 2 stable consonants were recorded in the adult forms of target words attempted by the children and this increased to only 5 at the 25 word stage. Even with all word positions considered the children had a mean inventory size of only 4.3 phonemes at the 4 word stage up to 9.9 at the 25 word stage. Kunnari observes that one of the reasons for such a small inventory might be that the Finnish language has relatively few consonant phonemes compared with other languages (e.g. only 13 compared to 24 for English, see Chapter 2). Thus children acquiring Finnish would be expected to hear less consonant sounds in their environment and this variation in input may effect their acquisition route.

In an attempt to define what exactly we might mean by 'acquisition' of phonemes Sander (1972) reanalysed data collected much earlier by Wellman (1931) and Templin (1957) and proposed age ranges (rather than a single age) within which normally developing children acquire phonemes. Sanders main concern was not with the children's acquisition of phonemes (i.e. using phones as contrasts) but the production of phones themselves. Single word elicited and imitated productions were compared to ascertain the ages at which 50% and 90% of the children tested demonstrated "acceptable" production of consonants. The age of 'customary production' was gauged to be at the 50% correct performance level and mastery of production when 90% of the children correctly produced the consonant phonemes at all possible structural positions (i.e. initial, medial and final position). For the English labials /p/, /b/, /m/, and /w/, the velars /k/, /g/, /ŋ/, the alveolar stops /t/ and /d/ and the consonant phonemes /h/ and /n/ customary production level was reached by age 2. All of these phonemes apart from /b/, /k/, /g/, /d/, /t/ and /ŋ/ had reached master of production level at age three with /b/, /k/, /g/ and /d/ reaching mastery at age four and /t/ and /ŋ/ at age 6.

The phoneme /s/, however, took from age 3 to age 8 to move from customary to production levels demonstrating perhaps that some phonemes are somehow being treated differently within the system than others. The study does not give any data on the position of phonemes within words i.e. in which positions the child first acquires these segments. Whether word or even syllable structure has a role to play in the acquisition route remains to be determined. Nor does it tell us about the acquisition of sequences of consonants or vowel phonemes. Treating acquisition as the presence of individual phonemes in this way does not really tell us much about the underlying system and might be regarded as purely 'a phonetic acquisition assessment'.

The relevance of the study of the acquisition of individual phonemes has also been raised (e.g. Grunwell 1980, Menn & Stoel-Gammon 1995). The child is acquiring a phonological system, not as individual sounds, which in themselves carry meaning but as a system of contrasts in order to be able to demonstrate lexical differentiation. A perhaps more valid way of assessing the child's phonological acquisition is in relation to which particular contrasts a child is actually making use of to signal contrasts in the words that he/she is using. Stoel-Gammon and Cooper (1984) in a study of lexical and phonological acquisition conclude that there is a very low degree of universality in phonological or lexical development with only a limited number of patterns underlying children's early phonological productions. They also stress the need for more empirically based studies.

Child realisations of phonemes when compared against the adult target forms provide an insight into the acquisition process. This relationship between the adult target which the child is assumed to be aiming at and the child's actual realisation has been expressed in a number of ways. For example in terms of distinctive features (Oller 1973), as phonological rules (Smith 1973) and as phonological processes (Olmsted 1971 and Stampe 1969 and 1979). Stampe's (1979) theory of Natural Phonology provides the framework for phonological process analysis. With this theory the child is assumed to 'possess' an innate set of natural phonological processes, which operate on language, facilitating phonological development. The processes are assumed to be innate and universal and the process of phonological acquisition is the accommodation of these innate and universal processes to the requirement of the ambient language. As

the child progresses towards the particular adult system, phonological processes are suppressed, limited or modified in some other way. In this way, the ambient language affects language acquisition. Even this approach has, however, been found to be inconsistent when languages from different language families are considered (see below).

Various studies on the early growth of children's vocabularies have tended to conclude that children typically acquire their first words round about nine to 12 months of age and by the time that they are two years old they may have acquired up to 500 words or more (Barrett 1995). The first 50 words or so have a simple syllable structure, e.g. CV, CVC, reduplication, e.g. CVCV, and a small phonemic inventory (Ingram 1999). As children develop their lexicons the length of target words in syllables is seen to increase, more words with consonant clusters are observed and the range of phonemes seen in words increases.

Several hypotheses have been proposed to explain this from articulatory or perceptual difficulty of certain targets (Olmsted 1971), subject-specific experiences (Ferguson & Farwell 1977) and frequencies of characteristics of the sound system of the child's native language (Mines Hanson & Shoup 1978). However, as Dobrich (1992) observed, already by age two children demonstrated phonological characteristics very similar to those of older speakers with respect to word length, word initial consonant clusters and constituent phonemes suggesting that selectional biases observed at younger ages are of quite short duration. The later development of word final consonant cluster usage is attributed to syntactic limitations rather than phonological development: the children produced relatively few grammatical inflections, copulas, auxiliary verbs, and other closed-class words at the youngest age. Interestingly this study demonstrated that frequency of phoneme in ambient language might have an important role to play. Even at age 2;0 it was quite rare for a high frequency phoneme to be avoided or for a zero frequency phoneme to be preferred or exploited. Individual differences were not found to be that great: no subjects showed a tendency to avoid multisyllabic targets, the children were similar in both their adult-like lexical choices with respect to initial consonant clusters and less frequent attempts to produce words with final clusters and their phoneme usage.

Children's earliest attempts at words are often sporadic and extremely variable in pronunciation. Whilst the child may have a systematic way of reducing adult words to forms which fit within their production capacities, children's earliest attempts at words are believed by some to be unsystematic in their phonological relation when compared to the adult target forms (Menn and Stoel-Gammon 1995). The common occurrence of homonymy in early speech means that many different words may actually be represented in the child's production by one form (Vihman 1981). The words attempted by children, rather than the actual pronunciation, are therefore good indicators of the developing lexicon. With words phonemes are recognised not simply as sequences of isolated units but within the context of the function they perform within the language to provide the contrast between words.

3.3 Cross-Linguistic Findings

Cross linguistic studies of language acquisition (e.g. Slobin 1985, 1992, 1997) enable some of the findings so far highlighted to be looked at across languages. By observing the similarities in development across different languages the concept of universality can be explored whilst at the same time the role of underlying features of the specific language systems, which might be responsible for the differences, are clearly visible.

Whilst there do appear to be universal tendencies in children's phonological acquisition routes that can be explained by universalist theories cross-linguistic research has revealed that there are in fact consistent cross-linguistic differences as well. Pye, Ingram and List (1987) compared the word initial phoneme inventories of English and Quiche children and found that the children showed the effects of their input language right from an early age. For example, English children do not acquire the affricate /tʃ/ until late in development. Jakobson also made this prediction based upon the fact that it demonstrates a marked feature and is rare in languages of the world. The Quiche children, however, used them among their earliest sounds. Pye et al argued that the only explanation for this difference could be the linguistic experience of the child (i.e. the ambient language surrounding them) and they went on to show that it was not the absolute frequency of a particular phoneme that determined its acquisition position but instead its functional load (i.e. its systemic usefulness). This explanation was also

supported by Ingram's (1988) study of word initial /v/. Whereas /v/ is acquired late in English it was found to be one of the earliest phonemes among Swedish, Estonian and Bulgarian speaking children.

According to Macken (1978) the early acquisition of /ʃ/ by Spanish children may be because it is more frequent in the corpus addressed to Spanish children by virtue of its high frequency in nicknames, diminutives etc. Frequency of sounds and sound patterns in adult language, language specific factors, are in fact highlighted in several cross-linguistic studies as a contributing factor to the stages of children learning those languages. Toivanen (1997), see above, cites that /d/ which is one of the earliest acquired obstruents in other languages is in fact acquired very late in Finnish, often the last consonant to be acquired and he also relates this directly to the fact that /d/ is in fact highly restricted in the adult Finnish language.

Cross-linguistic studies have not only shown the difference in acquisition routes of phones but have also highlighted the differences in the processes of acquisition. For example, whereas word-initial consonant omissions has often been regarded as an atypical process, Savinainen-Makkonen (2000) found that for the six Finnish children she studied it was in fact a general development process among early learners of Finnish. Another difference that she identified was that whereas English children demonstrate a high use of monosyllabic words in the early stages of acquisition (see above) with Finnish the characteristic use of longer multisyllabic words was observed. This is attributed in her study to the nature of the ambient language surrounding the children (i.e. that Finnish has more multisyllabic words to start with) and, as presented in Chapter 2, that one syllable words make up only a small percentage of Finnish words. Boysson-Bardies & Vihman (1991) in a study of children acquiring four different languages found that Japanese children tend to produce words containing three to five syllables at 15 months of age. Boysson-Bardies et. al. (1992) compared the phonological development of children learning French, Japanese, Swedish and English and found that mono-syllable word structure predominance only existed with the English children at the 25 word stage of development. Kunnari (1999) compared these findings with her findings for Finnish children and again concluded that at the 25 word point all of the non-English speaking children and especially the Finnish children,

produced more disyllabic words than the English children. English children on the other hand had twice as many mono-syllabic words than Finnish speaking children. Cross-linguistic studies of word structure have therefore provided a new way of viewing the earlier English results perhaps providing support for the language specific argument of acquisition based upon ambient language rather than a universal approach.

Iivonen (1995) studied the word structures used by two Finnish speaking children and concluded that even at the age of 0;10; to 1;7 the children predominantly utilised more complex structures than those predicted for English, finding VCCV, CVCCV, CVCV and CV the most frequently used by one child. Kunnari (2000) in her study of Finnish children aged 8.9 months to 18 months, found that the CV syllable was the most frequent form in the children's productions followed by CVV, CVC and VV. She also concluded that the proportion of open syllables was much higher than that of closed syllables.

Cross linguistic studies of children have also highlighted new information about differences between different language's acquisition rates. For example, Mowrer & Burger (1991) carried out a comparative study on Xhosa and English speaking children and found that Xhosa speaking children acquired some phonemes, common to both languages, earlier than English speaking children. In another study of Cantonese speaking children, So & Dodd (1995) found that while Cantonese speaking children acquired consonants in an order similar to that of English speaking children, the rate of acquisition was more rapid and they used some language specific rules such as affrication. Pye, Ingram & List (1987) studied five children learning Quiche, a Mayan language, and found that the Quiche speaking children's early phonetic inventories included sounds which were not acquired until much later by native English speaking children. Similarly Jimenez (1987) found that Mexican-American children acquiring Spanish as their first language acquired /t/ and /l/ much earlier than English speaking children. Anderson and Smith (1987) found that two-year olds learning Puerto Rican Spanish had already acquired palatals, showing much earlier learning than that evidenced by monolingual children learning English.

In summary, cross-linguistic research has highlighted problems with both the universalist and language specific approaches outlined above. Cross-linguistic phonological acquisition research has observed both the cross-linguistic similarities predicted by the 'universalist theories' and also the differences predicted by the 'language specific' theories. A role for both 'universalist' influences and the more stochastic relationship between certain statistical properties of sounds and sound patterns in particular languages and the stages of children learning those languages have both been identified. Many phonologists today accept that both underlying principles of 'language universals' and 'language specifics' contribute in some way. Several theories that fit somewhere between the two extremes have been proposed. One such theory, involving functional load, is now explored in more detail.

3.4 Functional Load

As discussed above functional load provides a way of recognising the importance of phonemes within the phonemic system. According to Grunwell (1980) a child does not acquire separately each of the phonemes of his language; he only uses these speech sounds as the 'phonemes' of the language when he has developed the complete adult system of sound contrasts. Functional load is one way of examining the role of phonemes in the language. The focus is thus on the phonemes as part of a system rather than phonemes in actual language use (i.e. frequency based).

The concept of functional load is well established in linguistics (e.g. Greenberg 1959, King 1967, Wang 1967, Meyerstein 1970). If we accept that the function of a phonological system is lexical differentiation, then functional load offers a way of measuring the phonemes that provide the most and the least differentiation; those that afford the greatest and least degree of contrast between linguistic units. Functional load thus characterises the extent to which a phoneme carries a particular 'weight' in terms of its necessity within a particular phonological system in its role of contrasting between meanings. It can be used to demonstrate the relative importance of phonemes in terms of their 'usefulness'. It is used to "characterise in a general way the extent to which contrasts among members of a set of phonemes or a set of contrastive features contribute toward signalling of significant differences." Greenberg (1959:7).

A phoneme is measured by the number of oppositions or minimal pairs that it occurs in within a given lexical system. Each language has its own phonological system and each language makes use of phonemes within the lexicons of their language in different ways. Whilst languages might indeed share phonemes (as universalist theories observe), each language makes unique use of these in their lexicons, thus resulting in languages having their own functional load rank orders. One prediction on phonological acquisition using the functional load concept as a basis is that children will acquire the phonemes with the highest functional load for their language (i.e. those phonemes that permit the most lexical differentiation), the earliest. The few studies that have used functional load in assessment of adult to child language relationship have had positive results (e.g. Pye et al. 1987). This still remains, however, an area for further analysis of more and different languages and hence the need for this study.

Early functional load research (e.g. Greenberg 1959) measured functional load based purely upon the 'possible' occurrence of a phoneme. These approaches provided information on the relative importance of phonemes in terms of their potential 'usefulness' within words of a particular system but they did not take into account the 'actual' use that the lexicon makes of the phonemic system. Whilst a phoneme combination might be possible it might be rare and possibly too rare for a child to hear in ambient language. Whilst all contrasts are deemed to be equal, certain contrasts are likely to be more important within spoken language than others.

Problems also exist with agreeing what actually constitutes a minimal pair. If we state that a minimal pair exists when one element contrasts (e.g./pæt/ to /bæt/, /bet/ to /bɪt/ or /bit/, /pæt/ to /pæd/) we might define a minimal pair simply as words consisting of sequences of phonemes which differ in only one respect. As has already been discussed in Section 2 the handling of sequences of phonemes that are seen to act as units within a language's phonological system need careful consideration. These directly support what constitutes a minimal pair and therefore what are the significant units within the phonological system for a functional load assessment. As Hockettt (1955) observed consonant clusters which could be argued to be one element of the sequence pose a particular problem; "English 'blown' and 'grown' would not count as

a minimal pair for an investigator whose native language was any of those of the West, but it might well function as one for, say, a Polynesian-speaking analyst". The approach already outlined in Section 2 will be adopted here.

There is also a potential problem when trying to quantify or rank the phonemes against a true reflection of the word base or lexicon likely to be used by speakers of the language. Some researchers (e.g. Greenberg, 1959) have attempted to incorporate a measure of 'use' based upon frequency of occurrence, sometimes termed functional yield. Thus in order to evaluate the functional yield the frequency of occurrence of the phonemes in contrastive minimal pairs is calculated. This is not simply to say that a phoneme that is used more frequently than another is regarded as being more important to the phonemic system than another but enables an analysis of the pairs of contrastive phonemes that are most frequently seen to be present in words. For example, assuming that two phonemes *x* and *y* (e.g. /p/ and /b/) contrast in words *a* and *b*(/pæt/and /bæt/), then the functional load carried by the contrast will reflect the frequency of the words *a* and *b*. The value of the contrast will be greater if both *a* and *b* have relatively high frequencies than if one has a high frequency and the other a low frequency or if both have low frequencies. For example, the /p/ to /b/ contrast would be expected to have a higher functional load when compared to the /p/ to /s/ contrast simply because there are a larger number of words differing only in the presence of /p/ or /b/ than /p/ to /s/ or /s/ to /z/.

One of the main problems with this approach to incorporate a measure of usage based purely upon frequency is what input you take for measurement of frequency of occurrence. The size of the word corpus used as input will determine the frequencies of words that are found and many early studies simply used text sources which may bear little relation to the frequencies seen in spoken corpora. Whilst we can hypothesise about the possibility of a word existing it tells us nothing about the likelihood of that word, that sequence of phonemes, being present in ambient language. With this approach no measure of the frequency of the contrasts is assessed in terms of likelihood of presence in spoken language, for example a contrast might be observed that rarely, if ever, actually occurs in spoken language and be weighted equally with the

another contrast which appears extremely frequently in spoken language. Also, as will be discussed in Chapter 5, using frequency of word types without any regard to overall frequencies carries the risk of moving the assessment further towards a purely frequency based assessment and away from the phonemes systemic usefulness to the phonemic system as a whole.

The measure of 'use', based upon actual data samples of spoken language and combined with the recognition of systemic 'usefulness', as measured with functional load, is scarce. Leinonen-Davis (1987) tested which of the processes observed in children's phonological development (e.g. weak syllable deletion, consonant cluster reduction) produced the most homophony in the lexical systems of children. A small set of children's words were assessed to see which processes (strategies adopted by children when unable to produce the required phonemes) neutralised the most contrasts and therefore most reduced the communicative adequacy of the system. This research, however, looked specifically at the words that children used rather than the words in the ambient language surrounding the child or at the language specific frequencies of particular phonemes observed in adult language.

In a later paper, Leinonen (1990) proposes that those phonological contrasts and combinations of contrasts which have a greater potential for homophony may emerge before those contrasts and contrast combinations which have less potential for homophony. This is in line with what Pye, Ingram and List discussed in 1987 and Ingram goes on to conclude in 1989.

The main benefit to using the functional load approach is that the phonological 'system' is recognised as an influencing factor in phonological acquisition. Different phonological systems offer different phonemic ranks the relative importance of phonemes within that system based upon their 'use' to the lexicon,. The main criticism is that functional load to date has not taken actual language specific features such as usage by speakers or frequency of occurrence.

This study aims to bring together the two influences of usage and functionality in a new method of assessment. The use of corpora of spoken language will directly provide

both the frequency information required and also will enable the 'actual' (as opposed to theoretically possible) minimal pairs of both Finnish and English to be assessed.
The data provided in this study may also be seen to have relevance for assessment and remediation of phonological disorders in children. Grunwell (1985) identified the word for prioritising therapy goals by considering the phonemes usefulness in language. In other words, the question of which contrasts should be introduced into a phonological system before others may be partly assumed by the FUSE data provided in this study.

3.5 Predictions for Phonemic Acquisition Order

3.5.1 Jakobson's Universalist Theory

The child's production of phonemes is the key focus with the universal approach and not how the phonemes are utilised within a system. The words that can be formed from the available phonemes at each stage of development are not explored by Jakobson (these would be language specific anyway) and nor is the fact that the underlying adult form behind the child's realisation may have a role to play in what phonemes are produced. As this study is not specifically looking at the route of acquisition of individual phonemes but instead the use of the phonemes within the system in order to contrast between words Jakobson's theory is only briefly explored in terms of the lexical predictions (i.e. the 'word forms') that might be deduced from the stages of phonemic development. The similarity of the children's usage of phonemes cross-linguistically, as a measure of the universality of usage of phonemes, is proposed as a more rigorous way of testing the universalist argument and this will be presented in Chapter 8. More detail on the background to Jakobson's theory is provided in Appendix 3.1. A simplistic presentation of the theory in terms of the English and Finnish phonological inventories is now defined.

3.5.1.1 Stage 1 - The Minimal System

The principle of 'maximal contrast', see Appendix 3.1, predicts that the first opposition to be observed in the speech of children should be the maximally contrastive 'open' feature of a back vowel contrasted with a maximally 'closed' front consonant.

In English this predicts that the child will develop the back vowel /ɑ/ and one of the bilabials /p/, /b/ or /m/.

The predictions for Finnish consonants would depend upon which of the systems described in Section 2.5 is used. The predictions on vowel acquisition would again be dependant upon the inclusion or not of length as a feature. Utilising the core system of 13 consonants it can be said that the Finnish child will develop the back vowel /ɑ/ (and with length /ɑ:/) and either one of the bilabials /p/ or /m/.

If we assume that the basic universal syllable structure type of CV is the first structure to be applied to phonemes by the child then the following sounds sequences would be predicted;

/bɑ/, /pɑ/ and /mɑ/	for English
/pɑ/ and /mɑ/, /pɑ:/ and /mɑ:/	for Finnish

The common feature of re-duplication in child's' speech patterns would also predict;

/bɑbɑ/, /pɑpɑ/ and /mɑmɑ/	for English
/pɑpɑ/ and /mɑmɑ/ etc.	for Finnish

The next most contrasting feature, noted by Jakobson as 'The first opposition on the axis of simultaneity' (Jakobson 1971:15) is the difference between those consonants (closed sounds) produced orally and those produced nasally. This predicts an opposition will be acquired on the axis of simultaneity of bilabialism between the nasal consonant phoneme /m/ and an oral phoneme, either /p/ or /b/ for English or /p/ for Finnish.

The next opposition to be observed would be predicted to be between labial and dental phonemes. In English this means that the nasal /m/ would be opposed by the dental /n/ or the orals /b/ and /p/ would be opposed by a dental, either /d/ or /t/. In Finnish this

means that the bilabial nasal /m/ would be opposed by the dental /n/ or the bilabial oral /p/ would be opposed by a dental, either /d/ or /t/.

For English the prediction would be;

/mɑ/ and /mɑmɑ/, or /nɑ/ and /nɑnɑ /,

/bɑ/ and /bɑbɑ/, or /pɑ/ and /pɑpɑ /,

/dɑ/ and /dɑdɑ /, or /tɑ/ and /tɑtɑ/.

For Finnish the prediction would be;

/ mɑ / and /mɑmɑ/, or /nɑ/ and /nɑnɑ /,

/pɑ/ and /pɑpɑ /,

/dɑ/ and /dɑdɑ /, or /tɑ/ and /tɑtɑ/.

The order of consonantal acquisition to this point can be summarised as;

[+ / - nasal] > [+ / - labial] > [+ / - dental]

Following these consonantal oppositions, the first vocalic opposition occurs with the narrow vs. wide vowel, the /i/ vs. /ɑ/. This extends the English child's CV sound combinations to;

/mɑ/ and /mɑmɑ/,	/nɑ/ and /nɑnɑ /,	/bɑ/ and /bɑbɑ/,
/pɑ/ and /pɑpɑ /,	/dɑ/ and /dɑdɑ /,	/tɑ/ and /tɑtɑ/.
/mi/ and /mimi/,	/ni/ and /nini /,	/bi/ and /bibi/,
/pi/ and /pipi /,	/di/ and /didi /,	/ti/ and /titi/.

and the Finnish child's CV sound combinations to;

/mɑ/ and /mɑmɑ/,	/nɑ/ and /nɑnɑ /,	
/pɑ/ and /pɑpɑ /,	/dɑ/ and /dɑdɑ /,	/tɑ/ and /tɑtɑ/.
/mi/ and /mimi/,	/ni/ and /nini /,	

/pi/ and /pipi /, /di/ and /didi /, /ti/ and /titi/.

The second vocalic opposition comprises either;

A. narrow vowel to front narrow /i/ and back narrow /u/ vowel.

or

B. narrow vowel to narrow /i/ and mid aperture/e/.

With reduplication the English child's output has now increased to some twenty-four possible different CV sound sequences;

/mɑ/, and /mɑmɑ/,	/nɑ/ and /nɑnɑ /,	/bɑ/ and /bɑbɑ/,
/pɑ/ and /pɑpɑ /,	/dɑ/ and /dɑdɑ /,	/tɑ/ and /tɑtɑ/,
/mi/ and /mimi/,	/ni/ and /nini /,	/bi/ and /bibi/,
/pi/ and /pipi /,	/di/ and /didi /,	/ti/ and /tɑti/.
/me/, and /meme/,	/ne/ and /nene /,	/be/ and /bebe/,
/pe/ and /pepe /,	/de/ and /dede /,	/te/ and /tete/.
/mu/, and /mume/,	/nu/ and /nunu/,	/bu/ and /bubu/,
/pu/ and /pupu/,	/du/ and /dudu /,	/tu/ and /tutu/.

At this stage the child has both the minimal vocalic system and the minimum consonantal system of all the languages of the world.

The Finnish child's output has now increased to some twenty possible different CV sound sequences;

/mɑ/ and /mɑmɑ/,	/nɑ/ and /nɑnɑ /,	
/pɑ/ and /pɑpɑ /,	/dɑ/ and /dɑdɑ /,	/tɑ/ and /tɑtɑ/.
/mi/ and /mimi/,	/ni/ and /nini /,	

/pi/ and /pipi /,	/di/ and /didi /,	/ti/ and /tɑti/.
/me/ and /meme/,	/ne/ and /nene /,	
/pe/ and /pepe /,	/de/ and /dede /,	/te/ and /tete/.
/mu/, and /mume/,	/nu/ and /nunu/,	
/pu/ and /pupu/,	/du/ and /dudu /,	/tu/ and /tutu/.

With the maximal system the consonant /b/ would be included and the CV sound sequences would match the English. With length taken as a feature and the Finnish long vowels included the total number of CV sound sequences would increase to 48.

Having assessed the predictions made by the universal approach it will be interesting to observe the words acquired earliest by the English and Finnish children to see if these predictions are in fact born out by the findings.

3.5.2 Frequency Based Acquisition Predictions

Chapter 2 provided details on the expected frequencies of English and Finnish phonemes which could, with a language specific frequency based approach, be said to predict the acquisition of phonemes by children acquiring these languages. This study will also produce its own frequency rankings for phonemes both overall and by word position with particular focus on word initial phoneme occurrence which can be used to directly compare with the child phonemic usage findings (Chapters 7 and 8).

Using a frequency based language specific approach it might be expected that the earliest phonemes to appear in the child's developing phonological system are those that occur most frequently in the language. For the purposes of this study where the age range being assessed is from 2 to 4 years (i.e. after the acquisition of many phonemes) the frequency of usage by the children will be assessed and compared to the adult language frequency findings. It would be expected that as the children develop towards the adult language the frequency rankings correlate more closely.

3.5.2.1 English Adult Phoneme Frequency Predictions

In summary, as seen in Section 2.4, for English the most frequent adult phonemes overall all word positions were found to be;

/ə, t, ɪ, n, s, d, l, r, ð, k, e, w/

Interestingly, the most frequently utilised adult words 'the, of, and, to, a, in, that, I, it, was' utilise all of these most frequent phonemes and additionally utilise the phonemes /aɪ/and /ɒ/ in word initial position.

3.5.2.2 Finnish Adult Phoneme Frequency Predictions

In summary, as seen in Section 2.7 for Finnish the most frequent adult phonemes overall word positions were found to be;

/i/, /ɑ/, /e/, /s/, /t/, /n/, /o/, /k/, /æ/ and /u/

It will be interesting to observe how these phonemes are used by the Finnish children and whether the most frequent words that they utilise demonstrate these phonemes in use.

This chapter has set out to explore some of the main issues surrounding child phonological acquisition and assessment of the role of ambient language. The main predictions for Finnish and English with both a purely frequency based approach and the Universalist approach of Jakobson have been explored such that later Chapters can assess the relevance and usefulness of these theories in explaining the findings of this study.

Chapter 4 now discusses the implications for using an empirical approach and large data sources for this study and assesses the corpora that are available.

Chapter 4 : Corpus Linguistics

4.1 Introduction To A Corpus Based Approach

According to Samson (2001) a corpus simply refers to a sizeable sample of real-life usage in English or another language, compiled and used as a source of evidence for generating or testing hypotheses about the nature of the language. Naturalness, in terms of spontaneously occurring speech that is a reflection of real life usage, is a basic premise for this study. Collecting naturally occurring spoken language is however a time consuming process. One of the main problems facing researchers has been detailed phonetic transcription. According to Crystal (1987) an hour of recorded conversational data can take 10 or more hours to transcribe, check, edit and type. As noted by Samson (2001) despite speech being a more natural, basic mode of language behaviour than written language much empirical linguistic research to date has focused on written language simply because research depends on the availability of sizeable language samples in machine readable form and it has been far easier to create electronic corpora of written than spoken language.

Four different types of spoken language are required for this study: adult English, adult Finnish, child English and child Finnish. Fairly large amounts of data are required in order to give a representational range of 'word types' for the adult data and in order to represent the child systems. The quality and quantity of input data will be important to this study and will influence the outcome. It was felt that a corpus-based approach, which would enable large volumes of already collected and transcribed naturally occurring data to be accessed, would be the most appropriate approach for this study. Utilising an empirical corpus-based approach will permit a quantitative examination of the phonemic sequences and frequencies observed in the two languages.

Not all four areas being assessed for this study have comparable amounts of language corpora available from which to select data. An exploration of what is available for each of the four areas, bearing in mind the particular requirements of this study, was completed and is detailed below.

Corpus linguistics has played a very important role in moving linguistics forwards from inward looking, intuition-oriented approaches, based upon one or two examples, towards a much larger picture and towards an appreciation of the communicative role of language (Sampson 2001). New patterns have been observed by analysing naturally occurring spoken texts which traditional descriptive frameworks are not normally able to account for or accommodate. Linguists can today accept evidence from the examination of large corpora in a systematic manner. They can use the examples of actual instances to demonstrate theories and they can underpin all of this with the knowledge that what they are citing is naturally occurring and real. Strings of phonemes are directly observable with large transcribed corpora. For this study it is proposed that they can be examined from an empirical and quantitative stand point using data processing methods usually observed in computational linguistics (Church 1993, Garside, Leech & Sampson 1987 and Atwell & Souter 1993).

Sampson (2001) cites the combination of an empirical (rather than theoretical) based approach together with the computational processing requirements needed for large corpora of naturally occurring language data as the key reasons why much corpus based linguistics is today placed within a computer science or computational linguistics framework rather than a linguistics or even humanities framework.

As presented in Section 3 there have been few previous attempts to utilise corpora in the cross-linguistic study of child phonology, as is proposed in this study. This may be partly due to the fact that suitable corpora have not existed nor has it been possible to process large quantities of data systematically and effectively until more recently. However, in order to follow an empirical approach, the processing of large amounts of phonemic transcriptions will be necessary.

Even with corpora containing phonemic transcriptions readily available, the processing overhead for the method of measure being proposed in this study is substantial. For example, as presented in Section 2, the English language has in excess of 90 phonemes and phonemic units that could occur at the start of words (i.e. the 24 consonant phonemes, 39 word initial two consonant clusters, 9 word initial three consonant clusters and 25 vowel and vowel sequences). For the processing proposed in this study

each of these phonemes and phonemic units would have to be systematically replaced with all possible alternative permissible phonemes and phonemic units. In order to check for minimal pairs a database of five thousand words would need to have each word checked after each change (i.e. 5000 x 96 checks). This amount of processing would need to be repeated for each of the four corpora.

A corpus is essentially a record of any naturally occurring spoken or written language. Early language collections were often in the form of citations or dictionaries (e.g. the Oxford English Dictionary) or paper based collections of records of text or speech. With the use of computers much larger samples of data can now be retrieved and stored for reference by multiple users. One of main advantages of modern computer corpora are that once they have been compiled they are readily accessible. Many modern corpora record many millions of words and are accessible directly online. The availability of computerised corpora makes it possible to rapidly and comprehensibly process larger amounts of data with more accountability.

Today many language corpora exist, reflecting a whole range of languages and types of language use. Some corpora are developed for specific purposes (e.g. for lexicographical research where a record of words in written texts might be compiled or for detailed prosodic research as with the Lancaster/IBM Spoken English Corpus), whilst others provide only a general record. Some corpora require specialised technical analysis tools in order to extract the specific requirements whilst others give a simple interface to orthographic representations of words. Spoken language corpora pose some special problems for the researcher (Gibbon, Moore & Winski 1997). In particular, it may be problematic to obtain an adequate transcription.

The identification of word forms and ease of 'reading' are also potentially problematic and thus spoken language corpora remain relatively rare. As Johansson (1995) observes, spoken material is still one of the most difficult types of data to both collect and record requiring as it does conversion to other than orthographic representation to give a phonemic perspective.

The various corpora that are available contain many special coding conventions using characters to represent and distinguish the features that they wish to enable to be drawn out of the specific corpora. Whilst there have recently been some attempts to standardise the presentation of data and codes for future corpora, this is not always the case with existing corpora. Care must be taken to not only select the closest matching corpora for the purposes of this research (i.e. with phonemic transcript, frequency information etc.) but also the coding conventions that have been built for the corpora need to be considered. Some corpora maintain the capital letters of words which start sentences as well as the capital letters of proper names in their orthographic form. Others maintain punctuation marks such as full stops and commas. As none of these features are relevant to this study, they need to be converted to phonemic form or removed from any corpora that is selected. For lexicographical research it may have been considered essential to group word tokens according to their membership of lexemes or lemmas and not by their spelling or pronunciation, whereas this study requires fully inflected forms of words. Each corpora must be therefore individually assessed in light of its suitability to purpose and its representational standards such that the most suitable corpora for each of the four areas is identified.

A summary of the main requirements from a spoken language corpus is now provided such that an assessment of the most suitable corpora can be made.

4.2 Corpora Requirements

"Any corpus must be seen against its intended uses" Johansson (1995).

Whilst using previously collected and recorded data appears to offer many benefits, the selection of corpora needs to be carefully made in light of the specific requirements of any study. This section outlines the main requirements from the corpora that are selected. It also presents some of the problems that utilising previously recorded data sources can present.

The requirements of the four corpora can be summarised as follows;

1. Spoken Language
2. Spontaneously produced / everyday speech (not formal or scripted radio/tv. broadcasts).
3. Phonemic transcriptions (either readily available or possible to produce).
4. Standard/consistent dialect – British English pronunciation for English, a recognised dialect for Finnish.
5. Fully inflected words - (rather than word lemmas).
6. Running text (such that word type frequency can be established) or word tokens with frequency.
7. Large word type range for adult corpora
8. Children's spoken language over a range of ages (i.e. same child at different ages)

As the frequency statistics presented for Finnish and English in Chapter 2 have shown, written and spoken language demonstrate very different characteristics. Not only will different words be represented in the collections of spoken and written language but different types of words reflecting formality, type of language sample etc., will be seen. The identification of corpora reflecting Finnish and English spoken language will be key to this work. Readily available phonemically transcribed spoken language corpora are still more rare than written language corpora (Johansson 1995, McEnery & Wilson 1996) and despite a large amount of spoken language corpora work many of the corpora have been developed as commercial products which are not generally available.

English pronunciation, unlike Finnish, is not phonematic, (i.e. it does not systematically follow a one to one relationship between phonemes and letters). Most orthographic characters can correspond to several different phonemes and characters often combine into sequences to form phonemes making the orthographic to phonemic transcription extremely problematic and as yet incalculable (Berndt, d'Autrechy & Reggia 1994). Whereas with Finnish the orthographic form of the word could be used to deduce the phonemic content, this is not the case with English and the orthographic representation of an English word gives relatively little assistance in arriving at its phonemic representation. In addition to this phonemic transcription is one form of annotation that

cannot be easily catered for by standard computer fonts and programs. Whilst most transcriptions represent the IPA characters needed for phonemic identification (as used in Chapter 2 above) with only basic computer character sets available different transcribers have adopted different transcription standards.

The most frequently produced words of English and Finnish adult language corpora will be taken by this study as the most representative words of naturally occurring speech. Word frequency information is readily available for a number of very large English language corpora and in a range of computerised dictionaries such as Collins COBUILD English Dictionary and the Longman Dictionary of Contemporary English. Care needs to be taken, however, in using this readily available frequency information which may use a written word frequency count or simply represent the word forms differently. For example, the COBUILD database has the 15k most frequent words identified and subdivided into five frequency bands or ranges whilst the LDOCE has the 3k most frequent words for speech and writing respectively, subdivided into three bands of 1k items each.

Computerised dictionaries often provide only one entry for a lemma, thus the frequencies of individual word forms are lost (e.g. all tenses of a verb are put under one frequency of the root form of the verb, singular and plural are put together etc.). With this approach one frequency entry would exist for several differently pronounced words (e.g. 'to go' would include 'went' /went/, 'going' and 'gone', one entry would exist for 'mouse' to include the differently pronounced words 'mice' and 'mouse' etc.)

The frequency information required as a basis for the selection of word types ideally has to relate to the pronunciation rather than the spelling of words. Sequences of phonemes that are identical (homophones) might not be represented in the same way orthographically (e.g. 'too' and 'two' both pronounced /tu/, 'eye' and 'I' /aɪ/) and the frequency count might therefore relate to the spelling and not the pronunciation. This study needs to regard homophones as one word type, as they have the same sequence of phonemes in the same order. For example, the words "buy', 'bye' and 'by' should be represented by one word type, /baɪ/. In order to capture such naturally occurring

homophony, the phonemic representations of word forms would need to be sorted by phonemic representation into like-types and the frequencies of the individual words added together to give a frequency by phonemic word type. In order to maintain the word type range and not to skew the FUSE rankings with naturally occurring homophony, identically pronounced word forms (representing naturally occurring homophony) will have to be extracted before any FUSE processing begins (Chapter 5).

Looking at this potential problem in frequency representation from another perspective, words that are spelt the same might not be pronounced the same. Homographs, that is words with identical spellings i.e. orthographic representation, but with two or more different pronunciations need also to be taken into account. Examples in English include the words 'abstract', which can be pronounced either as /æbstrækt/ or /əbstrækt/, 'minute', which can be pronounced either as /mɪnɪt/ or /mainut/ and 'present', which can be pronounced either as /prɛzent/ or /prɪzent/ depending on context. In order to capture all of the homography in corpora the various phonemic representations for same spelt words would be required together, ideally, with a frequency count relating to that specific pronunciation of the word. These phonemic representations might in turn form more naturally occurring homophony (i.e. identically pronounced words) where one variation of pronunciation is identical to another differently spelt word.

Whilst not a problem for this study (where the pronunciation is the prime identifier of a new word type) the handling of homonymy (several words with same meaning) and polysemy (one word with several meanings) present a severe problem for research which is needing to find or have identified underlying meaning. For example, COBUILD has a word entry 'buck (-s, -ed, -ing)' which is marked within the 7k most frequent words. With reference to the COBUILD CD-Rom, however, it appears that at least 25% of the total are for 'Bucks' (the shortened form of Buckinghamshire) leaving only half of the recorded instances to be subdivided between some 15 noun and verb senses, the predominant one being 'a fast buck'. Variation in meaning and grammatical usage are not the focus of this study. However, it must be born in mind that in order to get around the problem just identified some corpora have more than one entry (and

frequency) for words with the same pronunciation and spelling with one entry (and frequency) for each meaning difference. This study is interested in the pronunciation of words to form the list of word types. The frequencies of the various word types must be based upon the phonemic representation of the words (rather than the orthographic representation).

It is recognised that as well as finding a corpus of spoken language for each of the four areas, the adult corpora which provide the basis of the FUSE method of assessment for the two languages should be offering a large enough range of words for analysis so that the phonemes and phonotactic possibilities of the two languages can be represented. The size of the two adult corpora will therefore be of key importance to this study. The definition of 'large' will be governed to a certain extent by the corpora that are available. The relationship between the size of the corpora selected (in terms of running text) and the frequencies at which these words are observed are found to follow certain predictable rules (see Chapter 2). Even the largest corpora of running text often provide relatively few different word types. The requirement for a large range of different words could easily be satisfied by simply using all word tokens, the vast majority of which can be predicted to be hapax legomena word types (occurring only once). The benefit of having more word types to process against using words that appear only once and may not be good indicators of 'ambient' environment are weighed during the selection process.

Corpora that present a wide range of word types needs to be selected. The limited choice of Finnish corpora to a certain extent determine the actual size of English adult word bank required. In addition whichever adult corpora are selected should include the speech from of a variety of adult age groups and from both sexes. The corpora should ideally be sufficiently large enough that even after the removal of hapax legomena forms a wide range of word types remain, sufficient to represent the phonemic status within the language.

Whichever corpora are selected need to be readily accessible and available for repeated investigation, if required. Some of the corpora under examination will have readily available and already developed tools for the processing of the word forms. Whilst it is

not considered a criteria for corpus selection, these tools may prove useful for the initial data extraction where data will be presented in a particular format. It is anticipated that new procedures will need to be developed in this study in order to bring the four data sources to a consistent level of both frequency and phonemic detail.

Both Finnish and English have many dialects each with their own particular pronunciations of words and vocabulary (Wells 1986 and Paunonen 1994, 1995). For the purposes of the research it is important that a spoken language corpus that reflects one dialect form is found in order to give a consistent phonological structure and word base for frequency analysis. The approach taken towards the English database whereby orthographic word forms can be converted to a standard dialect, RP for the case of English, cannot be applied with Finnish where the orthographic representations of words will directly supply the phonemic transcriptions. For Finnish a corpus that reflects one dialect with sufficient numbers must be selected and it will be the word forms themselves seen in the transcription that will determine the correct pronunciation.

It will be important that the 'word form' as it actually appears in spoken language is recorded in the corpus. Finnish as an agglutinative language has many variations for one lexical root; taking the lexical root of the word would not provide the sequences of phonemes actually spoken. It is recognised that certain affixations may be more common than others but these reflect a syntactic rather than phonemic characteristic of the language.

In order to represent the true combinations of phonemes as they occur in spontaneous speech it is important that compound words and affixes are present in the word forms. Taking the lemma of these word forms would not provide an accurate measure of phonemic frequency in speech. As with the English database, it is essential to establish the actual word form spoken by the speaker, that is the actual sequence of phonemes e.g. when two words normally separated in written text are joined together and said as one in speech then one word form is recorded.

4.3 English Corpora

Whilst there are many English corpora from which to select word types, and more appear each year, very few corpora are both readily available and represent naturally occurring British English (as opposed to American English) spoken language in a form suitable for phonemic analysis. A selection of accessible English corpora available at the start of this study were assessed by the current researcher for suitability and are presented in this section.

4.3.1 English Adult Corpora

4.3.1.1 The Brown Corpus (1984)

One of the earliest attempts to produce a corpus of the English language was that developed at Brown University by W. Nelson Francis and Henry Kuçera during the 1960's (Atwell & Souter 1993). The corpus is a collection of one million words of written American English and was released in 1967 as the first computer corpus of American English.

4.3.1.2 Lancaster-Oslo/Bergen (LOB) Corpus

This corpus was developed as a direct British English equivalent to the American English Brown corpus described above. Collection of a one million word corpus of written British English commenced in 1970 by Geoffrey Leech at Lancaster as the CAMET (Computer Archive of Modern English Texts) project. The project was completed in 1978 in Norway with Stig Johansson of Oslo and Knut Hofland of Bergen. Analysis tools were developed using this corpus including analysis tools for tagging part of speech and word-classes known as CLAWS (Constituent Likelihood Automatic word-tagging System).

4.3.1.3 Survey of English Usage (SEU)

In original form this paper-based corpus represented one of the earliest attempts to compile a corpus of British English. The project commenced in 1959 at University College London as the Survey of English Usage (SEU) and approximately 1 million words were collected, half in written form and half in spoken. The spoken half now constitutes the London-Lund Corpus (see below). The orthographically transcribed spoken half and the written half were both originally filed and analysed on paper.

4.3.1.4 London Lund Corpus

In 1975 the Survey of Spoken English (SSE) was set up at Lund university in Sweden by Jan Svartvik initially to make available in electronic, machine-readable form the spoken part of the SEU corpus. The original 87 texts of transcribed speech totalling some 435,000 words were supplemented by 13 more texts to produce the complete London-Lund corpus (LLC) of one hundred 5,000 word texts (Svartvik 1990). The total of approximately 500,000 words was until recently the largest and most widely used freely available electronic corpus of spoken British English (Kennedy 1998).

The LLC has been used by researchers in many countries for studies of phonology, lexis, grammar and discourse structure and function. The main criticisms of the corpus (e.g. Kennedy 1996) are that the texts recorded were of speakers who were predominantly highly educated adults such as academics and that the range of genre categories is narrow.

4.3.1.5 The Lancaster / IBM Spoken English Corpus (SEC and MARSEC).

This is a corpus of modern spoken British English. Produced as a joint venture between Gerry Knowles of Lancaster and IBM (UK) it was completed in 1987. The corpus consists of approximately 53k words of the spoken standard British English (RP) of adults sampled between 1984 and 1987 from eleven categories mainly radio broadcasts including unscripted radio news broadcasts, university lectures, broadcast fiction and poetry and dialogue. As such it represents mainly formal speech. MARSEC is a

machine readable version of SEC with prosodic transcription. MARSEC was specifically designed for, and used most often for, detailed prosodic research and as such it shows features of stress, intonation and pauses.

4.3.1.6 Birmingham Corpus

This corpus of both spoken and written British English was released in 1988 (McEnery and Wilson 1996). It consists of approximately 20million words, of which 1.5 million words are spoken language.

4.3.1.7 The British National Corpus (BNC)

This corpus was created as a resource for research on the British variety of the English language and was compiled between 1990-1994 and released in May 1995. It was a collaborative national project involving the British Library, three publishers (Longman Group Ltd, Oxford University Press, Chambers-Harrap) and two academic institutions (Oxford University and Lancaster University).

This very large corpus of approximately 100 million words includes 90 million written words and 10 million words from spoken language. It includes a comprehensive range of text varieties. It is marked up according the guidelines of the Text Encoding Initiative and tagged with parts of speech, mainly done with Claws (see above).

4.3.1.8 MRC Psycholinguistic Database

The MRC Psycholinguistic Database, strictly speaking, is not a corpus. Rather than a record of running text it is a collection of previously compiled information brought together to make psycholinguistic analysis more feasible The words themselves come from a variety of different sources including the Oxford English Dictionary, the Edinburgh Associative Thesaurus, the Oxford Advanced Learner's' Dictionary. It is the provision of up to twenty six different linguistic properties (including frequency, psycholinguistic and phonetic features) which are provided for these words that makes the dictionary particularly useful for this particular research.

The original version of the database, provided as an on-line service in 1981, simply contained a dictionary of words and word association norms from the Edinburgh Thesaurus (Coltheart 1981). The second version, now available to language researchers as a computer usable resource, contains an updated and enlarged version of the word dictionary together with programs for accessing the dictionary (see Mitton 1986). As a database it contains information about approximately 150,000 English words (Wilson 1987). It differs from other 'dictionary' type corpora in that not only does it provide syntactic information (such as tagging codes for parsing) but also psychological data such as age of word acquisition for psycholinguistic analysis.

The features of the database that will be particularly relevant to this study are now presented.

The database meets the two basic requirements of

a. representing the frequencies of spoken naturally occurring language with
b. phonemically transcribed representations of words

The frequencies of both written forms of words and spoken forms of words are provided in the dictionary. The spoken frequency count is taken directly from a sub-set of 200,000 words of the spoken language recorded in the London Lund Corpus of English Conversation (see above) as produced by Brown (1984). This records frequencies for approximately 11000 different word types. The phonemic transcriptions are from Daniel Jones' Pronouncing Dictionary of English Language, 12th edition (1966) and have been converted to the U.K. Alvey standard for machine readable phonemic transcription (Wells 1986).

The dictionary file currently occupies 11MB as a sequential UNIX file, each line of the file representing the field for one word. The longest entry is 130 characters with a fixed composition of columns relating to different attributes. This includes the Brown Frequency ranking, if it exists for a certain word (position 22 to 25 of the field) and the phonemic transcription, again if it exists (at a position after 51). The dictionary file

contains 14,529 frequency entries for 8985 different word forms. The range of entries is from 0 (not recorded in the London Lund Corpus) to the maximum frequency of 6833 for the word 'the'. The mean frequency recorded is 35.

Words which have the same spelling but different pronunciations and syntactic classes are marked for identification (with 'O' in the word field) as are words with different pronunciations (with 'B' in the word field). Thus the various pronunciations of the same orthographically represented words can be ascertained. The frequency recorded on this dictionary file records spoken frequency for the orthographic form of the word and not for each of the pronunciation variants.

4.3.2 English Child Corpora

4.3.2.1 CHILDES

The CHILDES (Child Language Data Exchange System) Project (MacWhinney 1995) has established a uniform database of data on language development. The database is supplemented by a system for transcription and a set of programs for analysis. As such, the CHILDES project today represents the best electronic corpus for child language analysis.

CHILDES was originally conceived as a way of facilitating child language research. One of the main problems with all spoken language corpora is the collection and transcription and the collection of child data samples is extremely time consuming. CHILDES was set up to share data that has originally been collected for specific research projects in order to answer questions relating to both the universal nature of acquisition and the individual characteristics of particular children. Questions of the universal nature of acquisition such as whether all children babble, whether there are features of phonological acquisition that are common to all children or whether early word order is universal can only be asked with large data samples of many children. Questions relating to particular styles of language use need good data. In order to address such questions, data samples of many children of different ages and different languages are needed.

Data that had been collected for specific research projects were commonly transcribed to the standards of the particular research project and this made the sharing of data across researchers extremely difficult. MacWhinney recognised that computers would make it easier to duplicate, edit and analyse data from various sources in a consistent and repeatable way. However, one of the first tasks that Childes had to overcome was the standardisation of this data transcription process, so that data could be shared not only between researchers of the same language but also cross linguistically. Once data conformed to a recognisable format it could be collected together. Work on common tools for analysis could then begin such that individual program writing for specific needs could be replaced by a common, shared pool of tools which could be applied consistently to data from different sources and representing different languages.

In 1981 the original vision of Slobin, Levelt, Ervin-Tripp and MacWhinney was to produce "an archive for typed, handwritten and computerised transcripts" (MacWhinney 1995). Data entry commenced in 1984 at Carnegie Mellon University, USA and in 1994 internet access with free download of both data and tools for analysis was provided.

CHILDES today comprises 3 main elements;

1.A large database of spoken language of many different children including

- Spontaneous speech in 26 different languages
- Narratives
- Bilingual Acquisition (Dutch/English, Spanish/English, Chinese/English, Japanese/Danish, Spanish/Catalan)
- Impaired/Clinical
    - children with language impairments
    - adults with aphasia

2. A set of 36 tools for analysis known as CLAN (Computerised Language Analysis).

These tools, which are written in C, provide an automatic means of analysing data which has been built using the CHAT transcription codes. Facilities such as frequency counts of words, word searches, co-occurrence analysis, mean length of utterance counts and interactional analysis can be performed.

3. A transcription tool known as CHAT (Codes for the Human Analysis of Transcripts).

   The coding system enables phonemic representation of words, the coding of particular grammatical categories of utterances and a range of different linguistic features to be represented on the database files.

It is estimated that more than 60 research projects were underway in 1999 using the CHAT transcription code and more than 600 research projects had used CHILDES data up to that date (MacWhinney 1999).

English child data is represented in more than 200 different data files on CHILDES. Appendix 4.1 provides a summary of the main English child data files on CHILDES. The data files are grouped by name of researcher and a summary of the ages and numbers of the children contained in the various data files is given.

The selection of the most appropriate data file to use in this study relied on finding a sample of spontaneous British English child speech that demonstrated the same group of children's development over a time period of some years. The children should be speaking British English and if possible a phonemic transcription should be available.

4.4 Finnish Corpora

4.4.1 Finnish Adult Corpora

At present there is only one electronic corpus that exists for spontaneous adult Finnish speech, the Helsinki Spoken Language Corpus (Helsingin puhekielen aineisto).

4.4.1.1 The Helsinki Spoken Language Corpus

The Helsinki Spoken Language Corpus was collected during the period 1972-1974 by the Finnish Language Department of the University of Helsinki under the leadership of Terho Itkonen. Its primary objective was to study the differences observed between written Finnish and the particular dialect of Finnish spoken in Helsinki. Research during the 1970's was headed by Heikki Paunonen and looked specifically at Modern Finnish dialectic distribution.

The data was collected by recording informal conversations of 126 different Helsinki residents and it contains a total of approximately 600,000 running words. The corpus consists of approximately 60 hours worth of transcribed speech with each speaker providing approximately 30 minutes of speech. The data represents an even distribution in terms of age, sex, social class and home area; the 59 male and 67 female speakers come from two different areas in Helsinki (Töölö & Sörnäinen), from three different social classes (upper, middle and lower) and three age groups (15 –20 years, 40-45 years and 65+years).

The data is stored in the Multilingual Databank of the Department of General Linguistics, University of Helsinki, and is now accessible in computer form to authorised users working at the University of Helsinki. There is a file for each of the interviews, recorded with a code name that reflects the home address, gender, social status, age and name so that particular individual conversations can be accessed. Alternatively, and for the purposes of this research, the entire set of files can be accessed and processed together.

4.4.2 Finnish Child Corpora

There are no publicly available corpora or collections of spoken Finnish child data available in an existing electronic format. Despite the existence of other Finno-Ugric child language samples on CHILDES, such as those for Hungarian and Estonian, no samples have been presented for open use from Finnish acquisition studies.

Much research on Finnish child language has followed the case study approach with the individual research projects recording individual children's or small groups of children's acquisition (e.g. Savinainen-Makkonen 2000 and Kunnari 2000). The transcripts of many individual case study based research therefore exists, some in electronic form others paper based, however none are freely accessible. The only large collection of Finnish language acquisition data is that produced at Oulu University during the 1970's and this is mainly paper-based and is not freely available. In order to assess the suitability of this collection of Finnish language acquisition transcripts for this study, a visit to Oulu University was made by the current researcher.

4.4.2.1 Oulu Finnish Child Data Sample

The Oulu data sample (*Oulun yliopisto, Suomen kieli, Lapsenkielen materiaali* 'Oulu University, Finnish Child Language Material') contains material collected from the 1970's onwards. Today approximately 350 data samples for approximately 50 different children exists.

The transcripts reflect spoken, naturally occurring, spontaneous speech between parents/caregivers and children. Most of the children were from the Oulu area of Northern Finland and the Oulu dialect is therefore a feature of this collection of transcripts. Other '*lapsen kotipaikka*' 'child's area of residence' have been recorded as Merijärvi, Karunki, Reisjärvi, Kiiminki, Kempele, Hailuoto, Utsjoki, Kello, Ylikiiminki, Tampere (mid Finland), Aavasaksa and Espoo (near Helsinki in Southern Finland). In the main these are places in Northern Finland. The ages and names of the children have been collected together with the total lengths and dates of recordings such that the individual progress of particular children over time can be observed. Recordings from birth up to age 17.1 exist. However not all the children have been recorded at all ages.

Appendix 4.2 provides details of a selection of Finnish child data files from the Oulu data sample that were assessed for suitability for this study. The few files that had already been transferred to electronic format are identifiable in Appendix 4.2 with a file number. The majority, however, were paper based and required conversion to

electronic format by a process of scanning, selection of the child utterances and creation of word lists.

The preceding chapters have set out to provide a basis for the main objective of this study that of applying a new method of assessment of the relationship between child and adult phonological usage. Having established the phonemic inventories of the two languages Finnish and English and assessed the likely child development patterns that might be expected from the two languages following traditional approaches an attempt will be made to involve a measure of 'use' based around the empirical evidence of language as seen already existing corpora for the two languages.

The next chapter sets out the main basis of the new method of assessment before the detailed data processing is presented in Chapter 6.

Chapter 5 : Approach & Methodology

One objective of this research is to develop a new way of examining the relationship between ambient language and phonological acquisition. The phonological and phonotactic features of English and Finnish have been explored (see Chapter 2) such that an assessment of the relative systemic importance of particular phonemes and phonemic units can now be attempted. The FUSE method of assessment aims to combine this knowledge of the phonological and phonotactic systems of Finnish and English with the demonstrable use that speakers of the language make of these features, as seen in spoken language corpora selected for the processing (see Chapter 4).

The objective will be to develop a method of assessment which enables the phonemes and phonemic units to be ranked in terms of their relative systemic importance in spoken language. It must be designed in such a way that it can be applied to any language (to provide consistency in cross linguistic studies) and to both adult and child languages (so that the effect of ambient language can be observed on the child's developing phonemic system). The FUSE method of assessment will be achieved by applying the principle of functional load, a phoneme's potential importance to the language, and additionally incorporating a measure of use by utilising only the most frequently spoken words of the languages from the corpora. The development of the FUSE method of assessment is discussed in Section 5.1.

The FUSE method should enable a new way of exploring and observing the language specific and universal features of the relationship between the acquisition route adopted by children acquiring a particular phonological system and the adult language they are acquiring. The way that the FUSE findings will be used to assess this relationship is discussed in Section 5.2.

One additional result of analysing the spoken language corpora at a detailed level will be frequency statistics on the structure of words and syllables and the frequencies of particular phonemes at particular word positions. As much of these findings will be new, particularly for Finnish, these will also be recorded.

This study will be assessing whether children are alerted to the relative importance of phonemes within the phonemic system (i.e. their systemic usefulness) rather than frequency of occurrence of word tokens. The assessment of frequency for phonemic usage therefore needs to be looking at word types rather than tokens (i.e. are children somehow alerted by the presentation of a new word type to a phonemes ability to contrast within the system). The phoneme frequency information found in this study will therefore be very different to that previously presented based upon running texts. These phoneme frequency and word structure findings for word types of spoken English and Finnish will be compared with the frequency assessments presented in Chapter 2.

5.1 A Method For Measuring Functional Use (FUSE)

Two basic premises underlie the FUSE method of assessment;

Firstly, it is taken that for the purposes of this study, the best representation of ambient language that would be expected to be surrounding a child is naturally occurring, spontaneous, spoken language. It directly demonstrates the actual usage that speakers of a language make of the phonemic system of their language. The spoken representation of these words (i.e. the phonemic transcriptions), directly represent both the use of the phonemic system (which phonemes are being used to signal contrasts) and the phonotactic structures (which phonemes are being used in which syllable and word positions). Observing large quantities of spoken language provides a measure of the relative frequencies of particular word types and enables the most frequent word types to be extracted for further processing. It is these word types that are of direct importance to this study as the word types demonstrate the systemic usefulness within the phonological system of particular phonemes (i.e. their ability to provide contrasts) and also enables the relative frequency of phonemic usage and different word structures to be observed.

Secondly, it is accepted that certain phonemes within the phonemic system are systemically more important than others as they signal more meaning differences within the language. That is to say they occur in more minimal pairs as the

differentiating element or contrast between word types. Functional load as a concept (see Chapter 3) is generally applied only in theory to the phonotactic possibilities of a language. The measure of functional load would test a phonemes 'potential' ability to contrast between potential words of a language based upon meaning difference. As discussed in Chapter 3 few studies have actually applied functional load to large language samples of spoken language.

The method of assessment proposed here aims to combine these two elements; the potential 'usefulness' of a phoneme to the phonemic system (as assessed by functional load) with the use that speakers of a language typically make of that phoneme as seen in the word types they elect to produce.

The processing approach will be to systematically reduce the phonemic systems of the two languages. Each reduction will constrain the phonemic system to a greater or lesser degree, resulting in a loss of oppositions or contrasts. The effect of this loss is the ability to contrast between words, the loss of a meaning difference. With this study the effect of a loss of contrast will be measured as the number of identical strings of phonemes that are produced. For example, the English words 'cat' /kæt/, 'mat' /mæt/ and 'hat' /hæt' differ only in their first phoneme /k/, /m/ or /h/. In this example, these three phonemes are working as oppositions within the sets of minimal pairs ('cat' and 'mat', 'cat' and 'hat' etc.) to provide a contrast and to differentiate between the three words of the language. Without this ability to differentiate between the three words (i.e. without the contrasts of /k/ to /m/ or /k/ to /h/ or /m/ to /h/) pairs of identical word forms are seen to exist and a meaning distinction is lost.

Section 5.3 discusses in detail the processing approach adopted for the FUSE method of assessment and how the data processing developed from an approach that individually reduced phonemes and compared minimal pairs (as with functional load) to one whereby all word initial positions could be systematically and simultaneously reduced and the resultant homophony could be measured. For example, with the word initial phonemes reduced from the phoneme sequences /kæt/, /mæt/ and /hæt/ the string /*æt/ (* represents the contrast) would be presented three times. In other words,

identical phonemic strings, where previously different word types existed in the data, will be produced.

This research aims to assess the relative importance of phonemes for differentiating between word types. Words are seen as strings of phonemes, differing to a greater or lesser extent by the number of phonemes present, the actual phonemes present and the order of the phonemes. The frequency of occurrence of particular word types, whilst of importance for word frequency assessments (see Chapter 2) and the initial selection of the most frequent words for processing, is not then directly relevant for the systemic view of phonemic usage. Utilising word types will enable an assessment at the phonemic systemic, rather than frequency, level.

Phoneme frequencies reported in this study relate, therefore, not to the actual frequency of occurrence, as with word token assessments, but to the usage by the phonemic system (i.e. number of word types that make use of the particular phoneme or phonemic unit). The contrastive usage of particular phonemes for the two languages can be measured by assessing the minimal pairs that are found within the word types. As discussed in more detail in Section 5.3 word types will be grouped into minimal pair groups by word type length. Within each of these minimal pair groups strings of minimal pairs will demonstrate the contrastive phonemes. Each language will therefore have a group of minimal pairs comprising a number of minimal pair strings and within these strings a set of contrastive phonemes or phonemic units. All of this assessment will rely upon having a range of word types, rather than word tokens, from which to form the minimal pair groupings.

As discussed in Chapter 2 word forms of identical phonemic sequences which in effect have the same phonemic form despite having different orthographic spelling and meanings are known as homophones (Jackson 1988) and both Finnish and English demonstrate naturally occurring homophony (e.g. in English 'where' and 'wear' which are both pronounced /weə/, 'two' and 'too' which are both pronounced /tu/.

Reductions to the phonemic systems of Finnish and English will effectively mean a loss of ability for the languages to signal differences between words and an increase in the

amount of homophony within the system. The amount of resulting homophony is something that can be measured.

In summary, reductions to the phonemic systems of the two languages will result in a loss of differentiation between words and a reduction in the ability of the language to signal word differences through the phonological system. This assessment of the loss of the language to signal word differences will be measured by homophony. This assessment applied therefore to spoken language will recognise both a phonemes potential usefulness within the language to signal meaning differences and the actual use that the language makes of that facility.

Each reduction to the system will therefore result in a measure of that particular phoneme's functional use within the system. The functional use of each phoneme can then be compared with other phonemes' FUSE and a rank order of the relative importance of the phonemes to the system as a whole can be built.

One approach would simply be to reduce the phonemic system by deleting specific phonemes from the system wherever they occur (i.e. in all word positions) and then looking at the overall impact. It has already been suggested that word position affects a phonemes' functional load (Leinonen-Davies 1987). It is felt therefore that a systematic reduction of the phonemic system by particular position within the word or syllable (e.g. word initial syllable initial, word final syllable final) will provide a more detailed view of how a particular phoneme's importance to the system varies by word or syllable position. This study will look to reduce the phonemic system by deleting the phonemes seen at the start of words and then assessing the amount of homophony produced. It is accepted that word position may have a cross linguistic significance (i.e. one language may be found to present more homophony with the word initial phonemes reduced and another with the word final phonemes reduced). However, testing this any further is outside the remit of this study where only phonemes at the start of words are considered. Simply measuring the resultant homophony overall would only provide an indication of a phoneme's importance overall rather than a specific picture of how the phoneme was operating within and contributing to the

system. The approach will be therefore to systematically reduce the phonemes by word position and measure and record the resultant homophony by word position.

In order to assess the functional use of the various phonemes it will also be necessary to recognise there are certain characteristic features of the two languages which will affect the amount of resultant homophony. As has been discussed in Chapter 2, there are phonemic units that are observed within the two languages that appear to operate in contrast with singleton phonemes in minimal pairs, serving to differentiate between words. These phonemic units need to be treated as phonemes during the reduction process so that the units as a whole are systematically reduced from the system. The definition of 'phoneme' and the inventories of phonemes for Finnish and English given in Chapter 2 recognise these phonemic units and hence they will also be recognised during the phoneme reduction processing carried out in this study.

With the approach adopted in this study where the word position is to be taken into account during the phoneme reduction processing the phonotactic rules for word formation for the two languages must be considered. Special care needs to be taken in identifying the phonemic inventory to be reduced (including phonemes and phonemic units) within the context of the phonemic system of the language in question (i.e. by word position). For this initial application of FUSE where only word initial position is explored the phoneme reduction processing must therefore observe the word initial phonotactic rules of the phonemic systems.

As has been discussed in Chapter 2 both Finnish and English include digraphs (two transcription symbols) to represent one phoneme. In English the digraphs of affricates, diphthongs (e.g. /tʃ/, /ɪə/) and the trigraphs of triphthongs (e.g. /eɪə/) and in Finnish the digraphs for long vowels, geminates and diphthongs (e.g. /ɑɑ/, /kk/, /ou/) must be converted to be represented by one symbol so that the systematic reduction of phonemes includes those phonemes represented in corpora as digraphs or trigraphs. Each of the four areas of analysis presents its own requirements therefore for pre-processing as discussed in detail in Chapter 6.

This research aims to assess the relative importance of phonemes for differentiating between words. Words are seen as strings of phonemes, differing to a greater or lesser extent by the number of phonemes present, the actual phonemes present and the order of the phonemes.

Two methods for calculating FUSE will be attempted. Firstly, the FUSE assessment for a particular phoneme will be calculated by simply counting up the number of actual words that it provides the contrast in. Secondly, the count of contrastive phonemes in a minimal pair string will be used to give an indication of the work that a particular phoneme needs to do to provide the contrast. Both of these approaches to calculating FUSE are similar to functional load processing in recognising the importance of the phonemic system as a whole. The FUSE calculation differs, however, by directly including a measure of word and phoneme usage by taking only the most frequently spoken words rather than all possible words.

An alternative approach to this might be to use the token frequency count as a direct input to the FUSE calculation such that when a contrast is measured, say the phoneme /s/ in word initial position, it has the overall word frequencies involved in the minimal pairs taken into account. For example, the minimal pairs 'sat' and 'mat' would each have their relative word frequencies, as taken from the frequency count for the word form, included. With this approach the word initial /s/ phoneme count would be increased by whatever the frequency measure of the word 'sat' was rather than simply by the count of one indicating one word type. This approach would move the FUSE calculation much closer towards frequency based results. It would recognise frequency as a direct contributor to FUSE however its application would be problematic (see Section 5.4). It may also skew the FUSE results too closely towards frequency unless specific measures to integrate frequency in a principled way, rather than simply take the overall word frequency into account, were developed such that the weight of frequency and systemic usage were balanced.

The end results of this processing will be FUSE totals for each contrastive phoneme which will enable the phonemes to be ranked according to number of words they provided the contrast in. The FUSE rankings will include FUSE assessments for all

phonemes that are accepted as operating within the phonemic system, as well as FUSE assessments for the agreed phonemic units of the languages in question. The FUSE processing will be completed for both adult language corpora such that an assessment of 'ambient language' FUSE for each language can be ascertained. The same method of FUSE calculation that will be applied to the adult language can be directly applied to the smaller child data samples. The application of the same assessment for English and Finnish child language corpora at various stages of development, will provide data on how the FUSE rankings may or may not change during the acquisition process and at which stages of development different phonemes appear to be offering the most contrast within the child's developing lexicon.

5.2 Process for Correlating the Results

Having completed the application of the FUSE method of measurement for the English and Finnish adult and child data samples the next stage will be to assess how closely the children's developing phonemic systems correlate with this measure of the adult languages that surround the children as ambient language.

This can be done in two ways. Firstly, by simply comparing the adult FUSE rankings with the child frequency rankings and secondly, by comparing the adult FUSE rankings with the child FUSE rankings for the two languages.

The children can be observed both as a group, with their word types combined together to give a large word type base of unique words, and as individuals presenting their own word type base. The results of the group correlations are presented in Chapter 8 and for the children individually in Chapter 9.

The calculation of FUSE rankings for the child and adult data and the two languages will enable the following investigations to take place;

1. Assessment of the correlation between the adult FUSE rankings and the individual and group child word type word initial phoneme frequency rankings over the three ages for each language.

2. Assessment of the correlation between the adult FUSE rankings and the individual and group child FUSE rankings over the three ages for each language.

3. Assessment of the correlation of frequency of word initial phonemes between adult and child language.

4. Assessment of the correlation between the rankings cross-linguistically.

A close correlation between the adult FUSE rankings and the children's frequency rankings may be indicating that the FUSE results may tell us something about phonological acquisition routes for the children acquiring those languages. A low correlation at all of the three ages of children under observation may be indicating that there is actually little relationship between the ambient language FUSE ranking and child phoneme frequencies. Knowing that a particular phoneme is ranked high in word initial position according to FUSE, we can predict that as a relatively important phoneme to the phonological system and one frequently observed performing the contrast of words in spoken language, this phoneme would be acquired in word initial position before those with low FUSE adult ranking. Therefore we would expect the frequencies of word initial phoneme usage by the children to reflect the adult FUSE rank orders. A close correlation between adult FUSE rankings and child frequency rankings would indicate that the child is selecting to utilise those phonemes that do appear to be most important at signalling differences between words in ambient language. A movement towards the adult FUSE rankings might be expected over the three ages (as the children's phonological acquisition moves towards the adult system). It will be interesting to observe the differences between the languages that may be highlighted with this new assessment. One language may show more movement over the three ages or may start much closer to the adult rankings than the other initially.

The second correlation assessment will be between the adult FUSE rankings and the child FUSE rankings. The adult FUSE rankings when directly compared with the child FUSE rankings will enable the observation of the two systems in use. It will be possible to observe at which stage of development the FUSE rankings start to correlate

significantly with the adult rankings. An initial significant correlation between the adult and early child rankings of the same language will demonstrate that the lexical/phonemic structures seen in the adult language are also reflected in the child's early language use i.e. ambient language does have an early and measurable effect on the child's acquisition. A move towards the adult system would be predicted as the child's phonemic system develops and comes to more closely resemble the adult system.

Utilising the FUSE rankings for both the adult and child systems this movement can be observed at the systemic level (i.e. how close are the two systems of usage as opposed to frequency of usage) and the rate can be clearly identified in the correlation findings. If a correlation is established it will then be interesting to observe the rate at which the child FUSE rankings move towards those observed for adult speech, how quickly the ambient language system intercepts any possible universal basis. It will also be interesting to observe the differences between the two languages. A low correlation across all three ages for one language might be providing a clear indication that the adult language system has yet to be a feature of the child language system for that language. If a correlation is established between the adult and child of same language and for both languages with their different phonological systems then this may provide evidence that contrast usage is a driving force behind phonological acquisition.

The detailed frequency information gathered for both English and Finnish, child and adult data sources, will enable an assessment of the relative closeness in terms of word type word initial phoneme frequency between the adult 'ambient' data phoneme frequencies and the children's developing systems. A comparison correlation assessment will be made between the adult and child frequency rankings over the three ages such that frequency, as an indicator of the relationship between child phonology and ambient adult language, can also be assessed and compared with the earlier FUSE findings.

The fourth correlation assessment will observe the two languages' FUSE rankings both for the adult and child data. Close correlations between the adult FUSE rankings cross-linguistically would indicate the universal importance of certain phonemes to both

systems. Are certain phonemes ranked high for functional use in both languages? If so, then maybe these phonemes can be argued to have a universal significance and therefore might be expected to be seen in children's early speech sounds. If there is a significant correlation between the child FUSE rankings then this may be indicating a universal basis from phonological acquisition precedes towards the adult language.

In summary, for both the Finnish and English studies of spoken language there are two main areas of analysis; adult spoken language and child spoken language. This gives four distinct processing areas. The results from these four areas will then be tested for correlation as described above.

5.3 The Processing Method

Each of the four areas of processing will follow a set sequence of processing which can be divided into four stages which are detailed in the following sections;

Section 5.3.1 Stage 1 Creation of Word Files

- Corpus Selection and Data Extraction
- Pre-Processing of Word Forms
- Word Selection

Section 5.3.2 Stage 2 Data Processing

- Word Length
- Word Structure
- Minimal Pair Testing

Section 5.3.3 Stage 3 Calculation of FUSE and Ranking

Section 5.3.4 Stage 4 FUSE Correlations

The processing required for Stage 1 will vary considerably according to the corpora selected. As has been discussed in Chapter 4 available corpora for each of the four areas must be assessed for the closest fit to the requirements of this study. However,

each corpora is likely to have different transcription standards and offer a differing level of match to requirements. Processing to bring the four data sources up to a consistent standard (phonemic representation for actually spoken words) for further processing will require tailored programs for each of the areas to be written (as presented in Chapter 6).

Once the data for the four areas is in a consistent format, the remaining stages of processing can then follow a standard approach. In order to ensure consistent processing of the four areas, programs will be developed that can be applied to all four data areas. The data will be processed in a UNIX environment using simple text manipulation tools and the programming language AWK (Aho, Weinberger and Kernigham 1988). A sample of some of the programs written for this study are provided in Appendix 5.1.

The remainder of this chapter provides an overview of these common processing stages (as will be applied to all of the four areas) and then presents some of the general issues concerned with completing the processing.

The detailed processing which will be applied to each of the four different areas of processing is given in Chapter 6. Each of the four areas will produce its own set of results and FUSE rankings. Sections 7.1 to 7.4 details the results of this analysis divided into the four areas of processing of adult English, adult Finnish, group child English and group child Finnish. Chapter 9 presents the individual child results. Each of the four sections in Chapter 7 is divided into two main areas; a general findings section detailing frequency information and a section containing the homophony and FUSE totals for each of the phonemes and FUSE ranking information. The results of the correlation assessments are provided in Chapter 8.

5.3.1 Stage 1 – Creation of Word Files

- **Corpus Selection and Data Extraction**

As has already been discussed, the most important factor for the adult language

FUSE assessment is a representation of the phonemic content of frequently spoken words. A balance between utilising large adult word dictionaries (containing many word types but word types that are possibly never used in spoken language) or using smaller spoken language corpora with less word types but being more representative of actual spoken usage will be aimed at, so that the two adult data files are treated consistently. For the child word files the aim will be to fully represent a selection of children's phonological and lexical development over the time period from age 2 to 5 years. All the children's words at the three ages will therefore be processed.

The criteria of various corpora for counting a word form as a word had to be taken into account so that a corpora that provides access to the realised form of words is selected. Some data sources, particularly many of the dictionary based sources, count only a root form of a head word (or lemma) for inflected cases such as plural nouns and inflected verbs (see Chapter 4). For the purposes of this research the words as actually observed in ambient language (i.e. in their fully inflected form) is preferred, so that the actual sequences of phonemes present in ambient language can be assessed.

- **Pre-Processing of Word Forms**

Once the word forms have been extracted from the selected English and Finnish adult corpora they must be converted into a standardised form. The word forms must recognise the phonemic system and be representative of the phonemic inventories of the two languages identified in Chapter 2. Phoneme transcription codes must be amended to represent the phonemic units identified and the data must be cleaned to delete any extraneous material.

The various corpora will have been created by different sources and are likely to utilise different transcription standards. For example, the MRC database uses the Alvey transcription standard (see Appendix 5.2). Each of the corpora may also possibly represent the phonological system of the language in question in a different way depending upon the treatment of phonemic units. A certain amount of pre-processing

will therefore be necessary in order to represent the words of English and Finnish in a standard recognisable format, with the matching intra-language transcription standards outlined in Chapter 2 and reflecting the phonemic inventories described therein.
A certain amount of pre-processing will also be necessary to recognise the phonemic units. The phonemic representations of word forms will have to be scanned for these phonemic units and then a standard way of representing these units will need to be developed and consistently utilised. In order to test the impact of the various ways of assessing these 'units', several different approaches have been tested on the English adult data (Chapter 6). Phonemes represented with digraphs and trigraphs (e.g. /tʃ/, /tt/, /eɪə/) and phonemic units such as word initial consonant clusters, long vowels and geminates must be converted to a single transcription code.

As well as standardising the transcription notation and ensuring that both the adult and child data for Finnish and English fully and systematically represent the phonemic systems of the languages some data cleaning may be required to delete unwanted fields. Some transcripts contain extra fields, such as prosodic, phonetic or syllable information, that can be deleted as it is not required for this research.

- **Word Selection**

A selection of the word forms presented in the four corpora had to be made. With the adult corpora this selection was based upon frequency. In order to produce word files which represent the most characteristic words of ambient language, it was necessary to extract those words that appear most frequently in the corpora.

Whilst utilising hapax legomena words would provide a much larger word type base from which to test for FUSE (see Zipfian distribution in Chapter 4), it will present many uncommon forms of words. The size of the data files selected for processing must, as a minimum, be large enough to represent the phonemic systems of the two languages (i.e. have the majority of phonemes represented). A comparable number of phonemically transcribed word types will be sought from Finnish and English. However, both the selection of corpora that are available in the necessary format and the criteria of maintaining the representation of only most frequent word forms will

mean that the number of word form types will be reduced from the maximum and could mean a difference between the number of word types processed for the two adult languages. With the child corpora all word forms uttered by the child will be accepted as representing the child's language. The criteria for the child processing is not to find the most representative of words used by the child but to recognise all the words that the child is using in their developing system.

5.3.2 Stage 2 – Data Processing

By this point the data will have been pre-processed and cleaned, the word forms would have been selected and transcribed into the phonemic format of phonemes and phonemic units representing the phonemic systems of the two languages.

Common programs will then be written to apply the processing described below in a systematic way to all four areas under assessment. A sample of the programs themselves are provided in Appendix 5.1.

- **Word Length**

As minimal pairs are identical in every respect except one phoneme or phonemic unit, the point of contrast, it is true to say that homophony can only result from the processing of word forms of the same number of phonemes/phonemic units. Word forms will be first grouped into same length word forms. Thus all word forms of 2 phonemes/phonemic units (e.g. /æt/ 'at', /ɪt/ 'it', /hɪə/ 'here') will be grouped together, all word forms of 3 phonemes (e.g. /kæt/ 'cat', /tʃæt/ 'chat', /træp/ 'trap') will be grouped together etc. Sequences of phonemes that have been accepted as phonemic units will have already been converted to one symbol by this stage and will therefore be correctly grouped only with words that they may form minimal pairs with (e.g. /tʃɪp/ with /ʃɪp/ as both are word forms of three 'phonemes' in length).

This activity will enable both the range of word form lengths to be observed for each of the four areas of investigation and also the particular frequencies of word forms containing a certain number of phonemes for each of the four areas to be ascertained. Whilst not of direct relevance to this study, these do present new data and the results are provided in the general results of Chapter 7.

- **Word Structure**

The initial processing plan was to systematically reduce contrasts observed in words, one by one, and compare the amount of resultant homophony. In order to calculate which 'contrasts' operated the most on the words seen in the various corpora, the initial plan was to replace each phoneme with the other permissible phonemes for that word position (i.e. those that could, theoretically, be acting as a contrast to produce a minimal pair).

It was necessary with this initial approach to recognise the word in terms of its syllable structure and word structure (so that the phonotactic rules of the language were adhered to). The word forms were initially analysed for word structure (i.e. the number and type of phonemes within syllables and the number and types of syllables within words). The words were then grouped into similar structure types for further processing.

This processing was partially undertaken for all the data files and statistical information on the relative frequencies of particular word and syllable structures was produced. The processing was fully completed for the Finnish adult file producing phonemic usage by syllable and word structure. As most of these findings on statistical frequencies of words, phonemes, syllable structures etc. have not been previously available for the Finnish language they have been presented in this research (Chapter 7, General Results). Although a by-product of the FUSE processing, which is the main aim of this research, they provide a valuable body of work and give insight into the Finnish phonological system. A similar, but less detailed analysis of the English data, has been completed and provides information on the structure of words. This enables a comparison with the Finnish data.

- **Minimal Pair Testing**

The aim of this phase was to process the data files such that words could be grouped into minimal pair groups and the contrastive phonemes could be identified.

Initially it was felt that in order to assess FUSE, it would be necessary to systematically replace every phoneme in every word with all possible alternative phonemes for that word position from the phonemic system of the language. Processing was planned that not only took into account the different phonological systems of the languages and the phonotactic rules but also the rules for such elements as phoneme position within syllable, syllable position within word such that only acceptable replacements would be made. All the theoretically possible types of word and syllable structures and component phonemes of these structures for both languages would have had to have been calculated regardless or not of whether they actually occurred in the data. Complicated and lengthy processing to allow for every possible permutation would have had to have been developed. This was attempted for English and the processing required for English word initial syllables based on this theoretical model of how the words of the English language might present themselves was developed. The initial plan was for the resultant alternative word forms to be then looked up against the word lists to see if they existed amongst the most frequent words and, if so, homophony would result.

This initial approach proved extremely difficult and problematic. Firstly, it was very difficult to draw up the processing charts in a way that reflected both the phonotactic rules of the language and the actual sequences observed in the data. Attempts to define formal methods for automatic acquisition of phonotactics for speech processing systems based upon syllable structures have previously noted such difficulties (e.g. Goldsmith 1990). Word initial and word final consonant clusters demonstrate language specific co-occurrence constraints and mono-syllabic words are often highly idiosyncratic. Thus both 'within word' positioning of syllables and length of words affects phonological prediction. Even if such a complete definition were possible, previous analysis of word and syllable structures in spoken language have recognised the particular difficulties posed by spoken

language where consonant clusters that would not normally be acceptable in the system are observed in foreign words and where the processes of connected speech mean that unusual sequences are found.

As one of the main aims of this research is to provide new information about the phonological structure of spoken language ignoring the spoken language features actually demonstrated in the spoken language corpora was not an option; if they are seen in the data representing spoken language then they must exist as sequences and be acceptable to speakers of the language. The decision then had to be made whether to add new rules based upon possibly only one example existing in the data and assumptions would have to be made on which word positions to then apply these new permutations to.

Secondly, the number of processing rules was extremely great; each word position (up to position 21 for Finnish) had to have a set of permissible phonemes which would be based upon word position, preceding and following phoneme, syllable structure, syllable position within word, word structure and type of phoneme (vowel or consonant). The longest word forms would thus require in excess of 500 passes (21 x 30), a pass for each of the 30 plus possible phonemes that could contrast and for each of the 21 positions in the word. With word lists of more than ten thousand word types to process the processing overhead would have been large. A large proportion of the resultant word forms, whilst acceptable words, would not have resulted in homophony as the data file with which homophony was being compared was only of the most frequently observed spoken words.

This approach equally treated the rare and the common phenomenon of the two languages and meant that a processing was planned for some phonemes and phonemic units which, although permissible, did not actually exist in the data. For example, in Chapter 2 it was noted that Finnish recognises 17 or 18 diphthongs (Karlsson 1983 and Sulkala & Karjalainen 1992) and yet only 9 were found in the data in word initial position. Häkkinen (1983) lists 69 two consonant sequences and 25 three consonant sequences as being permissible in Finnish words and yet only 11 of these were found in the data at word initial position.

Following the initial approach therefore posed three main problems:

- Firstly, once the data begun to be analysed at the syllable and word structure level for phonemic content it was found that within the data of actually spoken language there were exceptions to the rules of Finnish phonology outlined in Chapter 2. Words that are generally accepted by a native speaker did not fit the theoretical frameworks outlined in Chapter 2 and sequences of phonemes that were previously thought impossible were observed. Usually for the Finnish data this was because of loan words or foreign names that had introduced a new 'sequence' into the language. Whilst these loan phonemes had been identified in some previous frequency assessments (see Chapter 2) they were not included in the phonotactic rules assessments by other researchers who may have simply accepted that they would be assimilated to an already acceptable phoneme (e.g. /t/ for /d/ etc.). Six, previously unaccounted for, word initial consonant cluster sequences were found in the data of spoken Finnish words these being /dr/, /br/, /fl/, /fr/, /gl/, /gr/ and two unexpected triphthongs, /yyæ/ and /ɑio/. With the basic premise that any element, be it a single phoneme or phonemic unit, might be functioning to contrast between words of the language every time a new sequence was observed this would have to be built into the processing as another possible substitute for the sequence being reduced.

- Secondly, working out exactly where a syllable boundary was proved difficult particularly for Finnish where examples such as sequences of vowels could be seen as sequences of vowels within one syllable or alternatively sequences of one vowel syllables.

- Thirdly, there were some sequences which have already been identified as being extremely rare although possible (e.g. English /gwen/). This initial approach to the processing meant that rare forms that may not exist in the actual data had to be allowed for in the same way as the more frequent

forms and it meant that processing had to be developed whether it would actually be required or not.

Having the word forms grouped by word length and word structure type enabled one to see very clearly where the similarities between word forms existed and thus where homophony was most likely to be produced. Rather than try to pre-empt any possible sequences that might or might not be observed in the data it was decided instead to use the data itself to provide the minimal pair word type groups and a second approach, as detailed below, a 'data-driven' rather than theoretical approach, was developed.

With the second approach it was recognised that already existing data of a sufficient volume would better represent phonemic sequences, syllable and word structures, not as purely theoretical possibilities but as the actual sequences directly observed in frequently spoken words. Utilising the words within the four corpora to find the 'minimal pairs', words that differed by only one phoneme (or phonemic unit), proved simpler in terms of the processing approach. It also meant that processing that directly represented the frequencies and features of the actual language sample was developed. By searching for word forms that were identical in every respect apart from one phoneme, groups of minimal pairs could be found and placed together for further processing.

In order to do this, the words were grouped into words of the same phonemic length such that possible minimal pair groupings could be found firstly based upon word length. Next the word forms were sorted by each position in turn to bring actual minimal pairs together, the position of the sort determining which position in the word was being tested for a contrast and with the number of sort orders being a factor of the word length being tested. This was done in order to bring groups of minimal pairs directly next to each other in the word lists. This meant multiple files were created for each word length, one for each sort position in the word. Thus the file of words which were 10 phonemes long would have 10 sort orders, running from sorted by first phoneme to the tenth. The Finnish data was found to contain words of up to 21 phonemes in length therefore 21 files, one for each sort position within the word where a contrast might potentially exist, was created.

The last position to be sorted by, the least significant part of the sort, in each instance, had whatever 'phoneme' appeared in this position converted to a common character 'X'. With 'X' replacing 'phonemes' that had previously operated as contrasts exact homophony was produced for the words that were previously minimal pairs. For example the English word types /trend, frend, spend, lend, send, tend and blend/ can all be seen to be minimal pairs differing only in their word initial phonemic unit. The sort would place these words together into a minimal pair group and replace the contrasting word initial phonemes with an 'x' thus giving all words grouped together as 'xend'. Where homophony resulted from this process the original phoneme could be demonstrated to have directly contributed to that homophony. A count of the minimal pair strings, word groups that form minimal pairs with each other, and also the phonemes involved as the contrasts in these minimal pair groupings could then be made. By repeating this processing for all words lengths and word positions each phonemes' contribution to the homophony could be totalled giving an overall phoneme total for each word position. With the files sorted by the second phoneme it was possible to assess which words created the minimal pairs (identical from position 2 onwards) and which phonemes were the word initial contrasts.

This enabled all minimal pair words (those that were identical in every respect apart from one phoneme) to be grouped together and counted such that the most frequently occurring identical phoneme strings seen in words could be observed. As well as enabling one to clearly see the loss of which particular phonemes provided the contrast in which word position this approach also enabled the word lengths most involved in minimal pair groups and the types of phonemes (whether vowel or consonant) to be explored. An initial attempt at grouping words into minimal pair groups was completed for all word positions and for both languages however only the word initial phoneme usage is reported here in Chapter 7.

The first phonemes of words can be easily identified as word initial. Consonant phonemes can always be described as WISI (word initial syllable initial) and vowels as V1's there was no need therefore with this study to separate the word length files by structure to obtain the overall word initial minimal pair groupings. If the remainder of

the word other than the first position was identical then the structure too would be identical.

In order to complete a fuller assessment of the phonemes that act as contrasts at other than word initial positions it would be necessary to consider syllable boundaries. Each phoneme would need to be assessed for its position in the syllable and the syllable would need to be assessed for position in the word before conversion of sequences of phonemes to phonemic units. For example words of 6 phonemes length might show many different word structures (e.g. CVCCVC, CVCVCV, VCVCVC, VCCVVC etc.). The structure CVCCVC can be clearly seen to contain 2 syllables and would typically be described as containing a WISI phoneme for the syllable onset, V1 and V2 for the two vowel phonemes, WISF(word initial syllable final) for the coda, WFSI for the fourth phoneme and WFSF(word final syllable final) for the final phoneme. The recognition of the syllable boundary between the two consonants would be necessary to ensure that the sequence of mid word consonants (e.g. 'cyg.net', 'dig.nity') is not taken as a consonant cluster attached to one syllable.

For both languages a list of the resultant word initial 'minimal pair groupings' are provided in Chapter 7 together with a list of the phonemes and phonemic units that provided the contrasts in word initial position. The minimal pair groupings can be observed by word length and the largest groupings (i.e. those with the highest number of contrasts) for each word length are also shown .

5.3.3 Stage 3 - Calculation of Functional Use (FUSE) and Ranking

As defined above, the FUSE method of assessment aims to assess the relative importance of phonemes within the phonemic system. This will be achieved by applying the functional load method to actual words representing the four areas of analysis and counting up the number of times that a particular phoneme provides the contrast between phonemic strings at a particular position within the word.

Two methods for calculating FUSE will be attempted. Firstly, the FUSE assessment for a particular phoneme will be calculated by simply counting up the number of actual

words that it provides the contrast in. For example, the English minimal pair 'mat' and 'sat' would give a count of 1 for the phoneme /s/ in word initial position and 1 for the phoneme /m/ in word initial position providing both of these words existed in the spoken language corpora. Each word type would be eligible for one minimal pair grouping and each phonemic unit would equally be eligible for a maximum count of one. Thus, the two factors of phonemic system (as represented by phonemic units) and word usage (as represented by most frequent spoken words) would be equally weighted in the calculation of FUSE.

The second method for calculating FUSE will use the count of contrastive phonemes in a particular minimal pair string to give an indication of the work that a particular phoneme is needing to do to provide the contrast. For example, the word initial minimal pair group of English words might contain the three phoneme minimal pair string /hæt/, /tʃæt/, /plæt/, /mæt/, /fæt/ and the two phoneme minimal pair string /græb/, /tæb/. The work that the phonemes are having to do to be contrastive in the first minimal pair string, where there are five words, could be argued to be more than the work being done by the phonemes /gr/ and /t/ with the second minimal pair string. With this approach, each of the phonemes would have its FUSE count incremented by the total number of contrastive phonemes in the minimal pair string (e.g. /h, tʃ, pl, m, f/ by 5 in the above example and /gr, t/ by 2). This second method for calculating FUSE recognises the fact that a phoneme may actually be carrying more weight in terms of its importance within the phonemic system based upon the number of words within minimal pair groups.

The results of this stage of the processing will be for each phoneme and phonemic unit to have its occurrence as the contrast in minimal pairs totalled to give an overall FUSE count by phoneme and word position. All word positions will be processed however only the word initial results are to be taken further in this study.

The phonemes and phonemic units will then be ranked according to this total. The rank being an indicator of the phonemes that occurred as the contrast in the most minimal pairs at word initial position.

5.3.4 Stage 4 - Fuse Correlations

In order to better assess the relationships observed, between the adult FUSE rankings and the children's frequency rankings, between the adult FUSE rankings and the child FUSE rankings and between the FUSE and frequency rankings cross-linguistically, the Spearman rank co-efficient will be used as a correlation method.

Comparisons will be made between the child rankings at each age in turn with the adult language rankings and this will be completed for both languages and for the children both as a group at each age and individually. An overall assessment of child frequency rankings will be also be produced by summing the child word initial phoneme frequency totals for each age. The general frequency correlations can then be compared with the adult FUSE and frequency correlation findings.

5.4 Processing Considerations & Limitations

There are a number of ways of viewing what may constitute a 'word form'. There are also different ways of counting the frequency of English word forms, roughly speaking divided between counting the orthographic or phonemic form both of which will have a different representation. This study will use readily available data which may restrict to a certain extent the options available for the processing. The limitations imposed by various data sources must be weighed up in terms of their usefulness and appropriateness to this study.

To a certain extent providing a consistent approach is adopted for the definition of a 'word form' and the frequency factor is handled systematically across the four areas of analysis the limitations that may be imposed by only have certain types of data available will be minimised. For example, the greater the range of words available the more likely minimal pair groupings will be found and the more phoneme contrasts will be counted. The only large British English adult data source that is readily available and provides phonemic transcriptions of words (the MRC Psycholinguistic database) provides spoken frequency information based upon orthographic representation. The phonemic transcription availability together with the variations in pronunciation that

this data source offers, far outweighs the limitation of each pronunciation variation not having its own frequency count.

These issues are now discussed and the approach to be adopted for the remainder of this study will be finalised.

5.4.1 Word Form vs. Lemma

The selection of 'word forms' for FUSE processing is dependant upon having a standard way of representing words across the four corpora.

Words may be represented in the form in which they actually appear in spoken language (i.e. with full inflection) or they may alternatively be simply represented as the base uninflected lemma form. For this study it will be important that variation in the realisation of a word based upon inflection (e.g. /kæt/ and /kæts/) which causes a change in the phonemic string used to represent the word form should be recorded as a separate word form. For example, the words 'run', 'ran' and 'running' all belong to the verb lemma 'run'. Corpora that recognise the various forms of the word as separate word forms will count three word form types. Dictionary files that have already processed the word forms may possibly only recognise one word form type. Likewise for nouns which will be recognised in both their singular and plural forms (e.g. 'cat' and 'cats' as two word form types) or as one type ('cat').

The aim of this research is to have a range of word forms which represent ambient language. The word lists should represent spoken Finnish and English and the 'word forms' should therefore be represented as they are used (i.e. by the actual strings that demonstrate the utilisation of the phonemic systems of the languages). The main criteria will be that whatever approach is adopted for the adult language assessment, must also be applied to the child language assessment so that an even representation of possible minimal pair groupings is maintained. The utilisation of word forms as opposed to the 'lemma' form gives a truer assessment of phonemic usage. However, it might also be simply reflecting, to a certain extent, the affixation rules of the language (see below).

Where previously compiled frequency counts are to be used, the basis on which the frequency has been calculated needs to be clear. Frequency counts that measure only phonemically identical word form occurrences, rather than counting occurrence of the lemma form ('run'), are required for minimal pair groupings in this study. All the corpora that have been selected for use in this study provide fully inflected forms of words.

5.4.2 Type/Token Considerations and Assessment of Frequency

The FUSE processing relies upon having a representative range of word form types (represented as different phonemic strings). Many different word form types will be sought so that the minimal pair groupings are as large as possible and more fully represent the contrastive usage of phonemes. This will mean searching out for alternative pronunciations of words where the alternative pronunciation would provide a different phonemic string.

The processing to convert the Finnish orthographic forms of words to phonemic strings means that there is a straight one for one 'orthographic' to 'phonemic representation'. Variations in pronunciation of a word (which would result in a different phonemic string) will be represented differently orthographically and the types of word form can be directly found from the input data of running text.

The English adult data source provides pronunciation variants in phonemic form (e.g. /mamut/ and /mmɪt/ for 'minute'). The English child data can utilise the English adult data orthographic word form to find all the possible variations to phonemic string representing alternative pronunciations of a word. The only problem with this approach will be that the English adult data source, whilst offering the advantages of phonemic transcription (which will be required to convert the English child data to phonemic strings) and pronunciation variants, does not provide frequency information for the pronunciation variations. One orthographic form of a words e.g. 'minute' would have one frequency recorded which must be applied to both pronunciation possibilities.

The actual frequencies of the word forms (i.e. the number of tokens) as they appear in the original data sources will be used for the initial selection of the word forms such that only the most frequent word forms are selected for further processing. Lists of word forms will be based upon word types and the measure of frequency for specific word types will not thereafter play a role in the FUSE calculations. This approach enables all phonemes to have an initially equal representation within the system and it is only their appearance or non-appearance as the contrasting element within the word types that determines their FUSE ranking (rather than their occurrence in the most frequent words). Lower frequency words that form minimal pairs demonstrate the contrastive ability of particular phonemes and usage of the phonemic system equally as well as higher frequency words. It is this approach that will be adopted in this study.

As has been discussed above, an alternative approach to this would be to use the token count as a direct input to the FUSE calculation so that when a contrast is measured, say the phoneme /s/ in word initial position, it has the overall word frequencies involved in the minimal pairs included. This approach is not possible for adult English where no pronunciation variation frequency figures exist and whilst it recognises 'usage' as actual word frequency, it would not provide a level comparison for measurement between the four types of data being analysed. A phoneme's FUSE count, for example, would not be indicating the phonemes overall importance in terms of its ability to provide contrasts between most frequently spoken words but instead it would be directly reflecting size of the data source (i.e. the larger the data source, the higher the frequencies of words). One way around this, should adult pronunciation variant frequency become available, would be to rank the words by frequency (i.e. assign a frequency rank rather than using the actual frequency of occurrence figure).

The child data is very likely have comparably lower frequencies for word usage and less word types and the adult data higher frequencies for more types. Whilst it could be argued that the individual FUSE rankings would even out any discrepancies between size of findings this approach is not followed here.

5.4.3 Treatment of English Homographs

As has already been discussed there is not a one to one representation in English between the orthographic form of words and their pronunciation and therefore many examples of homographs exist. Corpora that provide frequency information may count word forms based upon the orthographic form of a word or the phonemic form. With the first approach homographs would be served by one frequency count despite their being two pronunciations of the word. Also where the orthographic forms of words are the only evidence of pronunciation (as with the English child data), it will be necessary to ensure that all possible pronunciation variants are provided in the word lists.

The London-Lund database provides words that are orthographically the same under one entry with one frequency count as there is in fact only one orthographic word. The MRC Psycholinguistic database utilises these frequency counts (based upon the orthographic word) but goes on to provide the various phonemic representations. For instance, 'present' which can be pronounced either as /prɛzent/ or /prɪzent/ has one entry with one frequency but with the two pronunciation variants. The problem with this approach is that there is no way of establishing just what percentage of the frequency count applied to which pronunciation, instead it assumes the total frequency for both phonemic representations.

For the purposes of this research it is important to establish all the possible different pronunciations of the word if these may have fed into the frequency counts as with the London Lund database. Thus all the variant phonemic representations will be utilised. Any orthographic form of English words that carry a frequency of greater than one will have all its related phonemic sequences counted despite the possibility of some of them being hapax legomena or not even being present. The same problem does not arise for the Finnish data as different pronunciations will have different orthographic representations. Two identically pronounced words will carry the same orthographic sequence despite the word possibly signalling two or more different meanings.

5.4.4 The Treatment of English Homophones

The English homophones, words pronounced the same but with different spellings and meanings, present the opposite problem to that posed with homographs. With these identical pronunciations care must be taken that they do not feed into the FUSE calculations in a way that might possibly signal more of a contrastive role for a phoneme than there really is. Leaving naturally occurring homophony in the word lists will result in two identical phonemes being treated as contrasts (which they clearly are not) and they must therefore be identified prior to analysis of homophony produced by the processing. Identical phonemic strings to represent the two different orthographic forms will be reduced to one word form entry. The words involved in naturally occurring homophony will be listed so that analysis of the word initial phonemes in this naturally occurring phenomena can be seen.

If frequency of orthographic form has been used as the basis for frequency counts (rather than the creation of frequency counts based upon phonemically transcribed running speech), then the possibility for naturally occurring homophony must be recognised and the frequencies for identical phonemic occurrences must be summed together to give an overall frequency for that phonemic string. For Finnish, words which are pronounced the same are also represented the same orthographically and so identical orthographic occurrences will automatically indicate homophony.

5.4.5 The Treatment of Homonyms

Words with identical pronunciation and spelling orthographic representation but with two meanings are known as homonyms (Jackson (1988)). Both Finnish and English demonstrate naturally occurring homonyms in the words of their language. For example in English the word 'bank', pronounced as /bæŋk/, signals two meanings; bank of a river or a financial institution. Finnish homonyms with identical orthographic and phonemic representation but carrying two or more meanings include *säkeistä* meaning 'from the verses' and 'out of the sacs'. Often the Finnish homonyms have resulted from inflection e.g. *kirjaansa* 'his book' partitive and 'into his book' illative, *takkini* 'my coat' singular and 'my coats' plural.

The prime focus of this research is to establish the phonemes and phonemic units utilised in English and Finnish words. Variation in meaning and grammatical usage are not a focus of this study. If running text is used to form the word lists, as with the Finnish data and the child English data then homonyms, with identical pronunciation and orthographic representation will appear as extra occurrences of the same word despite the possibility of several word meanings being indicated. As sequences of phonemes are of prime concern here, and not word meanings being signalled by the sequences, this poses no problem. The frequency counts represent the phonemic word forms.

5.4.6 Processes of Connected Speech

Spoken language has the advantage of offering an insight into the phonemes and phoneme sequences actually used by speakers of a language. There are however several processes of connected speech such, as assimilation and elision, which must be taken into account when compiling word lists for the two languages.

In spontaneous connected speech sounds belonging to one word can cause changes in sounds belonging to neighbouring words. Assimilation, movement of one sound towards another (Roach 1991), affects phonemes, particularly consonant phonemes, either regressively (the one that precedes is affected by the one that comes after it) or progressively (where the one that follows is affected by what preceded it). In English, we see assimilation of place, manner and voicing in consonants. For example an alveolar consonant followed by a non-alveolar consonant, /t/ followed by /p/, usually results in the loss of the alveolar consonant, as in /ðæt persən/'that person' which is pronounced //ðæpersən/, /laɪt blu/ 'light blue' which is pronounced /laɪblu/ and /mit paɪ/ 'meat pie' which is pronounced /mipaɪ/. An alveolar consonant before a velar consonant (/t/ to /k/) is often omitted e.g. /ðæt keɪs/ 'that case' to /ðækeɪs/, /braɪt kʌlə/ 'bright colour' to /braɪkʌlə/. There are many such examples of how running speech can effect the pronunciation of words.

In Finnish the sentence *'oletko sinä ollut*? 'have you been?' made up of 5 morphemes (*ol* 'to be' + *e* present tense + *t* second person + *ko* question suffix + *sinä* 'you') could be shortened in two ways; *ots ollu* or *oks ollu* (Iivonen 1998). The first possibility *ots ollu* has shorted the first two words containing 10 phonemes to just three phonemes, each phoneme representing only one original morpheme – *o* for ol, *t* for second person, *s* for you. The second possibility *oks ollu* maintains three phonemes - *o* for ol, *k* for question suffix and *s* for you. The transcription of the word form may show the words in their original unshortened form, relying on a reader to perform the shortening, or alternatively it may already show the shortened word forms.

Spontaneous naturally occurring adult speech will also contain casual or slang forms of words which may not be found in a dictionary. Child speech may contain diminutive forms of words not found in the adult corpora. They, again, must be taken into account. Provided that the same approach to both the adult and child data forms of the two languages is adopted then, these processes are not a problem as the ratios of usage being measured within the languages will not alter.

This study utilises corpora of spontaneous connected speech for its word sources wherever possible. It is therefore expected that these will include words that demonstrate these processes of connected speech and exhibit a different phonemic string to that when the word is spoken in isolation. For the English data the RP accent used for word transcription will mean that much of these processes are in fact lost and a standard form of the word will be presented. For Finnish, where the orthographic forms of words provide the transcript, examples of words joined together and shortened (as above) will be included in the word lists. The volumes of data being investigated should mean that where such a word is presented there is also an example of its more standard form. The least frequent word forms will not be selected anyway for the processing and so the least representative words should not be included due to their frequency of occurrence anyway.

5.4.7 The Role of Syntax

No consideration for the grammatical role of the words is considered in this study. It is recognised, however, that the different use that different languages make of syntax to change the meaning of words will affect the phoneme frequencies observed in that language. For example, to change from a singular to a plural form the English language, in the majority of cases, adds the suffix 's' which can be pronounced /s/, /z/ or /ɪz/ dependent upon the phonemic context e.g. 'cat' /kæt/ to 'cats'/kæts/, slide /slaɪd/ to /slaɪdz/ and /rʌʃ/ to /rʌʃɪz/. It might be argued that word final /s/, /z/ and /ɪz/ could therefore be expected to be more frequently observed in spoken language simply because they have an important grammatical role to play. Similarly, the suffix 'ed', marks the past tense of many English verbs (e.g. he walked, she talked).
Whilst in one sense these rules skew the data towards more frequently observed phonemes, on the other hand, the phonotactic rules of a language actually prevent certain affixation, following a standard route (e.g. /s/ for plural), as seen in the above example.

In Finnish suffixation processes can be seen but obviously with different phonemes and for different purposes (e.g. *'ssa'* with *auto* as in *autossa* but ssä with *keittiö* 'kitchen' as in *keittiössä, ko* with *puhut* as in *puhutko* but *kö* with *kysyt* as in *kysytkö* 'do you ask?').

The English and Finnish languages make use of different affixation rules and different phonemes for common affixations (e.g. to indicate plural, past tense etc). Finnish, in particular, as an agglutinative language, has many forms of affixation. The frequency of these affixations and the phonemes that are used in the languages to represent them will affect the frequency of particular phonemes at particular word positions. The implications of these affixes on the word forms will be different for English and Finnish. As it is proposed that fully inflected word forms, including these affixes, are looked at for this research the affixes that are present in the words of the two languages will affect the FUSE rankings. This study considers only the word initial phonemes of words and where word initial affixes exist, as for the English data, they must therefore be considered as possible contributors to the FUSE rankings. Further research could

focus on how particular affixation phonemes influence phonemic frequency and may be influencing FUSE.

Not only do both languages follow different affixation rules, applying different phonemes, but they each make use of affixation for grammatical differentiation to a greater or lesser extent. Finnish relies heavily on affixation not only for plural/singular distinction and for verb tenses but also for prepositions, ownership, questions and degree. In addition to this each language uses prefixes and suffixes to a differing extent. English uses many the prefixes 'un', 'dis', 're' as seen in 'undo', 'disassociate', 'reassess', but relatively few suffixes. These syntactic requirements of the language do therefore contribute to both the individuality of a language and the phonotactic structures observed. The focus of this research is primarily on the word initial phonemes and it is accepted that the prefixes of the languages will affect the word initial phonemic frequency found.

The chapter has defined the FUSE method, detailed the common data processing required for each of the four areas of analysis and identified a range of considerations for the processing. As each corpora is likely to present data in a different format and with different transcription standards processing to bring the four data sources up to a consistent standard will be required. Chapter 6 now outlines the specific processing required for each of the areas of analysis.

Chapter 6 : Data Processing

The processing plan presented in Chapter 5 provides an overview of the sequence of processing which needs to be carried out in relation to each of the four data areas under consideration; adult English data, adult Finnish data, child English data and child Finnish data.

This chapter of the report provides details on the specific individual processing that was necessary to carry out for each of the four data sets, bearing in mind the different types of data input and the different phonemic systems of the two languages.

6.1 English Adult Language Processing

6.1.1 Corpus Selection and Data Extraction

As has been discussed in Chapter 4 there are a large number of English corpora available, each with differing degrees of suitability. At the initial stage of this research, when the selection of corpora needed to be made for this study, only one English corpora matched the requirements of being a record of British English pronunciation (as opposed to American English), containing spoken word frequencies with phonemic transcriptions for the words and with pronunciation variations given for homographs. Whilst orthographic representations of words were recorded in all of the corpora that were considered only one was readily available and contained phonemic representation of words.

The MRC Psycholinguistic Database was selected for the English adult data for this study. The main reason for its selection were that it provides a British English phonemic transcription for more than 38,000 word types and the spoken frequency of 14,529 words as directly recorded from a spoken language corpus (The London Lund Corpus). The phonemic transcription codes are presented in the U.K. Alvey standard for machine readable phonemic transcription (Wells, 1986). Appendix 5.2 contains these codes for ease of reference.

The entries with both a frequency count and phonemic transcription were extracted for further processing. This gave a large pool of phonemic transcriptions with a wide range of frequencies ranging from 6833 for the most frequent word 'the' down to 0.

In order to follow the principle of utilising only the most frequently spoken words initially only the non-hapax-legomena words (i.e. the word forms that had a frequency of greater than one) were extracted. This gave a sample of approximately 5000 orthographic word types from which to work. In order to observe whether significant differences were found with a larger sample of English word types (i.e. a greater range of words) the hapax legomena words were also extracted and added to the non-hapax legomena words in a separate file. With hapax legomena words included a data file of approximately 9000 orthographic forms of words was obtained.

The processing of the data was tackled in two phases. The first phase utilised the smaller file of non-hapax-legomena orthographic entries taken to be representing the most frequent word form. This was the main emphasis of this study and the FUSE results presented in detail in Chapter 7 relate to this size of file. A second phase of processing was completed for comparison purposes and additionally utilised the hapax-legomena words.

One useful feature of the MRC database is that where a word has several accepted pronunciations, all are provided, thus increasing the actual number of different phonemic strings for processing. However, one potential problem might be that one frequency count is given to relate to all variant pronunciations. As has been previously stated, the FUSE findings rely upon having a representative range of word form types (i.e. different phonemic strings). As many different word form types (i.e. strings of phonemes) will be sought so that the widest ranges of minimal pair groupings and thus use of contrasts is identified (see Chapter 5 above). This approach requires that alternative pronunciations of words are utilised even if they do not offer frequency information. Homographs, those words with two or more pronunciations, have therefore been included, as with other words, if they are non-hapax-legomena. This added an extra 69 phoneme strings as word types to the word list. Appendix 6.1 provides a full list of the homographs identified in this way; one frequency count,

relates to one word but several pronunciations. The most frequently observed were; 'a' pronounced /eɪ/ and /ə/, 'to' pronounced /tu/ and /tə/ and 'do' pronounced /du/ and /dəʊ/. The highest number of pronunciation variants were seen with the words 'contrast' (/kɔntrɑst/, /kɔntræst/, /kəntrɑst/, /kəntræst/) and 'graduate' (/grædjʊɪt/, /grædʒʊət/, /grædjʊeɪt/ and /grædʒʊeɪt/).

Following this expansion of the word lists to include homographs, the phonemic representations of word forms were then sorted by phonemic transcription. This resulted in an amount of naturally occurring homophony becoming visible. Words that had been spelt differently orthographically now had the same phonemic string to represent them. The FUSE assessment relies upon identifying contrastive phonemes in minimal pair groups. These naturally occurring homophones have one and the same phonemic string (i.e. there is no contrastive phoneme) and it was important, therefore, that only one entry for each identical string of phonemes was processed. Without the extraction of these naturally occurring homophones the later processing would have counted the phonemes involved in these naturally occurring homophones alongside those that signal a phonemic string difference.

Sixty seven identical phonemic strings were found on the data base representing 142 orthographic representations of words. Some of these were naturally occurring homophonous word forms with identical pronunciations but different orthographic spellings others were examples of homonyms (same pronunciation and spelling but different meaning). Appendix 6.2 provides the details of the homophonous forms showing the number of different orthographic representations relating to the one pronunciation, the pronunciation as the phonemic string representing the words (in MRC transcription format) followed by the orthographic representation of the words.

As can be seen from Appendix 6.2 the largest homophonous groups were found for; /ɑ/ ('are', 'R' and 'ah'), /aɪ/ ('eye', 'I', 'ay'), /bi/ ('be', 'bee', 'B'), /baɪ/ ('buy', 'by' and 'bye'), /si/ ('see', 'sea', 'C'), /fɔ/ ('for', 'four', 'fore'), /tʊ/ ('to', 'two', 'too'), /ðeə/ ('their', 'they're', 'there').

As has already been discussed, identical phoneme sequence constitute same word type for the current study. The adult English word form list was, therefore, reduced to provide only the one representation of each identical phonemic string for FUSE assessment. Each of the orthographic words has its own frequency count. The frequency counts for these word types would therefore be the combined orthographic frequency totals.

Homonyms are represented in the MRC database as two phonemic transcription entries but both carry the same frequency count. For example, the word 'frank', despite having only one pronunciation /frænk/, and one orthographic form, could be used as a verb 'to frank' or an adjective 'a frank man'. The word 'frank' appears twice in the MRC database both times with a frequency count of 4. Other examples include 'may/May' and 'arm(V)/arm(N)'. Homonyms representing these identical phonemic strings had also to be reduced to one word form on the word lists so that they would not be included in the later FUSE processing as minimal pairs.

With all of this actioned 4,713 different phonemic transcripts of words remained for further processing.

6.1.2 Pre-Processing of Word Forms

Once the phonemic transcriptions for words had been extracted a number of processes to clean the data and bring the word forms into a standardised format needed to be made.

The most significant reduction to the phonemic representations already extracted was to delete the syllable boundary and prosodic markers from the phonemic strings. As has been discussed in Chapter 5 the approach that has been adopted for the first application of the assessment measure whereby only word initial contrasts are to be measured does not rely on having the syllable markers for contrast identification. Stress is not regarded as a feature of contrast within the English phonemic system and so deleting the markers of stress did not therefore produce new word forms, for the purposes of this study, as the sequence of phones remained the same. It did, however, produce some

additional homophones. Where previously the only difference between two phoneme stings had been a stress marker, with the stress marker deleted, the phoneme strings became the same, thus representing one word form. As with other naturally occurring homophony these identical pronunciations were put under one entry in the word list for further processing. Whilst it is accepted that the pronunciation of these words would sound different, this difference is at a supra-segmental level and not one which provides a segmental contrast. This pre-processing further reduced the number of phonemic transcripts to 4670.

Most important to the pre-processing stage of the English adult data processing will be the amendment of the transcription codes so that the phonemic units of the English inventory, identified in Chapter 2 as potentially providing contrasts, are represented with one code. As discussed in Chapter 5, it will be necessary for the identification of minimal pair groupings that digraphs, such as the affricates and diphthongs, trigraphs of triphthongs and the word initial consonant clusters be converted to one code so that they are contrasted with phonemes.

For this first area to be processed two different pre-processing levels were applied to the adult English data and a pilot FUSE calculation was completed to test how the variations in phonemic treatment would affect the overall ranking structure of the phonemes. The two levels of phonemic unit recognition are now presented.

6.1.2.1 Processing Level 1 – Partial Recognition of Word Initial Phonemic Units

For the initial level of processing the two affricates /dʒ/ and /tʃ/ heard in English speech were both reduced to be represented by a single symbol; /dʒ/ was reduced to 'G' and /tʃ/ to 'C'. This enabled chaps /tʃæps/ converted to /Cæps/ to be matched as a near homophonous form with 'caps' /kæps/ and 'maps' /mæps/ despite it a phoneme count of one greater than them. Similarly 'jolly' /dʒɔlɪ/ was matched with 'folly' /fɔlɪ/ and 'golly' /gɔlɪ/.

The diphthongs, each originally represented with two elements (e.g. /ɪə/), and the triphthongs by three elements (e.g. /eɪə/), were also reduced to be represented by a single code (e.g. '1' and 'F'). Treating the diphthongs in this way meant that, for example, the words it /ɪt/, at /æt/, art /ɑt/, eat /et/, eat /it/ and ought would be grouped as minimal pairs with 'ate' /aɪt/.

Appendix 6.3 provides the processing notation for the vowel phonemes and phonemic units found in the data together with the MRC Notation (Alvey Standard – see Appendix 5.2) and the IPA representation originally presented in Section 2. These codes are used throughout the remaining processing. Appendix 6.3 also provides the frequency of the diphthongs and triphthongs as found in the adult English data. Without completing the processing in this way would have resulted in the vowel phonemes /o/ and /a/ normally only seen in English in the diphthongs /ɔɪ/, /aɪ/ and /aʊ/ and the triphthongs /ɔɪə/ and /aɪə/, /aʊə/ appearing as candidates for FUSE ranking despite them never appearing singly in English speech. An approach that would therefor be counter to the phonemic system of the language.

This first level of the processing made no alteration in the phonemic transcription to represent consonant clusters as phonemic units of one code. Two consonant clusters were treated as two elements and three consonant clusters were treated as three elements etc. Thus /hæt/ would not group into a minimal pair with /flæt/, /slæt/, /bræt/ or /spræt/; sprat with its CCCVC word structure would not be grouped with CCVC, or CVC structured words.

This initial approach was completed so that the frequencies of individual phonemes involved in word initial consonant clusters could be clearly observed and compared with previously reported frequency findings which used this method (see Chapter 2).

6.1.2.2 Processing Level 2 - Reduction of All Word Initial Phonemic Units

This level of FUSE assessment acknowledged the need to treat word initial consonant clusters as phonemic units (identified in Chapter 2). All the word initial consonant clusters that existed in the data were therefore reduced to be represented as one code for further processing. For example, the two consonant cluster /pl/ was changed to 'I', the two consonant cluster /sl/ to '{' and the three consonant cluster /spl/ to /P/.

Appendix 6.4 provides the processing notation for the consonant phonemes used throughout the remaining processing together with the MRC Notation and the IPA representation originally presented in Section 2. Appendix 6.4 also provides the frequency of these word initial consonant clusters as seen in the English adult data.

Many consonant clusters may have missed being matched to their minimal pair groupings prior to this level of processing. For example, a three consonant cluster such as /skr/ would have only been matched with other three consonant clusters in minimal pair groupings. Only one element of the three consonant cluster would have been identified as the contrast (two phonemes would have had to be the same) with the earlier level of processing.

6.2 Finnish Adult Language Processing

6.2.1 Corpus Selection and Data Extraction

The ambient language that young children experience is of a spoken nature. The primary focus of this research should therefore be on the words and phonemes presented in adult spoken language with the assumption that a child acquiring Finnish will be exposed to those words presented most frequently in spoken Finnish rather than those appearing most frequently in written Finnish.

A database that records spoken language is therefore required. As previously mentioned, there is only at present one corpus that exists for adult Finnish speech, the

Helsinki Spoken Language Corpus (Helsingin puhekielen aineisto). Following an initial review of the corpus it was decided that this corpus is sufficient for the needs of this research, particularly as the alternative would be to utilise a written language corpus (e.g. PAROLE).

Some consideration of the features of the corpus had to be taken into account. It is easy to observe that colloquial forms of Finnish spoken language show characteristics particular to dialectical area (Paunonen1994 &1995). For example, in the Helsinki dialect it is acceptable to shorten many words by the omission of phonemes e.g. *minä* to *mä, sinä* to *sä* and *nyt* to *ny*. In connected speech it is also acceptable to join together into one word form words that would be written separately e.g. *se oli* 'you are' to *soli*. In addition a range of assimilation processes account for the systematic change of spoken sounds in connected speech. Word final /n/ may change to /m/ before a word initial /p/ or /m/ as in *menem pihalle* 'I'm going into the garden' and word final /n/ may change to /ŋ/ before /k/ as in *menen kotiin* 'I'm going into the house'. For the purposes of this study the words as presented in the Helsinki Spoken Language Corpus were used. It was felt that the large range of words offered on the corpus would most likely present a good range of standard word forms as well as a range of connected speech word forms.

6.2.2 Pre-Processing of Word Forms

Once the phonemic transcriptions for words had been extracted (i.e. excluding comments contained within '/ /') a number of processes to clean the data and bring the word forms into a standardised format needed to be made.

An initial scan of the Helsinki Spoken Language Corpus reveals a total of approximately 600,000 running words i.e. tokens. The transcription standards used during data entry used markers for phonetic features and other symbols for comments, proper names, unfinished words etc. Word types that are in fact phonemically identical may therefore look different initially due to these extra markers. A certain amount of data cleaning to remove these markers and find phonemically identical word forms was therefore necessary.

The Helsinki Spoken Language Corpus contains a number of special transcription standards to represent descriptive qualities in addition to phonemic representation. For example, words are marked with '>' to indicate aspiration, with '+' for particularly short consonants, with ' " ' to indicate the looseness of diphthongs and with the numbers 1 to 9 for a whole range of other phonetic characteristics. Thus the word form 'yksi' might in fact be recorded 'yk>si', 'yksi8' or a range of other alternatives. Appendix 6.5 contains a fuller description of the corpora and the codes utilised. These markers are not required by this study and so were deleted from the phonemic strings representing word forms.

Non-standard ASCII characters which were used in the transcription as alternatives to ASCII characters (e.g. '{' for 'ä', '|' for 'ö', '[' for 'Ä' and '\', 'for 'Ö') were initially converted into standard Finnish orthography for ease of reading.

At the same time in order to give long vowels phonemic status as opposed to simply treating them as two occurrences of their short vowel equivalents and to make the identification of them easier in future processing, long vowels were converted to be represented as follows;

'aa'	changed	to	'A'
'oo'	"	"	'O'
'ää'	"	"	'B'
'öö'	"	"	'C'
'ii'	"	"	'I'
'uu'	"	"	'U'
'ee'	"	"	'E'
'yy'	"	"	'Y'

Geminates were not found in word initial position and so did not need to be converted to phonemic units. Their appearance within words that formed minimal pairs would be identical whether left as 2 codes (as is done here) or converted to one. Minimal pairs grouped by word initial position contrast, where a word is identical in every respect apart from the first phoneme, will only occur if every other part of the word is identical.

Whether the word includes a geminate or not will not affect this assessment and no phonemic status needs to be given to within word geminates and hence providing them with a conversion code is not necessary.

The transcriptions were made using orthographic transcriptions for the phoneme representation. As there is a one to one relationship between all Finnish graphemes and phonemes (apart from the velar nasals), this enabled the orthographic transcriptions to provide the phonemic values. Orthographic sequences of the velar nasals 'ng', which can be represented phonologically as /ŋ/, were therefore converted to the transcript code 'N' and sequences of 'nk', which can be represented phonologically as /ŋk/, were converted to 'Nk' in order to distinguish them from the alveolar nasal /n/.

According to Finnish research (see Section 2 above) a total of 18 different diphthongs were expected to be seen. The original plan therefore was to convert these to one code, as with the English processing. Without this conversion to a code then only words that contained the identical number of phonemes would be matched in the minimal pair groupings (i.e. a singleton vowel would not be matched with a diphthong, and only diphthongs containing one identical vowel phoneme in the same position (1st or 2nd) would be matched).

An initial scan of the data extracted from the Helsinki Spoken Language Corpus revealed a total of 41 different 2 vowel sequences. Only ten of these 'sequences' of vowels, 9 diphthongs and one VV sequence, previously explored in Section 2, however, appeared in word initial position;

/ɑi, ɑu, æi, ei, eu, oi, ou, ui, iæ and yœ/

Included in the forty one sequences of two vowels seen within words were the 18 expected diphthongs plus an additional twenty three sequences of two vowels.

Many different three vowel sequences were observed both within words and at word endings. Of the six three vowel sequences found at the start of Finnish words

(/ɑɑi, eei, eii, ɑio, oii and yyæ/) only one was taken as acceptable by a native Finnish speaker who assessed the word lists. The words 'aai', 'eii', 'oii' and 'eei' were dismissed as interjections and 'yyällän', which occurred only once in the data, was taken to be a transcription error. This left only one Finnish three vowel sequence 'aio' in the words 'aiotte' and 'aiot' to consider as a phonemic unit for reduction to a code for processing or alternatively to keep as a two syllable sequence e.g. 'ai-otte'.

As the main focus of this research is to test the word initial phoneme usage, it was decided to only substitute those diphthongs that appear in word initial position with a symbol for processing. This would still enable minimal pair word groups, where the only difference was at word initial position, to be detected whilst avoiding an assessment of where a syllable boundary might be between VVV sequences. As has already been explored, Finnish rules for determining syllable boundaries depend not only upon the presence and sequence of vowels but also on their position within word or syllable. Appendix 6.6 contain the processing codes used for the vowels, diphthongs and the one VVV sequence treated as a unit hereafter in the processing.

As with the English processing in order to be able to detect minimal pair groupings involving sequences of consonants that could be considered as units at word initial position it was necessary to convert consonant clusters found at word initial position to one code.

An initial scan of the data extracted from the Helsinki Spoken Language Corpus revealed a total of 16 different two consonant clusters (CC) and 1 three consonant cluster (CCC) at word initial position;

/br, dr, fl, gl, gr, kl, kr, pl, pr, sk, sl, st, sv, tr, ts, fr and skr/

Appendix 6.7 gives the original IPA codes together with the orthographic representations and the codes that were used in the processing to represent the consonants and consonant clusters. A sample of the words with initial consonant

clusters is also provided together with frequency information for the number of words containing these clusters.

The many other sequences of consonant phonemes which were observed within word and at word final position have not been analysed further for this research. The issues of whether to treat geminates in a similar manner to consonant clusters and how to assess syllable boundaries so that sequences of consonants, which are not clusters, are not treated as such would need to be further addressed for analysis of FUSE ranking that concentrated at positions other than word initial position.

Once all of this pre-processing had been completed the compound words, marked with '_' between the words, could have the '_' removed to give the compound word as one word form. Completing this process before the conversion of the long vowels, diphthongs/triphthongs or consonant clusters would have resulted in a sequence of two short vowels seen at the juncture of two words being converted to a long vowel or diphthong and sequences of consonants (as opposed to clusters) being treated as clusters which would have been incorrect. As with the English stress markers the '_' does not represent a phoneme or phonemic unit and so could be deleted without change to the phonemic string representing the word form.

Once the data had been pre-processed and cleaned of extraneous fields the words themselves could then be extracted and their frequencies counted up such that for each word type a frequency was obtained.

6.2.3 Data Selection

Following this initial pre-processing and cleaning of the data a total of approximately 34,250 word forms, i.e. types, were found to exist in the corpus. Hapax legomena words were removed from further processing. Eight foreign words, words of non-Finnish orthography, containing the letters 'x', 'z' or 'c', were also removed (as their pronunciation would have been estimated rather than precise – see Chapter 2). With foreign and hapax legomena word forms removed and with identical strings of phonemes grouped together, the total number of different word forms was 10,306. This

figure includes 475 proper names which were considered as a string of phonemes as with any other word form. These proper names had a code to mark their status as such, which needed to be removed, so that they had the phonemic transcription codes used in this study.

Written Finnish was used for the transcription process and as this represents all phonemes directly orthographically (apart from the velar nasals which had already been converted to their phonemic representation), it was possible to utilise the orthographic transcription to find the phonemic representations. Naturally occurring Finnish homophonous forms, two word forms that are pronounced the same but carry two meanings (e.g. *säkeistä* 'from the verses' and 'out of the sacs'), would also be spelt the same in the corpus and would therefore be counted together as one word form which means that naturally occurring homophonous forms cannot easily be extracted from the total 'phoneme string' count . Neither can a word with 2 meanings be easily seen without a semantic interpretation. The focus of this research is to consider sequences of phonemes and not the semantic range associated with them and so neither of these pose a problem.

6.2.4 Initial Approach

In order to assess which phonemes were contrasts in minimal pair groupings and be able to describe them in terms of their phonotactic position (e.g. WISI, WISF etc.) in addition to their word position (e.g. 1, 2 etc. as above) it would be necessary to consider each phoneme within its position in the word together with the overall word structure. For example words of 6 phonemes length showed 104 different word structures (see Chapter 7). The most frequently occurring structure CVCCVC can be clearly seen to contain 2 syllables and the following could be used to identify each phoneme in position e.g. WISI for the first phoneme, S1 vowel for the second phoneme, WISF for the third phoneme, WFSI for the fourth phoneme, S2 vowel for the fifth phoneme and WFSF for the final phoneme. A word of format CVCVCX, whilst still containing 6 phonemes, can be seen to contain 3 syllables and therefore require different description categories e.g. WISI for the first phoneme, S1 vowel for the second phoneme, WWSI for the third phoneme, S2 vowel for the fourth phoneme,

WWSF for the fifth phoneme and S3 for the final phoneme. Thus each word form would have to be analysed in terms of structure before the homophony findings for all similar types e.g. WISF, WFSF etc. could be added together.

To demonstrate this point, for word forms of 6 phonemes the consonant phoneme /t/ was seen to occur at WISI position in 30 homophonous forms, phoneme /s/ in 30 forms, /m/ in 27 etc. These consonant phonemes represent syllable onsets. However the vowel phonemes /e/ and /a/ which are also seen in the list must represent initial syllables without onsets. The vowel phoneme /e/ was seen to occur in 4 homophonous forms, in V1 position and the vowel phoneme /a/ was seen to occur in 3 homophonous forms in V1 position. When looking at the same 6 phoneme words in another position e.g. position 4 the word structure would have to be considered in order to work out phonotactic descriptions i.e. V1 or WISI cannot be assumed.

As discussed above, for the initial application of FUSE in this study only the FUSE results for the phonemes in word onset position (WISI or V1) have been reported. As word initial consonant phonemes can always be described as WISI and all word initial vowels as V1's there was no need to separate the word length files by structure to obtain the overall word initial FUSE rankings. Word onset FUSE rankings relate to a count of the amount of homophony presented by the reduction of first phonemes in minimal pairs. Some information has been gathered for phonemes in word final position which would assist in future work to look at contrastive phonemes by word positions. The information now exists to complete the calculations for all phonotactic description types by word length and overall if required in future work.

6.3 English Child Language Processing

One objective of this study is to assess the words that children typically utilise such that the phonemic system can be analysed in terms of frequency and compared with the adult frequency findings. In addition, an assessment of the children's own FUSE rankings at several stages of development will provide a new way of observing the changes that occur during acquisition.

6.3.1 Corpus Selection and Data Extraction

In order to assess development it was important to identify samples of English child language which enabled the progress of particular children to be followed over a period of several years. Rather than a snapshot of the sort of words used by many children at a particular age, of which there are many samples on the CHILDES database, data for children over several ages were sought. As was discussed in Chapter 4 the selection the corpus from CHILDES was based around naturally occurring, British English, spoken data. However, none of the suitable corpora actually provided phonemic transcripts of the children's realisations of the word forms.

By age 2 the children will probably have acquired many of the phonemes, although not all of the phonemic units such as consonant clusters, of the language they are acquiring (see Section 3). Also, they may not have acquired the phonemes in all word positions. This will be a good point to start the assessment and compare the word type phoneme frequencies observed in adult language with those in child language. English children will have the words that they use analysed for phonemic features at three stages of development such that their relative closeness to their adult language (in terms of word type word initial phoneme frequency) and the new measure of functional use can be assessed.

The aim was thus to find examples of the same children's words at roughly aged 2, 3 and 5 years. Whilst Childes contains many files of child language no British English phonemically transcribed child data of naturally occurring spontaneous speech currently exist. Those that are phonemically transcribed either represent American English or else are not spontaneously produced (i.e. word elicitation games etc).

As Chapter 3 has discussed children's earliest attempts at words are often variable in pronunciation and may not be a good indication of the word forms being attempted. It was felt therefore that utilising the target adult forms of words would provide a better way of recognising the child's lexical system. In order to recognise the target adult forms of words care-givers or native speakers' interpretation of the word forms will be necessary.

As the children's words would in the main be expected to exist already in the adult English word lists it was decided therefore to create the child phonemic representations of words based upon the English adult word files. This way naturally occurring homographs would be included as with the adult data (i.e. several pronunciations would be included). The frequency counts for these homographs in the child data thus relate to the orthographic forms of words and not the actual pronunciation of words. Naturally occurring homophones would also be easily detectable with this approach. The frequency of occurrence for each orthographic form will be summed to provide an overall frequency by word type (phonemic string).

Having the orthographic representation of words also meant that problems to do with accent (Wells contains data samples collected in Bristol) were avoided as the standard RP form of the word, as represented in the adult word lists and used for the adult processing, would be used. The Wells data sample on CHILDES was selected as it contains a relatively large pool of data relating to many different children. It records British English naturally occurring spontaneous speech, and records the same children over a time span of some 3 years. The data collection project, headed by Gordon Wells, was collected during the 1970's for a project entitled "The Bristol language development study: language development in pre-school children". It contains 299 files from 32 British children (16 girls and 16 boys) aged 1;6 to 5;0 recorded in a naturalistic setting. Each child was observed a total of ten times at three monthly intervals. The samples were recorded by tape recorders that turned on for 90 second intervals at a time. The main part of each recording was the spontaneously occurring conversation of the child. Recordings were transcribed orthographically.

A sub-set of 5 children, Benjamin, Betty, Elspeth, ,Geoffrey and Jason were selected. Three samples existed for each of these children at three roughly consistent ages of 2 years, 3 years and 5 years and they gave samples of language use by both boys and girls. The children's words produced at each of three ages, approximately aged 2 years, aged 3 years and aged 5 years, were extracted from the data sample.

Table 6.1 provides information on the ages that the various samples were taken, and the total duration of the time under examination.

Table 6.1 – English Child Data Samples

Name of Child	Age	CHILDES Sample id.	Ref. No	Time between Recordings	Total Time Span
Benjamin	1;11	Wells	benjam04		
Benjamin	2;11	Wells	benjam08	12 months	
Benjamin	5;0	Wells	benjam21	25 months	3 years 1 month
Betty	2;0	Wells	betty04		
Betty	3;0	Wells	betty08	12 months	
Betty	4;11	Wells	betty21	23 months	2 years 11 months
Elspeth	2;0	Wells	elspet04		
Elspeth	3;0	Wells	elspet08	12 months	
Elspeth	5;0	Wells	elspet21	24 months	3 years
		Wells			
Geoffrey	2;2	Wells	geofre04		
Geoffrey	3;0	Wells	geofre08	10 months	
Geoffrey	4;11	Wells	geofre21	23 months	2 years 9 months
Jason	2;0	Wells	jason04		
Jason	3;0	Wells	jason08	1 year	
Jason	5;0	Wells	jason21	24 months	3 years

6.3.2 Pre-Processing of Word Forms

The orthographic representations of words extracted from CHILDES were converted into phonemic transcriptions for the words utilising the adult English word lists as a look up table. Where a word existed in the child word lists but didn't exist in the adult word lists this was marked for future reference and the larger MRC database was used instead.

The word lists were expanded to include alternative pronunciation options e.g. 'the' pronounced both as /ði/ and /ðə/ (represented 'D@' and 'Di' for the processing) and 'a' pronounced /ə/ and /eɪ/ (represented '@' and '5' for the processing) . Interjections such as 'mm' and 'eh' were deleted as were unrecognisable word forms such as 'qqqq'. Naturally occurring homophony that occurred within a particular data sample for one

child e.g. /tu/ for both 'two' and 'to' and /ðeə/ for 'there' and 'they're' were grouped together as one phonemic string and one word type as with the adult language processing.

Searching for minimal pair groupings based upon the inflected forms of verbs and the plural forms of nouns would mean that only those minimal pairs carrying exactly the same inflection or plurality would be identified and that this would prove less useful than utilising the root forms of verbs and the singular forms of nouns. Where a plural form of a noun had been used by a child it was assumed that the child also had the singular form within their lexicon and so this was added to the word list (e.g.'cakes' assumes 'cake') in order to increase the word types recorded for each child. The root form of regular verbs were similarly assumed from an inflected form e.g. 'jumping' assumes 'jump'. The inflected and plural forms were also left in the word lists so that should they also occur in minimal pairs then this would be identified.

As with the adult English data a number of processes to bring the word forms into a standard representation recognising the phonemic inventory of English and cleaning of the data to remove unwanted fields needed to be completed. The representation of the phonemic system had to be altered such that the components of the phonemic system were recognised in the same way as for the adult English data. The phonemic units of word initial consonant clusters and the phonemes represented by more than one transcript element (the affricates, diphthongs and triphthongs) were converted to the codes utilised for the adult English data (Appendices 6.3 and 6.4).

The result of this processing was a set of fifteen word type data files representing the English child data with one data file of word types for each child at each age. Following a standard linguistic approach to child phonemic analysis the resultant word lists were then analysed non-parametrically for frequency and for FUSE such that the individual children's phonemic usage could be observed over the three ages. The results of these individual analyses are given in Chapter 9.

The word type files for each age were also combined to provide one word type data set for each of the three ages under assessment. These group data sets contained identical

word forms where a particular word type had been utilised by more than one child. An assessment of the common word usage (i.e. words used by more than one child) could then be made and the frequencies of particular word usage could be combined to give an overall word type frequency ranking. Once this had been completed identical word forms were deleted to provide a data file of word types for further phoneme frequency and word structure assessment. The results of the group findings are presented in Chapter 7.

6.4 Finnish Child Language Processing

6.4.1 Corpus Selection and Data Extraction

There are no publicly available corpora of spoken Finnish child data available in an existing electronic format (Section 4). In order to be able to include Finnish child data in this study, the decision was made therefore to invest the effort required to transfer a selection of files from the Oulu collection, as representative of a well established existing collection, and convert these into a format for electronic analysis and processing. A small number of files that had already been converted into text files were used alongside the majority which were paper versions of typed transcripts and needed converting into electronic text format.

The objective of this research is to assess how the children's phonemic usage changes during language acquisition. As with the English processing it was therefore essential to identify samples which would enable the progress of particular children to be followed. Rather than a snapshot of the sort of words used by 2, 3, 4 and 7 year old children (which individual word lists would give), data samples for children at each of these ages were sought. As with the English child data a sub-set of 5 children, including both boys and girls, were selected based upon their proximity of data collection timing. The typed transcripts for Harri, Riikka, Sami, Teppo and Virpi, were thoroughly analysed and transferred as word lists to electronic files.

The children's words produced at each of four ages, approximately aged 2 years, approximately aged 3 years, approximately aged 4 years and approximately aged 7

years were extracted from the data samples. Only ages 2, 3 and 4 were actually used for this study, however, in order to approximately match the ages being tested for the available English child data and still show continuity with data for same child. Future research may further utilise the older Finnish age data that is now available in electronic text format. The word lists were further checked by a native Finnish speaker to minimise the number of possible transcription or transfer errors included. Table 6.2 provides more detailed information on the ages that the various samples that were used for this study were taken at, the original source of the data files and the total duration of the time under examination.

Table 6.2 Finnish Child Data Samples

Name of Child	**Age**	**Electronic File No.**	**Ref. No**	**Time between Recordings**	**Total Time Span**
Harri 1	2.04	Transcript			
Harri 1	3.03	1011	Merk01	11months	
Harri 1	4.04	1012	Merk01	13 months	2 years
Riikka	2.05	Transcript			
Riikka	3.05	2081	Merk07	1 year	
Riikka	4.05	2082	Meri07	1 year	2 years
Sami	2.04	Transcript			
Sami	3.04	1111	Merk09	1 year	
Sami	4.04	1112	Merk09	1 year	2 years
Teppo	2.07	Transcript			
Teppo	3.07	1131	Merk11	1 year	
Teppo	4.06	1132	Merk11	11 months	1 year 11 months
Virpi	2.08	Transcript			
Virpi	3.08	2161	Merk16	1 year	
Virpi	4.10	2162	Merk16	14 months	2 years 2 months

Although ideally the child and adult corpora should have been of the same Finnish dialect such that word forms were identically presented this was not possible for this study due to the lack of corpora from which to make the selection.

6.4.2 Pre-Processing of Word Forms

As with the adult Finnish data a certain amount of pre-processing needed to be completed to ensure that the long Finnish vowels retained their phonemic identity and

that the velar nasals seen in the orthographic strings 'nk' and 'ng' were treated as such (see the arguments used for this assessment given in the adult Finnish data section above). Sequences of 'aa' were converted to 'A' to represent the long /a:/ vowel sound, 'oo' to 'O', 'ää' to 'B', öö' to 'C', 'ii' to 'I', 'uu' to 'U', 'ee' to 'E', and 'yy' to 'Y'. Sequences of 'nk' were changed to 'Nk' and 'ng' to 'N'. Consonant clusters seen at the word initial position were converted to be represented as one symbol (as were diphthongs at word initial position) in order to duplicate the adult Finnish processing (Appendices 6.6 and 6.7). Other sequences of vowels or consonants seen within words were not changed in any way.

Once the lists of words had been created for the five children at 3 ages each it was clear that there were several compound words where it could be argued the child had acquired the individual components of the words. The reasoning behind this being that a word of two components e.g. 'kaunokirjotusta', literally 'nicebookcase', did demonstrate acquisition of the two components 'kauno' and 'kirja'. These two words are more likely to be seen in minimal pairs than the full word form where a much larger phonemic sequence would have to be matched. The two shorter words would all be tested for minimal pairs. These words were left in their compound forms as well as being divided into the individual components. For example, the word *kuormaauto* 'lorry'(literally 'load car') was accepted as three words *kuorma, auto* and *kuormaauto.*

Homophonous forms that would be pronounced the same would also be spelt the same, so no two words could be spelt the same but pronounced differently as was evidenced in the English database. Frequency of word types could therefore be directly ascertained by totalling the number of times a particular phonemic string occurred.

As with the English child data the result of this processing was a set of fifteen word type data files representing the Finnish child data with one data file of word types for each child at each age. The results of the analysis of these individual child data sets are given in Chapter 9. The word type files for each age were combined as with the English child data in order to provide one word type data set for each of the three and the results of the group findings are presented in Chapter 7.

Chapter 7 : Results

This chapter explores the results of the data processing divided into the four areas of analysis;

- Section 7.1 English Adult Data
- Section 7.2 Finnish Adult Data
- Section 7.3 English Child Data
- Section 7.4 Finnish Child Data

The results found for each of these four areas is divided into two main sections;

- General Findings
  - number of word types processed
    - most frequent word types
  - frequency of phonemes by word type
    - phoneme frequency over all word positions
    - word initial phoneme frequency
  - word type lengths, structures and frequencies

- FUSE Findings
  - Minimal pair word type groups
    - number of word types in a group
    - range of minimal pair string lengths
  - FUSE assessment by phoneme
  - FUSE assessment ranked

This chapter will be reporting on the findings for the English and Finnish children with the children as a group. For the purposes of further linguistic analysis the individual child data findings are also presented in this study in Chapter 9.

The first section of the general findings will assess the ultimate size of the word lists to be processed for each of the four areas. The number of different non-hapax-legomena

word types the adult data contains and the total number of word types the child data, with the children as a group, contains will be reported. The most frequent word types will also be reported for the four areas. This will enable an overview assessment of whether the predictions that Jakobson's Universalist Theory makes for Finnish and English acquisition (Chapter 3) matches the child word type findings.

The phoneme frequencies will be assessed based upon the presentation of word types (i.e. each word type will have its component phonemes counted once for each time the phoneme appears in the word type). This approach should present some new findings as much previous phoneme frequency has reported on usage based upon tokens. The underlying premise for this approach, as discussed in Chapter 5, is that according to phonological theory a child will hear and assess a phonemes role within the system based upon its occurrence in different words rather than its specific frequency of occurrence overall.

Findings both on phoneme frequency over all (i.e. all word positions) and word initial position for both adult and child data will be reported. The phoneme frequencies for the child data will be presented for the children as a group at each age here and for the children individually in Chapter 9.

Word type structures in terms of their component vowels and consonants will be presented together with the frequency of each structure for all four data sets.

The FUSE processing will provide a number of new, previously unreported, findings to do with the overall frequency of minimal pairs for each of the four areas. Minimal pair groups, containing strings of minimal pair words that differ by word initial phoneme, will be produced for every word length. The minimal pair groups for the four areas under analysis will be assessed in terms of the size of the minimal pair group, the range of word lengths involved in the minimal pair group and the range of minimal pair string sizes (i.e. the number of contrast phonemes). It will be interesting to observe whether the two languages make the same sort of use of phonemes as reflected in minimal pairs.

The FUSE counts for each phoneme will then be totalled for each of the four areas and the phonemes will be ranked according to their FUSE counts. Each of the four areas will thus present a FUSE ranking which can be compared with both frequency rankings (adult to child) as well as the other FUSE rankings (see Chapter 8). The child data FUSE rankings are presented individually in Chapter 9 such that patterns can be more closely examined.

7.1 English Adult Results

7.1.1 General Findings

From approximately 5000 non hapax-legomena orthographic forms of word types selected for processing a total of approximately 4700 different word types remained after homophonous word forms (both naturally occurring and stress variations) were deleted and variations in pronunciation were added (see Chapter 6).

7.1.1.1 Word Types and Frequency

The English word frequency figures for this research come directly from the MRC database (see Chapter 4). As stated in Chapter 4 the spoken frequency count is taken directly from a sub-set of 200,000 words of the spoken language recorded in the London Lund Corpus of English Conversation as produced by Brown (1984). This records frequencies for approximately 11000 different word types.

The most frequently produced word types are shown in table 7.1.

Table 7.1 Most Frequent English Adult Word Types

Frequency	*Word*	*Frequency*	*Word*	*Frequency*	*Word*
6833	THE	3653	IT	1652	BUT
6797	I	3169	THAT	1621	KNOW
5453	AND	3123	IN	1582	THIS
5006	A	2675	YES	1464	HE
4817	YOU	2100	IS	1371	IT'S
4434	TO	2079	WAS	1365	THEY
4253	OF	1753	WELL	1365	HAVE

7.1.1.2 Phoneme Frequencies

Once the word file of different word types had been cleansed of extraneous fields and pre-processed to represent the English phonemic inventory as defined in Chapter 2, the frequency of phonemes could be counted.

The phoneme frequencies have been counted both by position within word and overall (including all word positions). These counts include the allowance for diphthongs/triphthongs, affricates and word initial consonant clusters to be represented as 'units', as already outlined in Chapter 6. Within word sequences of consonants, some of which may be regarded as consonant clusters, have been treated as strings of phonemes for the purposes of this count.

As the emphasis of this study is on the phonemic system and the relative importance of phonemes within that system (rather than word frequency) each word type was presented once for the count rather than the count being made of a running stream of phonemes. That is to say that the individual word type frequencies have not been included in assessing the phoneme frequencies.

The results of this assessment are shown in Appendix 7.1 which gives the frequencies of particular phonemes for each word position. Appendix 7.1 shows the number of word types of each phonemic length found in the English adult data and also enables both the range of word lengths and the range of phonemes by each word position to be observed. As can be seen words ranging from one to fifteen phonemes in length were found.

In summary, the ten most frequent phonemes over all word positions were;

/ɪ, t, n, ə, s, l, k, d, r, z/.

This matches very closely with previous phoneme frequency findings (see Chapter 2).

Knowles (1987) also included the phoneme /ð/ in his ranking of the top ten most frequent phonemes replacing the phoneme /k/ which he ranked 11th. The high frequency of /ð/, ranked 9th in Knowles' study but ranked 38th in this study can probably be accounted for by the fact that /ð/ occurs in many of the most frequent words, namely 'the', 'this', 'that' and 'they'. With this study these word types would be counted once whereas Knowles would have counted the phonemes in each word occurrence. Interestingly despite the non-recognition of consonant clusters and triphthongs in Knowles' study the ten of the eleven most frequent phonemes are the same in the two studies.

Mines, Hanson and Shoup (1978) also included the vowel phonemes /i/ and /e/ within the top ten most frequent phonemes in their study of American English phonemic usage but they excluded the phonemes /z/ and /k/. Wang & Crawford's (1960) assessment of consonant phoneme frequency included the phonemes /ð/, /m/ and /w/ but ranked the phoneme /z/ 12th.

Appendix 7.1 shows that out of the total of 97 possible phonemes and phonemic units (the 24 consonant phonemes, 39 word initial two consonant clusters, 9 word initial three consonant clusters and 25 vowel and vowel sequences provided in Chapter 2) only 87 were observed being utilised in the most frequent word types of the spoken data corpora used for this study. Six word initial two consonant cluster sequences /ʃr, gw, θw, dw, sj, sf/, three word initial three consonant cluster sequences /spj, skl, skj/ and the vowel triphthong /eɪə/ did not occur at all in the word forms selected for processing. As the objective of this study is to select the most frequent word types as a representation of ambient language, rather than a range of word types in order to demonstrate all of the phonotactic possibilities of the language, then this lack of representation of all the possible phonemes in all possible word positions is acceptable.

Appendix 7.1 also enables the range of usage of each of the phonemes by position within word type to be observed. As can be seen the phonemes /z/and /ɪ/ are used the most extensively across more word positions than any other phonemes. They are the only phonemes to occur in within word positions ranging from 1 to 14 and therefore display the greatest range of occurrence. The phonemes /ə, t, l, ŋ, r/ all occur in thirteen word positions.

7.1.1.3 Word Initial Phoneme Frequencies

The frequency listing provided in Appendix 7.2 relates specifically to the number of different word types that had a particular phoneme at WI position.

As Appendix 7.2 shows the ten most frequently observed WI phonemes were found to be;

/k, s, ɪ, r, d, f, m, p, ə, b/

As has been shown above there were 87 different phonemes observed in the data. All except the consonant phoneme /ŋ/ are possible in word initial position and the consonant phoneme /ʒ/ is extremely rare therefore a total of 85 different word initial phonemes might have been expected in word initial position. Not all the possible phonemes and phonemic units actually occurred in word initial position however. The vowel phonemes /u, ʊ, ʊə, əʊə, ɔɪə, aɪə/ did not occur at word initial position thereby reducing the word initial phoneme inventory by a further 6 to 79. Each of these 79 phonemes and units would be possible candidates to act as contrasts in minimal pairs and could therefore potentially carry a FUSE ranking.

7.1.1.4 Word Structure

Word structures ranging from one to fifteen phonemes in length were found.

Appendix 7.1 shows the number of word types of each phonemic length found in the English adult data. As can be seen the most frequent length of word type was four phonemes, followed by words of five, three and then six phonemes.

Eleven word types of one phoneme in length, in each case a vowel phoneme as would be expected, were found. These are /aɪ, eɪ, eə, ɪə, aʊə, əʊ, i, ə, ɔ, ɑ, 3/ (I/eye, a/A, air/heir, ear, our/hour, oh/owe/O, E, a, or/ore, are/R and err'). One word of a length of 15 phonemes was identified this being /ɪnkɔmprɪhensæbl/ 'incomprehensible'.

From the total of approximately 4700 different word forms a total of 313 different word structures were identified. With vowel phonemes represented as V and consonant phonemes and word initial consonant clusters as C table 7.2 shows the 20 most frequent word structures observed.

Table 7.2 – English Adult Word Structures

Frequency	Structure	Frequency	Structure
498	CVC	118	VCC
378	CVCC	112	CV
364	CVCVC	102	CVCCC
216	VC	99	VCCVC
173	CVCV	96	CVCVCVC
151	CVCCVC	75	VCVCC
147	VCVC	74	VCVCC
132	CVCVCC	74	CVCCV

Chapter 2 outlined how the phonotactic rules of English would create the possibilities for these vowel and consonant sequences to be seen in English words. It is interesting to observe the use that the language has made of the different possibilities. For example, twenty five vowel phonemes could potentially have been found in nucleus only syllables (V) and yet only 11 different one vowel phoneme words were identified (see above).

The rule of maximal communicative intent with minimal effort (see Chapter 2) predicts that words of one syllable should be seen more frequently than words of more than one syllable. As has previously been shown above the 20 most frequent word tokens are indeed all of one syllable in length and table 7.2 shows that the one syllable word

structures CVC and CVCC are in fact more frequent than any other word structures. However, when word type structures (as opposed to word tokens) are assessed the third most frequent type of word structure is in fact comprised of two syllables (CVCVC). When assessing the range of word structures that an English child is exposed to it must therefore be remembered that the English child hears structures of more than one syllable the third most frequently in different words.

7.1.2 English Adult FUSE Ranking

7.1.2.1 Minimal Pair Findings

Word forms were first collected into minimal pair groups (based upon number of phonemes in the word form) and then sorted to bring minimal pairs together into minimal pair strings (see Chapter 5). As can be seen from 7.1.1. above the English data includes word forms of up to 15 phonemes in length. Files of the minimal pair groupings were created for every word position (i.e. 15 contrast files, one for each position within the word where a contrast might potentially exist). Minimal pair groupings have been created for every word length and for every position within word although it is only the word initial findings that this study will be reporting on.

A total of 1610 word types were involved in groups of minimal pairs where the contrastive phoneme occurred at word initial position i.e. words that differed from one or more other words only in the word initial phoneme. Appendix 7.3 provides the minimal pair groups by word length. The minimal pair strings for each word length can be clearly seen as can the contrastive phonemes. Appendix 7.3 also shows the number of minimal pairs involved in a minimal pair string, the contrastive phonemic units and the word types themselves (the 'x's mark the position of varying phoneme as described in Chapter 5). For example, under the heading 'word length of 4 phonemes' there are a number of minimal pair strings relating to minimal pairs. The sequence 'x6nd' relates to the phonemic string /xamd/ and shows that there are 8 minimal pairs within the minimal pair string of word types ending /amd/. These 8 word forms have the phonemes /gr, f, k, l, m, s, w, bl/ in word initial position forming the words

/gramd/, /famd/, /kamd/, /lamd/, /maɪnd/, /samd/, /wamd/, /blamd/, (representing 'grind', 'find', 'kind', 'lined', 'mind', 'signed', 'wind', 'blind') and so on.

As can be seen in Appendix 7.3 words of up to 8 phonemes in length were involved in minimal pair groupings. Words of three phonemes in length demonstrated the most word types in the minimal pair groups with 685 word types involved, followed by words of four phonemes in length with 472 word types involved (see appendix 7.3). Within each minimal pair group (based upon word length) there are a number of minimal pair strings representing the minimal pair words.

As would be expected, the largest groups of minimal pairs were for the shorter words although each group would obviously be restricted by the phonotactic rules of the language. English, for example, has the possibility to utilise 3,350 different two phoneme sequences (1750 onset and nucleus and 1600 nucleus and coda) but three times this amount of 2 syllable words (CVCV, VCVC, VCCV).

Whilst words of two phonemes in length demonstrated smaller minimal pair groups (188 word types), this length of word demonstrated more contrastive phonemes involved in minimal pair strings (i.e. less minimal pair strings but a greater range within a string).

As Appendix 7.3 shows, sixteen word types of the format /xɔ/ (where x represents the contrast) were found. The two phoneme minimal pair word forms found ending in the phoneme /ɔ/ are;

/flɔ, dʒɔ, stɔ, bɔ, dɔ, hɔ, lɔ, nɔ, pɔ, dr, fɔ, jɔ, kɔ, mɔ, sɔ, wɔ/ representing 'flaw/floor, jaw, store, bore/boar, door, haw/whore, law/lore, gnaw/nor, pore/pour/poor/paw, drawer/draw, for/four, your, core/caw, moor/more, saw/sore, war/wore'.

The phonemic transcriptions for the 16 words clearly demonstrate the contrast phonemes whereas the orthographic representation does not. The orthographic form of

the word forms are provided here for clarity in reading, however, as can be seen even within this small subset of words there are several naturally occurring homophones recorded.

One other large two phoneme minimal pair groups is the group of 16 word forms ending in the phoneme /u/;

/klu, flu, bru, tru, kju, dru, du, hu, lu, fu, nu, ju, vju, dʒu, tu, dju/
representing 'clue, flew, brew, true, queue, drew, do, who, loo, few, new, you view, Jew, to/two/too, due'.

7.1.2.2 FUSE Calculation

Once the words involved in the minimal pair groupings had been identified then a count of the individual phonemes involved in the contrasts could be made. There were 1610 word forms involved in minimal pairs thus 1610 contrastive phonemes. To the totals for each phoneme were added the one phoneme words. A total range of 74 different phonemes/units were observed as the contrast phoneme in minimal pairs. It is these 74 phonemes that form the basis of the FUSE analysis and ranking.

The phonemes were counted for FUSE in two ways. Firstly each of the phonemes acting as the word initial contrast (e.g. /gr, f, k, l, m, s, w, kl/) would have their individual phoneme counts increased by one. With this approach, approach A, each contrast is deemed to start with an equal weighting regardless of the number of contrastive pairs within a minimal pair string. The second approach, approach B, incremented the phoneme count by the number of total words in the minimal pair string. Thus, in the example above where sixteen words were in the minimal pair string, each of the 16 contrastive phonemes would have its count incremented by 16. With this approach the traditional functional load method of calculating a phonemes 'work load', the work that a phoneme is deemed to have to do in order to contrast one word with another, is taken into account.

Word types of one phoneme in length are automatically going to form a minimal pair group with all other one phoneme words (they differ in only one element) and contain only the one minimal pair string. The eleven one phoneme word types have therefore been added to the FUSE totals for the phonemes, with approach A with the increment of one and with approach B incremented by 11.

Appendix 7.4a provides the totals for each phoneme involved in word initial position minimal pairs. The 74 phonemes involved as the contrasting phonemes of minimal pairs are ranked by the number of times they appeared as the word initial contrasts within minimal pairs (approach A). Each phoneme's count for all word lengths is totalled to give an overall FUSE total for each individual phoneme. Appendix 7.4b provides the totals for each phoneme with the FUSE count incremented by the minimal pair string length (approach B). Appendix 7.4c provides the totals for both of these FUSE count methods together with the ranks of the phonemes with the different approaches. The difference between the rank structures with the two approaches can be observed.

In summary, the phonemes ranked the highest for FUSE (i.e. judged to be the most important phonemes for the adult English system to differentiate between words) with the first approach were found to be;

/s/, /l/, /f/, /w/, /t/, /b/, /m/, /k/, /r/, /h/, /p/ and /d/.

With the second approach, using the number of contrastive phonemes in a minimal pair string to form the count, they were found to be;

/s/, /w/, /f/, /l/, /h/, /t/, /b/, /k/, /m/, /p/, /r/ and /d/.

It is interesting to observe that all of these highest ranked FUSE phonemes are singleton consonant phonemes. It is interesting also to observe that the 37 highest ranked phonemes are, in fact, all consonants or consonant clusters with both approaches. The highest ranked two consonant cluster is /st/ at rank positions 13 and 14 and the highest ranked three consonant cluster is /str/ at rank positions 32 and 36.

Both of these consonant clusters have as one of their components the /s/ phoneme which ranked the highest for FUSE over all with both approaches. Significantly the first vowel phoneme to be ranked for FUSE, does not occur until ranking position 38 with the first approach (the vowel phoneme /æ/) and position 37 with the second approach (/e/). This could be a reflection of the higher frequency of consonant phonemes in word initial position overall (see Appendix 7.1) and the higher frequency of words that have the structure with a syllable onset for the first syllable (see Section 7.1.1.4 above).

7.1.3 Summary of English Adult Findings

The most frequent English phonemes over all word positions were found to be;

/ɪ, t, n, ə, s, l, k, d, r, z/.

The most frequently observed word initial phonemes were found to be;

/k, s, ɪ, r, d, f, m, p, ə, b/

The phonemes ranked highest for FUSE were;

/s, l, f,w, t, b, m, k, r, h/ (with approach A and each phoneme counted once)

and

/s, w, f, l, h, t, b, k, m, p/ (with approach B using minimal pair string count).

As can be seen a different phoneme rank order is produced for each of the two FUSE assessments and the two phoneme frequency counts, any of which could be correlated with the corresponding child data findings. The word initial frequency and approach A FUSE ranking will be directly compared to the child data frequency and approach A FUSE rankings in order to assess correlation in this initial application of the method. The correlation results are presented in Chapter 8.

7.1.4 Large to Small Data Sample Comparison

For a direct comparison between two different size word files, the processing was also completed for a larger English word file. This file of words extracted again from the MRC Psycholinguistic Database contained an additional 4685 phonemic transcriptions for hapax legomena word. They may not therefore be a realistic indicator of speakers' usage of the system but are included here to provide a direct comparison between using a smaller and larger data set of words.

As would be expected from the larger word basis there are more minimal pairs observed and the overall frequencies of the individual phonemes involved in minimal pair strings are therefore higher. Despite the larger range of words from which to calculate FUSE the findings are similar. The larger English adult data file including hapax legomena words contains a larger range of words and its size more closely matches the Finnish word file size. However the large file has the disadvantage of including words that are unlikely to appear frequently in spoken language.

With regard to the difference in the rankings of the phonemes of smaller to larger dataset it is interesting to observe that regardless of the size of the dataset the same phonemes consistently appear within the top ten FUSE rankings.

The same ten phonemes /s, l, f,w, t, b, m, k, r, h/ are involved in both FUSE rankings despite their actual ranks being slightly different. The phoneme /s/ remains the most frequently utilised contrast phoneme followed by /w/ for the larger corpora and /l/ for the smaller corpora. Both the large and small data sets demonstrate that /s/, /f/, /l/ and /w/ are consistently important for signalling contrasts within English. As there is such a close correlation between the ranking orders in spite of the different corpora sizes it was decided to utilise the smaller adult English data file containing only non-hapax legomena words as a better representation of ambient language from here on in the processing.

7.2 Finnish Adult Results

7.2.1 General Findings

From the running word total of approximately 600,000 word tokens and with hapax legomena words removed a total of approximately 10,300 different word types remained for further processing. Whilst this figure is much higher than the number of word forms processed for the English data it was felt that reducing the word forms further, by perhaps excluding words with a frequency of 2 as well as hapax legomena words, in order to better match the number of English adult word types would not give any advantages (see Large to Small English data set comparison above). The higher word count gives a good representation of the most frequently spoken words and at the same time provides a good range of phonemic use for analysis.

7.2.1.1 Word Types and Frequency

The 20 most frequently produced word types ranked in terms of frequency are shown in table 7.3.

Table 7.3. Most Frequent Finnish Word Types

Rank	Frequency	Finnish Word Form	English Translatior
1	7658	ja	and
2	3683	oli	was
3	3341	se	it
4	2175	ku	when
5	1997	sitte	then
6	1710	ei	no
7	1653	mä	I
8	1591	että	that
9	1497	ni	so/ok
10	1416	ne	those
11	1374	et	that
12	1270	on	is
13	1109	niin	so/yes
14	1047	joo	yeah
15	981	sitä	that
16	970	sit	then
17	941	mutta	but
18	870	kyllä	yes
19	857	siinä	there
20	831	nyt	now

A full listing of the 100 most frequent word types in the Finnish adult data set, including frequencies and a translation into English, is provided in Appendix 7.5. The word type /ə/ also occurred within the top 20 word forms occurring 1126 times. It is taken here to represent a pause or hesitation and is deleted from further processing. The word type /əə/ occurs 549 times and the word type /mm/ occurs 404 times: both are taken to represent pauses or interjections and are deleted from further processing.

The most frequent Finnish spoken word forms, when translated, provide an almost identical list to the most frequent English word forms presented in Chapter 2 and analysed above. The only words that do not exist in the Finnish list are not separate word forms anyway in Finnish (e.g. 'the' as Finnish has no determiners, 'she' as there is no gender difference for pronouns in Finnish, 'of' which is represented as an affix etc.). As with the most frequently produced English words all are closed class function words. However, unlike the English findings several of them are of more than one syllable in length e.g. *ole, sitte, että, sitä, mutta, kyllä and siinä.* This is discussed further in 7.2.1.4 below.

7.2.1.2 Phoneme Frequencies

With each word type represented once a count of the phonemes seen at each position within all of the words was made. Appendix 7.6 contains frequency information for each word position, a count of phonemes by word position (up to position 21), and demonstrates not only the frequencies of particular phonemes at each word position but also the range of phonemes observed at each position.

In summary the ten most frequently observed phonemes were;

/t, i, ɑ, s, l, e, n, k, o, m/

An inventory of 61 phonemes and phonemic units that might operate as contrasts within the Finnish phonemic system have been identified so far in Chapter 2 and Chapter 6; 33 consonant phonemes/units (16 singleton consonant phonemes, 16 word initial two consonant sequences, 1 word initial three consonant sequence) 28 vowel

phonemes/units (8 short vowels, 8 long vowels, the schwa vowel /ə/, ten WI diphthongs and one WI three vowel sequence (VVV)). As can be seen from Appendix 7.6 all these possible phonemes and phoneme units were observed being utilised in the most frequent words of the spoken data corpora.

The frequency findings of this study concur closely with the findings of Pääkkönen (1973) who ranked the phonemes /ɑ, i, t, n, e, s and l/ as the most frequently found. Vainio (1996) and Pajunen & Palomäki (1984) both ranked the phoneme /i/ as the most frequently occurring phoneme (more frequent than /ɑ/) whereas this study, using word types as opposed to tokens in running text, found the consonant phoneme /t/ more frequently than either of the vowel phonemes.

Vainio (1996) ranked an almost identical top ten phoneme list to this study, even though his figures were for running text rather than for word types. He included the phoneme /æ/ in his top ten listing, whereas this study replaced this phoneme with /l/ in the top ten. Vainio's top ten frequency list of tokens includes the phonemes /æ/ (ranked twelfth in this study) and /u/ (ranked eleventh) whereas this study replaces these with /l/ and /m/. This could be a reflection of the frequency of particular words within the running text.

Häkkinen's (1983) study also produced an almost identical top ten phonemes rank list. Whereas the phoneme /l/ was included within this top ten most frequent phoneme rank list /o/ was not, again being replaced with /æ/ (ranked twelfth in this study). The expected frequency for the ranking of diphthongs was also very close as were the long vowel rankings. Whilst the most frequent consonant clusters seen with both studies was /st/, thereafter there is some difference between the consonant cluster rankings. This is probably to do with the different treatment of geminates and within-word sequences.

It is perhaps obvious that following the representation and recognition of phonemic units adopted for this study, the diphthongs, triphthongs and consonant clusters should have lower frequencies as these sequences are only represented in word initial position.

It is interesting to observe how the short vowel phonemes tend to be more frequent overall than the long vowel phonemes and the loan phonemes /b/ and /g/ are less frequent than other singleton consonant phonemes.

With regard to phoneme usage by position within word Appendix 7.6 enables the range of usage of each of the phonemes to be seen. The phonemes /ɑ/ and /t/ are used in all 21 positions and are therefore the most extensively used phonemes across all word positions. The phonemes /l/ and /e/ occur in twenty within word positions.

These findings, when compared with the findings for English adult phoneme frequency, show a more frequent use of the vowel phonemes. The English adult data had only 2 vowels within the top ten most frequent phonemes, /ɪ/ and /ə/ at positions 1 and 4, whereas 4 of the top ten most frequently used phonemes for the Finnish adult data were vowel phonemes, /i/, /ɑ/, /e/ and /o/ at positions 2, 3, 6and 9 respectively. It must be kept in mind, however, that Finnish has a higher proportion of vowel phonemes than English (29 out of 61 for Finnish compared to 25 out of 97 for English).

7.2.1.3 Word Initial Phoneme Frequencies

As Appendix 7.6 demonstrates, certain phonemes do occur more frequently and possibly have more significance at certain positions within the words. As this research is focusing on word initial phonemes, an analysis of the phonemes seen in word initial position is now provided. The frequency listing provided in Appendix 7.7 relates specifically to the number of different word types that had a particular phoneme at WI position. As Appendix 7.7 shows the ten most frequently observed WI phonemes for Finnish were found to be;

/k, t, s, m, p, v, l, n, h, j/

An inventory of 61 phonemes has been identified so far and has been seen in use in the Finnish words selected for this study. The phoneme /ŋ/ is not permissible in word initial position and so would not be expected to appear in the list as the start of a valid word

form. It did appear in two words *'ng' 'ngyt'* both of which are taken to be either interjections or invalid word transcriptions (probably *'nyt'*). The long vowel phoneme /œ:/ (represented orthographically as 'öö') is the only phoneme that is permissible but not found at all in word initial position in the word lists used for this research.

It is interesting to note that only nine diphthongs out of the eighteen diphthongs identified in Chapter 2 were actually seen in word initial position;

/ɑi, ɑu, æi, ei, eu, oi, ou, ui and yœ/

As has already been discussed in Chapter 6 one two vowel sequence /iæ/ and one three vowel sequence /ɑio/, as in 'aiotte', were also observed in word initial position. Many other vowel sequences were observed both within words and at word endings, possibly representing several syllable nuclei rather than dipthongs or triphthongs.

The frequency ranking for word initial phonemes given by Häkkinen (1983) is /k, s, j, p, t, m, v, n, h, o/. Whilst this study found a slightly different rank order only one phoneme was different in the top ten phonemes when comparing these two studies. Whereas Häkkinen (1983) ranked the phoneme /o/ in position 10 this research placed the phoneme /j/ in that position. Considering that this study is counting phoneme frequency based upon word types and Häkkinen's study is based on tokens, it is very interesting to observe such similar findings on the most frequent word initial phonemes. Interestingly, despite the overall high ranking of vowels for frequency in the adult Finnish data none appear in the top ten most frequent word initial phonemes.

7.2.1.4 Word Structure

Words from 1 to 21 phonemes in length were found. Appendix 7.6 shows the number of word forms of each phonemic length found in the Finnish data. As can be seen, the most frequent word length was six, followed by five, seven and then eight phonemes. The interesting thing here is that all these most frequent word lengths are longer than those found in the English adult data.

Three Finnish words of one phoneme in length were identified; /ei, ui, yœ/. As would be expected each of these one phoneme length words is a vowel phoneme. It is interesting to note that all of these one phoneme words are, for Finnish, diphthongs. Each one phoneme word type forms a minimal pair directly with all others. The minimal pair group string, therefore, for words of one phoneme in length, contains 3 contrastive phonemes. Two words of 21 phonemes were identified; *diakonissalaitoksella* 'at the Diakonissalaitos hospital' and *seitsemäsluokkalaiset* 'seventh year school students'. As there is very little published information about the frequency of particular word structure types for Finnish it was decided to analyse these results in greater detail for the Finnish adult data than for the English adult data.

Using the notation of C for consonant phoneme, V for short vowel phoneme and X for long vowel phoneme a total of 1586 different word structures were found. The most frequent word structure was 'CVCCVC' with 355 occurrences. This figure increases to 493 when the 138 long vowels are also included (CVCCXC). The second most frequent word structure was CVCC with 277 occurrences, followed by CVCVC with 257 occurrences and CVCVCCV with 230 occurrences. It is notable that long vowels did not feature within the 11 most frequent types of word structures. Table 7.4 shows the most frequent word structures (consonants are represented with 'C', short vowels represented with 'V' and long vowels with X).

Table 7.4 – Most Frequent Finnish Word Structures

355 CVCCVC(+138 CVCCXC)
277 CVCCV
257 CVCVC
230 CVCVCCV
229 CVVCVC
187 CVCCVCCV
156 CVVCVCCV
155 CVVCCV
152 CVCV
149 CVCCVCVC
139 CVCCVCV
138 CVCCXC
131 CVCCX
129 CVVCV

As can be seen from the above the most common word form length contained 6 phonemes. As would be expected from statistical norms the word lengths cluster around this peak frequency. It is interesting to note that short vowels were far more frequent in word forms than long vowels.

The rule of maximal communicative intent with minimal effort (see Chapter 2) predicts that words of one syllable should be seen more frequently than words of more than one syllable. Overall, an analysis of the most frequent Finnish words contained in the adult Finnish corpora shows the use of multi-syllable words even in the those most frequently utilised. When analysing Finnish word types it is clear to see that not only do longer words appear more frequently but that their structures tend to be of more than one syllable in length.

Even for one word length there were many different word structures observed. For example the most frequently occurring word length of 6 phonemes had a total of 104 different word structures. Table 7.5 shows some of the most frequent of these.

Table 7.5 – Most Frequent Finnish Six Phoneme Word Structures

Frequency	Word Structure
355	CVCCVC
229	CVVCVC
155	CVVCCV
138	CVCCXC
105	CVCCVV
99	CVCVCV
73	VCVCCV
71	CVVCXC

Similar analysis has been completed for each word length such that future research might consider the variation of word structures by word length or syllable structures by word or syllable length.

7.2.1.5 Phoneme Frequency Within Word Format Type

In addition to completing the analysis of the most frequent phonemes by word position and the word type, the frequencies of phonemes at each position within each specific

word format were also recorded for Finnish. For example, the most frequent word format type was CVCCVC (see above) and the most frequent phonemes for this word structure were found as shown in table 7.6.

Table 7.6 – Finnish Phoneme Frequency for Six Phoneme Words

1st Position		2nd Position		3rd Position		4th Position		5th Position		6th Position	
Phon.	Freq.	Phon.	Freq.	Phon.	Freq.	Phon.	Freq.	Phon.	Freq.	Phon.	Freq.
t	67	a	78	l	95	k	60	e	91	n	90
k	62	e	78	n	65	n	59	a	72	t	55
s	56	i	74	h	41	l	58	i	44	s	53
m	40	u	47	r	40	t	35	ä	40	m	42
p	28	o	33	t	40	s	33	u	37	v	21
v	25	ä	32	s	31	m	22	o	37	N	20
j	23	y	13	N	14	r	20	y	29	l	19
l	17			k	12	d	19	ö	5	j	16
n	16			m	8	j	18			k	16
h	12			p	6	h	14			p	11
r	7			d	2	v	11			r	7
f	2			f	1	N	4			h	4
						p	1			d	1
						f	1				

From these tables it is easy to see which phonemes appeared the most frequently in specific word structures by word position.

This level of detail has been produced for all word and syllable structures and might be of use to future research. It was initially prepared as a way of better understanding the format of Finnish syllables and words in preparation for FUSE processing. However, it was no longer required when the data driven approach explained above in Chapter 5 was adopted.

7.2.2 Finnish Adult FUSE Ranking

7.2.2.1 Minimal Pair Findings

As with the English processing the Finnish words were formed into a minimal pair groupings each containing a number of minimal pair strings of varying length and containing different contrastive phonemes. Appendix 7.8 provides a full list of the

words involved in the minimal pair groupings sorted by word length and minimal pair string size.

A total of 1278 Finnish words were involved in groups of minimal pairs where the contrast phoneme occurred at word initial position. This compares with 1610 words for English. Despite the adult English word file containing less actual word types than the Finnish word file there were more minimal pairs where the initial phoneme provided the contrast found in the adult English data. This may be indicating something about the different nature of minimal pairs in English and Finnish (i.e. that English tends to have more word initial contrasts whereas perhaps Finnish utilises another word position such as word final position).

Words of up to 10 phonemes in length were involved in minimal pair groupings. Words of 5 phonemes in length demonstrated the largest minimal pair grouping with 317 words involved, followed by words of four phonemes in length with 268 words involved. Once again both of these counts of word types are lower than the English findings (i.e. less words involved in minimal pair groupings) however the words themselves tend to be longer in terms of number of phonemes than for English.

Whilst words of two phonemes demonstrated less minimal pair word types than words of 5 or 4 phonemes in length words of 2 phonemes, as with the English data, again demonstrated the largest minimal pair strings (i.e. greater range of phonemes within a string). For example, nine spoken word forms of 2 phonemes in length were found to end in the phoneme /ɑ/ (e.g. *sa, ta, va, ja, pa, ma, la, ka* and *ska).* Seven contrasts were identified for word forms of three phonemes in length (e.g. *kai, lai, mai, sai, tai, vai, tsai).* Similar lists of the word initial minimal pair groupings for every word length are presented in Appendix 7.8.

7.2.2.2 FUSE Calculation

Once words involved in minimal pair groupings had been identified then a count of the individual phonemes involved in the contrasts could be made. There were 1278 words involved in Finnish minimal pairs, thus providing contrasts involving 1278 phonemes.

The results for approach A, where FUSE has been counted once every time a particular phoneme appears in a minimal pair group, are given in Appendix 7.9a. The word length totals are shown for each phoneme and the phonemes are then ranked for FUSE. These results are presented alongside a summary of approach B (where the phoneme count has been incremented by the number of phonemes within the minimal pair string) in Appendix 7.9c.

In summary the highest ranked Finnish phonemes according to the FUSE assessment with approach A were found to be;

/s, m, t, k, n, v, j, p, h, l/.

With approach B the highest ranked phonemes were;

/s, m, t, k, n, v, j, p, h, o, e, l/.

As with the English FUSE findings all the top ten phonemes with approach A are singleton consonants.

The highest ranked vowel is /o/ at position 11 (with approach A) and 10 (with approach B) followed by /e/ at positions 13 (approach A) and 11 (approach B). Both of these vowel phonemes rank considerably higher than with English perhaps an indication of more open first syllable usage (i.e. nucleus without a preceding consonant). The highest long vowel is /o:/ (ranked positions 17 and 14) . After position 10 both vowels and consonants appear (unlike English which shows more consonant usage). It is not until rank position 17 that the first diphthong /ei/ appears and rank position 13 that the first consonant cluster /ts/ appears.

As well as the /ŋ/ phoneme which is not permissible and the /œ:/ phoneme which did not occur in word initial position there were several phonemes/units that were identified as potential word initial contrasts but did not occur in minimal pairs.

Three vowel phonemes (/œ, u, ə/), two diphthongs (/æi, eu/), seven two consonant clusters (/dr, fl, gr, kr, pl, sl, tr/) and the one three consonant cluster (/skr/) did not appear in minimal pairs and are not therefore ranked for FUSE. All of these phonemes and phonemic units occurred only infrequently in the corpora in word initial position (see Appendix 7.6).

This leaves a total of 47 phonemes which are involved as the word initial contrast in minimal pairs. It is the approach A FUSE rankings for these 47 phonemes that will be used hereafter for comparison with child frequency and FUSE findings.

7.2.3 Summary of Finnish Adult Findings

In summary, the most frequent phonemes observed in Finnish word types over all word positions were found to be;

/t, i, ɑ, s, l, e, n, k, o, m/

The most frequently observed word initial phonemes as found in the word types analysed were found to be;

/k, t, s, m, p, v, l, n, h, j/

The phonemes ranked highest for FUSE were;

/s, m, t, k, n, v, j, p, h and l/ (with approach A)

and

/s, m, t, k, n, v, j, p, h and o/ (with approach B).

As can be seen once again a different rank order is produced for each of the FUSE assessments and the phoneme frequency counts. The word initial frequency findings and the approach A FUSE rankings will be compared to the child data frequency and FUSE findings to assess correlation (Section 8).

7.3 English Child Results

The spoken words of the five English speaking children at the three ages under investigation were extracted from the CHILDES data samples.

As discussed in Chapter 6, the adult English orthographic word forms as represented in the full MRC database were used as a look up table to provide the phonemic transcriptions for the words used by each child at each age. Pronunciation variants were added, interjections were deleted and word lists were expanded to include root forms of vowels and singular nouns. This resulted in a list of phonemically represented word types for each child at each age (i.e. 15 different word lists, one for each of the five children at each of the three ages under assessment). Each child demonstrated an individual set of word types used, a different number of word types and also a different frequency of particular word types. The individual child findings are explored further in Chapter 9.

Once the individual word type lists had been ascertained they could then be grouped together by age to provide a larger word type set for the three ages under assessment. In order to follow the previously presented approach of utilising word types (rather than tokens) identical word types which occurred when more than one child had used an identical word were reduced to one representation of the word type in the group file. For example, if the five children's word totals for age 3 had been simply added together a total of 671 words would have been presented however in searching for unique word types only 411 different words were found to exist. With the words grouped in this way a total of 256 word types were processed for age 2, 411 word types for age 3 and 762 for age 5.

7.3.1 General Findings

The general findings on the English child data, with the children treated as a group, are now presented.

7.3.1.1 Word Types

Whilst each child demonstrated a preference for certain words at each stage of development (as Chapter 9 explores) and therefore has a different frequency ranking for the individual words, two factors emerge when the children's words are grouped together and then compared to the adult word frequency rankings;

- There is degree of similarity between the words used by children at each age. The individual word totals totalled for age 2 suggests 369 words however when word types are counted only 256 remain (i.e. 31% of the words are produced by more than one of the children). For age 3 the individual word total of 671 is reduced to 411 word types (39%) and for age 5 the word total of 1319 is reduced to 762 different word types (58%).

- The most frequent words for the children as a group at each age match closely to the most frequently spoken adults words. The most frequently spoken words with all the child word counts grouped together are now given (those which are frequent but only spoken by one child are marked with '*').

  At age 2 'no, mummy, a, want, I, oh, Betty*, you, mum, I'm*'

  At age 3 'I, you, that, want, in, the, it, mummy, my, can, there'

  At age 5 'I, you, it, to, the, look, that a, and, mum*.'

Another way of assessing word type usage is to assess those words that are used by all five of the children under assessment at one age. This provides a way of assessing which words are generally the most produced at each age rather than being a reflection of a smaller number of children using particular words more frequently.

At age 2 the words 'mummy', 'no', 'yes', 'look' and 'it' were used by all five of the children. At age 3, in addition to those given already for age 2, were the words 'one', 'what's', 'what', 'want', 'to', 'my, 'you', 'got', 'don't', 'up', 'is', 'in', 'that', 'a' and 'I'. By age 5 a total of approximately 60 words were being used by all the five children. These include, in addition to the words above, 'was', 'see', 'said', 'some',

'put', 'me', 'can't', 'have', 'get', 'them', 'this', 'the', 'they', 'there/their', 'I'm', 'of', 'on', 'and', 'at'.

The words predicted with Jakobson's Universalist Theory (see Chapter 3) relate to child realisations or attempts at adult forms of words whereas this study has looked specifically at the adult realisations behind the child forms. Only a very brief analysis of the results of this study compared to the theory is therefore possible and certain assumptions on the likely child pronunciation behind the words is necessary.

Interestingly, from the 20 words predicted as early words with the universalist approach the word /mɑ/ 'ma' is already found in the English children's words at age 2. Additionally the words 'mummy', 'mum' and 'mumma' which are amongst the most frequently spoken words by the children at age 2 could be said to relate to this phoneme sequence.

The words 'daddy', 'be' and 'tea' also appear in the children's words at age 2 and could be representing the predicted /dɑ/ and /dɑdɑ /, /bi/ and /bibi/ and /ti/ and /tɑti/ phoneme sequences respectively.

By age 5 the children are also using /mi/ 'me' and /bɑbɑ/ 'baa-baa' as well as the word 'baby' which could be representing this phoneme sequence.

7.3.1.2 Phoneme Frequencies

Once the word files of word types for the children grouped together at each age had been created the frequency of phonemes could be counted.

The phoneme frequencies have been counted by position within word and overall (including all word positions). As the emphasis of this study remains the phonemic system and the relative importance of phonemes within that system (rather than word frequency) each word type was presented once for the count rather than the count being

made of a running stream of phonemes. That is the individual word type frequencies have not been included in assessing the phoneme frequencies.

The results of the frequency assessments for the three ages are shown in Appendix 7.10. These give the frequencies of particular phonemes for each word position. The table enables the range of word lengths to be seen, as well as the range of phonemes by each word position.

In summary, at age 2 the ten most frequent phonemes over all word positions were;

/t, ɪ, n, s, k, d, æ, b, m, ɒ/

Interestingly, all of the consonant phonemes in this list are also included in the top ten most frequently used adult phonemes (see 7.1 above) perhaps suggesting a closeness in the frequency of phoneme usage between the children at the adult language even at this young age. As well as the vowel phoneme /ɪ/, ranked 2nd for the child data but 1st for the adult data, the children additionally utilise the vowel phonemes /æ/ and /ɒ/ both of which are open vowels.

In summary, at age 3 the ten most frequent phonemes over all word positions were;

/ɪ, t, n, d, æ, k, m, s, z, ə/

In summary, at age 5 the ten most frequent phonemes over all word positions were;

/l, n, ɪ, t, d, ə, m, k, ʒ, p/

Appendix 7.10 shows that out of the total of 97 possible phonemes and word initial phonemic units a range of 58 were being utilised by the children at age 2, a range of 66 at age 3 and 73 at age 5.

At age 2 all the singleton consonant phonemes and fifteen of the consonant clusters, including one 3 phoneme consonant cluster /spl/, were in use. Both /spr/ and /pr/ occurred frequently in the adult data but were not seen in the child data for this age. All the vowel phonemes and diphthongs, apart from /ʊə/, were in use already at this age but only one of the triphthongs /aʊə/ was in use.

By age 5 the only phonemes and phonemic units that had been utilised in the adult words but were not being utilised in any of the children's words were /θr, pj, mj, dj, kj, hj, vj, skw, sm, spr, stj/ and the triphthongs /ɔɪə/ and /əʊə/. All of these also occur only relatively infrequently in the adult data.

7.3.1.3 Word Initial Phonemes

In order to ascertain the usefulness of the earlier proposed acquisition theories and to complete the correlation of adult to child frequencies for word initial position some general statistics about the phonemes seen in word initial position for the five children under analysis were gathered. The individual child data findings are provided in Chapter 9.

With the children treated as a group and only unique word types included the word initial frequency rankings for the 3 ages can be seen in Appendix 7.10.

In summary with this approach at age 2 the most frequent word initial phonemes were;

/b, h, k, m, w, t, s, d, p, n/

At age 3 the most frequent word initial phonemes were;

/w,b, h, d, t, m, k, p, f, r/

At age 5 the most frequent word initial phonemes were;

/b, h, m, w, s, p, d, r, t, k/

At age 2 the group results show a vary different top ten ranking to the adult word initial top ten phoneme frequency findings. The phonemes /h/, /w/, /t/ and /n/ are ranked 2nd, 5th, 6th and 10th for the child data but do not appear in the adult data. By age 5 the children's top ten word initial phoneme frequency rankings contain only 2 phonemes that are not in the adult data. Whereas the top ten of the overall frequency findings for age 2 children were already very close to the adult findings these word position specific findings seem to more clearly show a movement. They may be indicating a gradual move towards the adult frequency findings over time which the correlation results in Chapter 8 will set out to explore.

From the list of 79 different phonemes observed in the adult data at word initial position, a range of only 45 different phonemes were observed in the English child data at age 2, 54 different phonemes were observed at age 3 and 64 at age 5. Not all the possible phonemes and sequences acceptable as units therefore occurred in word initial position in the child data.

As shown above the number of different phonemes used by the children as a group increases from 45 at age 2 to 64 at age 5. As would be expected, the number of phonemes not utilised at all therefore decreases as the children get older.

Another way of looking at this data is to assess which phonemes are used by all 5 of the children at each age. This gives a better indicator of the phonemes acquired generally by children as opposed to a particular child's frequency of use or idiosyncratic route.

In summary;

- At age 2 the word initial phonemes used by all 5 of the children were; /b, m, h, k, d, t, ð, n, j, ɪ, l, and əʊ/;

- At age 3 the word initial phonemes used by all 5 of the children were; /w, h, m, d, ð, b, t, k, g, p, ɪ, n, r, l, j, ɒ, aɪ, ə, əʊ, ʌ, eɪ/;

- At age 5 the word initial phonemes used by all 5 of the children were;

  /w, h, b, m, s, ð, l, d, k, t, r, p, n, f, g, ɪ, j, ʃ, ə, aɪ, st, tr, dʒ, æ, əʊ, ɒ, br, ɔ, ʌ, θ, fr, pl, dr, ɑ, eɪ, aʊ/;

7.3.1.4 Word Structure

As can be seen from Appendix 7.10 word structures ranging from one to 7 phonemes in length at age 2, from one to 7 at age 3 and from one to 9 phonemes at age 5 were found perhaps indicating that as the children developed the length of the words they tended to utilise got slightly longer.

The most frequent word type length was 3 phonemes at ages 2 and 3 and 4 phonemes at age 5 indicating again that as they developed the children more frequently utilised longer forms.

At age 2 the total of 256 word types were represented by 49 different word structures. With vowel phonemes represented as V and consonant phonemes as C table 7.7 shows the most frequent word structures observed.

Table 7.7 – English Child Word Structures – Age 2

Frequency of Structure	Word Structure
61	CVC
20	CVCV
20	CVCC
20	CV
16	CVVC
13	CVV
13	CCVC
12	VC
9	CVCVC
6	VV

At age 3 the total of 411 word types were represented by 58 different word structures.

Table 7.8 shows the most frequent word structures observed.

Table 7.8– English Child Word Structures – Age 3

Frequency of Structure	Word Structure
94	CVC
32	CVVC
29	CVCC
26	CVV
26	CVCVC
25	CVCV
23	CV
15	CCVC
12	VC
10	CVCCVC

At age 5 the total of 762 word types were represented by 109 different word structures. Table 7.9 shows the most frequent word structures observed.

Table 7.9 – English Child Word Structures – Age 5

Frequency of Structure	Word Structure
155	CVC
58	CVCC
50	CVVC
43	CVCV
41	CVCVC
31	CVV
27	CCVC
25	CV
19	VC
15	CVVCVC
15	CVVCC

At all three ages the most frequently used word structure was thus found to be CVC. This matches the most frequent adult English word structure observed in the data.

At age 2, a feature of early child speech, re-duplication, where a first syllable structure is repeated may explain the high frequency of CVCV structures (position 2). By age 5 the children are utilising a range of multi-syllabic word structures that more closely match the adult word structure usage.

It is interesting to observe the high frequency of consonant cluster codas even from the earliest age under assessment. A consonant cluster was ranked 3rd even at age 2 and a consonant cluster was ranked 2nd at ages 3 and 5. Word initial consonant clusters would here be represented as C alongside other consonantal phonemic units but a reference to the earlier phoneme frequency tables above also indicates that consonant clusters are a feature of English children's speech even from the earliest age under assessment.

7.3.2 English Child FUSE Ranking

Each of the five children presents their own set of minimal pair groupings at each age under assessment. These are based upon their own individual use of the phonemic system and the words they produce. These are provided in Chapter 9.

In addition to the individual children's FUSE rankings for each of the phonemes at each age two sets of combined FUSE total rankings were created. The first analysis assessed each individual child's particular FUSE total (based upon the range of words in use by a particular child) and then added the phoneme totals for the same phoneme by all the other children at that age together. This combined FUSE count, Child FUSE count A, takes into account individual word range usage.

Chapter 9 utilises the individual FUSE totals further in order to assess whether there is a degree of similarity between the children's minimal pair usage.

The second child FUSE analysis adopted the approach taken for the adult data samples whereby the word types of all the children combined were processed such that minimal pairs within the children's words overall could be assessed. With this approach a group FUSE count was produced for each phoneme thus enabling the phonemes to be ranked in a similar way to the adult data. The overall group rank based upon the group FUSE counts and with each contrastive phoneme counted once (i.e. not for minimal pair string length) are presented fully in Appendix 7.11.

In summary, at age 2 the following phonemes were ranked highest for WI FUSE;

/b, t, h, m, k, s, ð, g/.

At age 3 the following phonemes were ranked highest for WI FUSE;

/w, b, h, m, ð, d, tr, t, k, s, g, f/.

At age 5 the following phonemes were ranked highest for WI FUSE;

/b, t, m, h, k, w, r, s, ð, d/

An initial look at the top ten rankings of the adult data compared to these group child findings tends to suggest that there is some similarity between the adult and child FUSE rankings. At age 5 the English children are only using two phonemes /ð/ and /d/ which are not in the adult top ten ranking. The correlation assessments provided in Chapter 8 should enable this relationship to be more clearly observed.

7.4 Finnish Child Results

7.4.1 General Findings

The spoken words of five children at ages from 2 years to 7 years were extracted from the Oulu corpora. The children varied in their use of language in term of the word forms and phonemes utilised. They also varied in their volume of words spoken at each age, as would be expected. The individual child findings are explored in Chapter 9.

It is interesting to compare the words most frequently spoken by the children with the words most frequently spoken by adult Finnish speakers and represented in the larger adult Finnish corpora and word listings. It is also interesting to compare and contrast the phonemic contents of the words uttered by Finnish as compared to English children.

The lists of children's word types at each age were combined into one list and identical word forms (where 2 children used an identical word) were reduced to one representation of the word type in the group file. For example, if the children's word totals for age 3 were simply combined a total of 760 words would exist in the list however there are only 585 different words (i.e. word types). With the words grouped in this way a total of 581 word types were processed for age 2, 585 word types for age 3 and 820 for age 5.

7.4.1.1 Word Types

Unlike the English child data where the children were found to demonstrate the same most frequent word types the Finnish children demonstrated individual preferences for certain words at each stage of development. Each child therefore has a different frequency ranking for the individual words.

At age 2, for example, there was no common word used by all the children from the top ten words of each. With all 581 word types analysed only the words *ei, en, joo, on, pannaan, siellä, tiedä* and *tuo* were found to used by all five children. At age 3 words used by all the children were *ei, katso, kun, minä, olen, on, se, tämä, tässä.* At age 5 words used by all the children were *ei, en, ja, kun, me, minä, niin, ole, olen, on, se* and *tuo.*

At age 5 the most frequent words are beginning to reflect the words most frequently seen in adult spoken language with the words *en* 'no', *on* 'is', *se* 'it/he/she' being found in the top ten words of all 5 children.

As with the English data there is degree of similarity between the words used by children at each age however there are less commonly used words with the Finnish children. This may be reflecting the highly inflected and agglutinative nature of the Finnish language whereby many more word variations may be presented. The individual word totals totalled for age 2 suggests 982 words however when word types are counted only 581 remain (i.e. 41% of the words are produced by more than one of the children). After this age the common words as a percentage of the total number of

word types is less than the English findings. For age 3 the individual word total of 760 is reduced to 585 word types (24%) and for age 5 1057 is 820 word types (23%).

7.4.1.2 Phoneme Frequencies

Once the word files of word types for the children grouped together at each age had been created the frequency of phonemes for the children as a group at each age could be counted.

The phoneme frequencies have been counted by position within word and overall (including all word positions). As the emphasis of this study remains the phonemic system and the relative importance of phonemes within that system (rather than word frequency) each word type was presented once for the count rather than the count being made of a running stream of phonemes. That is the individual word type frequencies have not been included in assessing the phoneme frequencies.

The results of the frequency assessments for the three ages are shown in Appendix 7.12. These give the frequencies of particular phonemes for each word position. The table enables the range of word lengths to be seen, as well as the range of phonemes by each word position.

In summary, at age 2 the ten most frequent phonemes over all word positions were;

/ɑ, n, t, i, k, l, e, s, o, u/

Only one of these phonemes /u/ does not occur in the adult top ten phoneme usage ranking perhaps suggesting that already the frequency of usage of phonemes by the children, even at age 2, is close to that of adult usage.

At age 3 the ten most frequent phonemes over all word positions were;

/ɑ, t, i, n, k, s, l, o, e, u/

At age 5 the ten most frequent phonemes over all word positions were;

/ɑ, ɪ, t, n, k, l, s, e, u, o/

Eight of these also rank in the adult Finnish data top ten phoneme usage but the phonemes /u/ and /ɪ/ do not.

Appendix 7.12 shows that a range of 37 phonemes and phonemic units were being utilised by the children at age 2, a range of 39 at age 3 and 42 at age 5.

The phoneme /œ:/ was not used by any of the children at any of the three ages under assessment. At age 2 all of the singleton consonant phonemes and all the remaining long and short vowels were utilised. Only five word initial triphthongs and diphthongs and only one of the consonant clusters, /tr/, were utilised at this age. By age 5 the consonant clusters /br, pl, st, tr/ were also in use as were two more of the diphthongs /oi, ou/.

7.4.1.3 Word Initial Phonemes

In order to ascertain the usefulness of the earlier proposed theories, some general statistics about the phonemes seen in word initial position were collected. With the children treated as a group and only unique word types included the frequency rankings for the 3 ages can be seen in Appendix 7.13.

In summary with this approach at age 2 the most frequent word initial phonemes were;

/p, t, k, m, s, l, n, v, o, ɑ/

At age 3 the most frequent word initial phonemes were;

/k, t, p, m, s, n, v, l, o, h/

At age 5 the most frequent word initial phonemes were;

/t, k, p, s, m, l, v, o, h, ɑ/

Only the phoneme /o/, which occurs in the top ten word initial phoneme frequency rankings for all three ages under assessment, does not occur in the adult top ten word initial findings (it occurs at position 11).

With all the children's word initial phoneme frequencies combined the range of word initial phonemes used by the children varies from 28 different word initial phonemes at age 2 to 33 at age 5.

Another way of looking at this data is to assess which phonemes are used by all 5 of the children at each age. This gives a better indicator of the phonemes acquired by children as opposed to a child's idiosyncratic route. Chapter 9 analyses the individual children's usage of particular phonemes. It is interesting to observe that the WI phonemes utilised by all 5 of the children, rather than one or two children, are generally the most frequently observed in these group findings.

In summary;

- At age 2 the word initial phonemes used by all 5 of the children were; /t, p, k, m, s, o, n, l, v, ɑ, j, i, h, r, e and ei/;

- At age 3 the word initial phonemes used by all 5 of the children were; /t, k, m, p, s, n, o, v, l, j, e, r, i and ei/;

- At age 5 the word initial phonemes used by all 5 of the children were; /t, k, m, s, p, o, l, v, n, j, h, ɑ, i, e and ei/;

It is interesting to observe how the children seem to demonstrate a preference for certain long vowels in word initial position at different ages. At age 2, for example, only the long vowels /ɑɑ, ii and uu/ appear in the children's words, whilst at age 5 only the long vowels /ee and ää/ are present.

Diphthongs appear at age 2 with /ei/ being produced by all 5 of the children, /au/ by 4 of the children and both /ai and äi/ by two of the children at age 2. At ages 3 and 4 /ei/ is again produced by all 5 children. The only other diphthongs to be seen at age 3 are /au, ai and äi/ which are produced by 4 of the children. At age 4 /au, ei, ui, ai and äi/ are also produced.

Only two consonant clusters are seen at word initial position in the whole data sample. The consonant cluster /tr/ appears at each of the three ages but only for one child at ages 2 and 4 and for two children at age 3. The consonant cluster /pl/ appears at age 4 for only one child.

7.4.1.4 Word Structure

As can be seen from Appendix 7.12 word structures ranging from one to 14 phonemes in length at ages 2 and 3 and from one to 17 phonemes in length at age 5 were identified. The number of word types of each phonemic length found in the Finnish child data can be found by referencing the totals columns. All of these are considerably higher than the word lengths found for the English child data.

As can be seen the most frequent word type length was 6 phonemes for all the ages, followed by words of 5 phonemes at age 2 but by words of 7 phonemes at ages 3 and 5. All of these word lengths are higher than the findings for the English child word types.

A total of 166 different word structures were identified at age 2. With short vowel phonemes represented as V, long vowels as X, diphthongs as VV and consonant phonemes as C table 7.10 shows the most frequent word structures observed.

Table 7.10 – Finnish Child Word Structures – Age 2

Word Structure Frequency	Word Structure
66	CVCCV
28	CVCV
24	CVCVC
19	CVVCCV
14	CVVCV
14	CVCT
12	VCCV
12	CVV
11	CVCCVC
10	VCVC

If the long vowels (marked T) are taken as vowel phonemes and added to the above then an addition to the 66 CVCCV structures are 7 CVCCT structured words, 6 CTCCV words and 3 CTCCT words giving a total of 82 words with the structure CVCCV. The word structure CVCV can be increased by 24 (14 CVCT, 7 CTCV and 3 CTCT) to 52.

It is interesting to note that the seven most frequent word structures are all multi-syllable words with the first potentially single syllable words structures CVV and CVVC not occurring until ranks 8 and 18. The first single syllable word structure appears at rank 16 and interestingly includes a long or double vowel, CTC, rather than a single vowel (CVC). In fact, CVC, which occurred most frequently across all the English child word structures does not occur until rank 28. These findings reflect the adult word structure usage in particular of multi-syllable words and long word length.

A total of 188 different word structures were identified at age 3. Table 7.11 shows the most frequent word structures observed. Once again all of the most frequent word structure types are for multi-syllable words including CVCCV as the most frequent as at age 2. The common English child word structure CVC occurs at rank position 18 in the age 3 Finnish child data.

Table 7.11 – Finnish Child Word Structures - Age 3

Word Structure Frequency	Word Structure
51	CVCCV
40	CVCV
17	CVCVCCV
16	CVVCCV
16	CVCVC
14	CVVCV
13	CVCVCVC
12	CVCCVC
12	CVCCTC
11	CVVCVCCV

A total of 251 different word structures were identified at age 5. Table 7.12 shows the most frequent word structures observed.

Table 7.12 – Finnish Child Word Structures - Age 5

Word Structure Frequency	Word Structure
66	CVCCV
34	CVCV
25	CVCVC
24	CVCCVC
22	CVVCV
22	CVVCCV
18	CVCVCCV
17	CVVCVC
15	CVVCVCCV
15	CVCVCV

Across all the ages the most frequent word structure was found to be CVCCV, in all instances a two syllable word structure. As with the Finnish adult data there is more use of multi-syllable word structures than with the English data. The highest ranked one English child structure, CVC, across all three ages does not occur until rank position 40 at age 2, rank position 18 at age 3 and rank position 33 at age 5 in the Finnish child data.

7.4.2 Finnish Child FUSE Ranking

Each of the five children presents their own groups of minimal pairs based upon their own use of the phonemic system and the words they produce. Each child will therefore have their own phoneme frequency rankings for each age of assessment.
The fifteen individual child word initial FUSE ranks are presented in Chapter 9.

In addition to the individual FUSE findings the children's FUSE counts for each of the phonemes at each age have been combined into a group FUSE count which has also been ranked. The overall group rank, based upon a sum of the individual child FUSE counts and with each contrastive phoneme counted once (i.e. not for minimal pair string length) are discussed in Chapter 9.

Appendix 7.13 provides the child data FUSE rankings with one group child word file and the children's words taken as a group through the FUSE processing.

In summary, at age 2, the following phonemes were ranked highest for WI FUSE;

/s, m, t, p, n, k, l, v, j/

This shows an identical match to the top nine Finnish adult FUSE ranked phonemes.

At age 3 the following phonemes were ranked highest for WI FUSE;

/n, m, s, t, o, p, ɑi, ei, ɪ, j, r, u:, v, ui, e/

At age 5 the following phonemes were ranked highest for WI FUSE;

/s, m, t, n, v, j, k, l, o, p/.

As Appendix 7.13 shows a range of 15 phonemes were seen as the word initial contrast at ages 2 and 3 and 16 at age 4.

7.5 Summary of Results

This chapter has presented a range of different findings that have been produced from running the processing for FUSE assessment.

Two main types of data have been presented for each of the four areas under analysis and with the children treated as a group.

Firstly, some general findings regarding the word types and their frequencies, phoneme frequencies both at word initial position and over all word positions, and word structures have been presented. Table 7.13 presents a summary of the top ten ranked for frequency word initial phonemes for the four types of data.

Table 7.13 – Summary of Word Initial Phoneme Top Ten Ranks

Top Ten Adult English WI Phoneme Frequency	**Top Ten Adult Finnish WI Phoneme Frequency**
/k, s, ɪ, r, d, f, m, p, ə, b/	/k, t, s, m, p, v, l, n, h, j/
Top Ten Child English WI Phoneme Frequency	**Top Ten Child Finnish WI Phoneme Frequency**
Age 2 – /b, h, k, m, w, t, s, d, p, n/ Age 3 - /w,b, h, d, t, m, k, p, f, r/ Age 5 - /b, h, m, w, s, p, d, r, t, k/	Age 2 - /p, t, k, m, s, l, n, v, o, ɑ/ Age 3 - /k, t, p, m, s, n, v, l, o, h/ Age 5 - /t, k, p, s, m, l, v, o, h, ɑ/

It is interesting to note the higher usage of vowel phonemes by the Finnish children. Secondly, the findings that are more specifically related to the FUSE assessment have been presented such that the numbers of word types involved in minimal pair groups, the lengths of minimal pair strings and the range of contrastive phonemes involved can be identified.

The top ten ranked FUSE phonemes for the four areas under analysis are summarised in table 7.14.

Table 7.14 – Summary of Top Ten Ranked FUSE Phonemes

Top Ten Adult English WI FUSE Phonemes	**Top Ten Adult Finnish WI FUSE Phonemes**
/s, l, f, w, t, b, m, k, r, h/	/s, m, t, k, n, v, j, p, h, l/
Top Ten Child English WI FUSE Phonemes	**Top Ten Child Finnish WI FUSE Phonemes**
Age 2 – /b, t, h, m, k, s, ð, g/	Age 2 - /s, m, t, p, n, k, l, v, j/
Age 3 - /w, b, h, m, ð, d, tr, t, k, s, g, f/. /n, m, s, t, o, p, ɑi, ei, ɪ, j, r/	Age 3 - / u:, v, ui, e/
Age 5 - /b, t, m, h, k, w, r, s, ð, d/	Age 5 - /s, m, t, n, v, j, k, l, o, p/.

Chapter 8 will now assess the level of similarity found between the adult and child data for the two languages.

Correlation tests will be run on both the purely frequency based data (word initial phonemes) and on the new FUSE assessment measure to see whether any of these throw any light onto the role of the adult language in the children's phonemic development.

Chapter 8 : Application of the FUSE Method of Assessment

The FUSE method of assessment, as applied in this study, aimed to combine both a measure of the systemic usefulness of a phoneme to the phonemic systems of Finnish and English (i.e. each phoneme's ability to contrast between words) with a measure of the use that the languages make of this potential in terms of the most frequent words speakers use.

Now that the various frequency and FUSE rankings have been produced, for both adult and child data samples and for each language, they will be assessed for closeness of fit and similarity. This will be done by ranking the results and then completing correlation calculations using the Spearman rank correlation method (Coolican 1994). This will enable an analysis of the extent to which the FUSE method of assessment assists with an understanding of the relationship between phonological acquisition (as expressed in the child data samples) and ambient language (as expressed in the adult data samples). It will also enable the new FUSE measure to be compared directly with earlier purely frequency based theories of the interaction between phonological acquisition and ambient language.

It is accepted that the limited English and Finnish data samples that have been utilised for this study would need to be extended in order to apply the method of measurement more rigorously to more representative, larger data collections. The method of assessment should also ideally be applied to further different language samples before any conclusions about its usefulness could be drawn.

As defined in Chapter, 5 the first stage of the FUSE analysis will be to test whether the adult FUSE rankings can assist in explaining the patterns found in phonemic and phonotactic acquisition. At age 2 children would be expected to have already acquired many phonemes of their language but not necessarily all the phoneme sequences of consonant clusters and diphthongs/triphthongs (see Chapter 3). Their selection of words, which utilise these phonemes, will therefore be a direct demonstration of their acquisition of the phonological system as whole. The phonemes observed in the child

words relate to word differences and it is the contrastive abilities of phonemes together with the phonotactic rules of the language that determine the selection the child makes.

The first assessment will compare the ranking of the phonemes and phonemic units based upon frequency of use by the children as a group at the three ages with the adult FUSE rankings for the language they are acquiring. The closeness of the correlation between the children's frequency rankings and the FUSE ranking for the adult language will provide an indication of whether the children select phonemes that are functionally more useful and more frequently seen in the adult language or not. With this approach the FUSE calculations are being used as a prediction tool for phonological acquisition; the phonemes ranked the highest for adult FUSE are predicted to be utilised the most frequently by the children. A measure of how the correlation closeness changes over the three ages will provide an indication of the rate and direction of acquisition. These findings are presented in 8.1 below.

Another way of utilising the FUSE assessment for the exploration of phonological acquisition will be with the children's FUSE rankings as a measure of the children's use of a phonemic system. The children's group FUSE ranking for each age under assessment will be compared with the adult FUSE ranking to see the level of correlation at the three ages, the starting closeness of correlation and the direction of any correlation change. These findings are presented in 8.2 below.

A correlation of the frequency of adult word initial phonemes to child word initial phonemes for the children as a group at each of the three ages will be completed in 8.3. The correlation results for the two different applications of FUSE described above (i.e. adult FUSE to child frequency and adult FUSE to child FUSE) will be compared both with each other and also with the frequency to frequency correlation.

In order to observe the correlation findings in perspective of the two languages a measure of the two languages' starting similarities will also be made and presented in 8.5 below. This will be done firstly by comparing the two adult languages' frequency rankings and then by comparing the two adult languages' FUSE rankings. The rankings are here representing different and unrelated phonemic systems and word

usage. Any correlation that is found here between the two unrelated languages might actually be indicating some universal principles that apply to the two languages and possibly languages as a whole.

As an exploration of the usefulness of the various frequency and FUSE correlation measurements to test phonological acquisition the frequency and FUSE rankings for the English and Finnish children will be assessed for correlation. A close initial correlation between children acquiring Finnish and English may suggest that children select similar phonemes and phonemic sequences in the early stages, perhaps demonstrating a universal base of phonemes. Any initial closeness should be seen to reduce as the children develop towards the phonological system of the ambient language. It will be interesting to observe whether the FUSE rankings provide a way of assessing any movement of the system towards the ambient language system surrounding the children as the children develop. If the findings for both Finnish and English similarly show such a movement then this might provide evidence for a universal approach adopted by the children acquiring the two languages whereby the phonemic contrasts within the system are being sought out by the children and used more consistently as they develop their phonological systems.

Finally, in order to more thoroughly test the new method of assessment the adult rankings of both languages will be compared cross-linguistically to the child rankings of the oppositve language (i.e. adult English to child Finnish and adult Finnish to child English). Here a lower correlation than the intra-linguistic findings would be expected. It will be interesting to observe whether any movement away from the opposite adult language towards to the same adult can be detected.

In order to assess the amount of similarity between various ranking structures (child FUSE ranking, frequency ranking etc), the Spearman Rank Co-efficient will be used as a measure of correlation. Applying the correlation calculations provides an indication of how closely two rankings correlate and are therefore similar (Coolican 1994). Using standard tables (of rho measures) we can see whether a positive correlation at the chosen level of confidence is detected. Roughly speaking the closer the co-efficient produced is towards +1 the closer a correlation is indicated. This figure must be read in

light of the number of components that are actually being compared in order to give a level of confidence for the results.

In summary, the following rank correlation comparisons will be undertaken;

8.1 Adult WI FUSE to Group Child WI Phoneme Frequency

8.1.1 English Adult WI FUSE to Group Child WI Frequency

8.1.2 Finnish Adult WI FUSE to Group Child WI Frequency

8.2 Adult WI FUSE to Group Child WI FUSE

8.2.1 English Adult WI FUSE to Group Child WI FUSE

8.2.2 Finnish Adult FUSE to Group Child WI FUSE

8.3 Adult WI Frequency to Group Child WI Frequency

8.3.1 English Adult WI Frequency to Group Child WI Frequency

8.3.2 Finnish Adult WI Frequency to Group Child WI Frequency

8.4 Summary of Correlation Results

8.5 Cross-Linguistic Assessments

8.5.1 English Adult Frequency to Finnish Adult Frequency

8.5.2 English Adult FUSE to Finnish Adult FUSE

8.5.3 English Child Frequency to Finnish Child Frequency

8.5.4 English Child FUSE to Finnish Child FUSE

8.5.5 English Adult to Finnish Child

8.5.6 Finnish Adult to English Child

8.1 Correlation of Adult FUSE Rankings to Group Child Word Initial Phoneme Frequency Rankings

The first stage of the FUSE correlation analysis will be to see what FUSE indicates and whether the rankings can assist in explaining the children's phonemic acquisition routes.

The FUSE rankings that have been calculated for the two adult languages will be compared with the children's frequencies of phonemes at the three ages with the combined frequency rankings.

8.1.1 English Adult FUSE to Group Child Frequency

The word initial FUSE findings for adult English have already been presented in Chapter 7. The English children's group word initial phoneme frequencies have also already been presented as a combined frequency figure based upon word types for each age (Section 7.1).

As an initial observation on the combined frequency child ranks compared to the adult FUSE ranks the most frequently observed word initial phonemes that English children at age 2 used were found to be;

/b/ (which ranked 6.5 for adult FUSE) ,

/h/ (which ranked 10.5 for adult FUSE),

/k/ (which ranked 8th for adult FUSE),

/m/ (which ranked 6.5 for FUSE),

/w/ (which ranked 3.5 for adult FUSE).

Followed in decreasing order of frequency by;

/t/ (ranked 5 for adult FUSE), /s/ (ranked 1 for adult FUSE) ,

/d/ (ranked 12 for adult FUSE), /p/ (ranked 10.5 for adult FUSE) and

/n/ (ranked 15 for adult FUSE).

All of the most frequent word initial phonemes utilised by the children do occur within the top 15 adult FUSE rankings (out of a total of 74 ranks). The phonemes /h/, /d/, /p/ and /n/ appear to be used more frequently by the children than the FUSE systemic

importance ranks them. They occur within the top ten word initial child ranks but not within the top ten adult FUSE rankings. On the other hand, the phonemes /f/, /r/ and /l/ all ranked within the top ten for adult FUSE but did not appear within the child top ten word initial frequency ranking, being ranked 16, 10.5 and 20.5 respectively. All of these are used less frequently by the children at this age, and perhaps most significantly the phoneme /l/ which ranked 2nd for FUSE is only ranked at position 20.5 for the children at this age.

Another observation is that with the adult data very few of the vowels were observed in word initial position providing the contrast between words whereas the children's frequencies tended to reflect the use of more vowels in word initial position indicating more open syllables without a consonant onset.

At age 3 the children used mostly the word initial phonemes;

/w/ (ranked 3.5 for adult FUSE),

/b/ (ranked 6.5 for adult FUSE),

/h/ (ranked 10.5 for adult FUSE),

/d/ (ranked 12 for adult FUSE),

/t/ (ranked 5 for adult FUSE) and

/m/ (ranked 6.5 for adult FUSE).

Followed in decreasing order of frequency by;

/k/ (ranked 8 for adult FUSE) , /p/ (ranked 10.5 for adult FUSE),

/f/ (ranked 3.5 for adult FUSE), /r/ (ranked 9 for adult FUSE).

Out of this list of word initial phoneme frequency rankings only the word initial phoneme /d/ was found to be outside of the top 20 adult FUSE phonemes ranking again indicating a fair degree of similarity between the two rankings. Interestingly, the

phoneme ranked highest for adult FUSE, /s/, appears here only ranked at position 11 for frequency of usage and the second highest FUSE ranked phoneme, /l/, is ranked 13.5 for child word initial frequency usage.

At age 5;

/b/ (ranked 6.5 for adult FUSE),

/h/ (ranked.10.5 for adult FUSE),

/m/ (ranked 6.5 for adult FUSE),

/w/ (ranked.3.5 for adult FUSE),

/s/ (ranked 1 for adult FUSE),

Followed in decreasing order of rank frequency by;

/p/ (ranked 10.5 for adult FUSE) , /r/ (ranked 9 for adult FUSE),

/d/ (ranked 12 for adult FUSE), /t/ (ranked 5 for adult FUSE) and

/k/ (ranked 8 for adult FUSE).

This shows that once again the phonemes used the most frequently in word initial position are those that rank the highest for FUSE. Only one of the children's top ten phonemes, /d/, is not included in the adult top ten FUSE ranking. On the other hand, the phonemes /f/ and /l/ both rank high for adult FUSE but not for frequency. The highest ranked phoneme for adult FUSE is only positioned at 5th place whilst the second highest, /l/ is only ranked 11th for the children's word initial frequency rank.

Whilst these initial comparisons of the top ten ranked child word initial frequency and adult FUSE phonemes enable us to observe some relationship between the adult FUSE and the child frequency rankings it is difficult to asses how strong the relationship is at the three ages and how it may be developing as the children get older. In order, therefore, to more thoroughly assess the amount of similarity for these adult FUSE to combined child frequency rankings the Spearman Rank Co-efficient has been used as a measurement of correlation. The full findings are given in Appendix 8.1 and show how

the various ranks correlate both individually and then overall for the whole ranking structure and for the three ages.

As can be seen, all the correlation results show a good amount of correlation. In summary at age 2 the correlation between the 2 ranks, adult FUSE and overall child frequency, was 0.74, at age 3 it was 0.77 and at age 5 it had reduced slightly to 0.72. With the total of 77 ranks correlated in this way we can be confident at a high level that the rankings are truly correlated even at the earliest age of 2 years. The correlations are only marginally different over the three ages but interestingly the child frequencies do more closely correlate at the middle of the ages under assessment than for age 2 (0.74) or the latest age, age 5 (0.72). The average correlation measure across the three ages is 0.74.

The results overall show that the children as a group tend to use more frequently the phonemes that rank high for FUSE even at age 2. As the frequencies and range of the phonemes utilised increases as the children get older they move slightly closer towards the adult usage and functional load represented by FUSE.

8.1.2 Finnish Adult FUSE to Group Child Frequency

The adult Finnish FUSE rankings and the group child frequency have already been presented in Chapter 7.

Initial observation shows that at age 2 the phonemes most frequently observed in word initial position were;

/p/ (ranked 8 for adult FUSE),

/t/ (ranked 3 for adult FUSE),

/k/ (ranked 4 for adult FUSE),

/m/ (ranked 2 for adult FUSE) and

/s/ (ranked 1 for adult FUSE).

Followed in decreasing order of rank frequency by;

/l/ (ranked 10 for adult FUSE), /n/ (ranked 5 for adult FUSE),
/v/ (ranked 6 for adult FUSE), /o/ (ranked 11 for adult FUSE) and
/ɑ/ (ranked 15 for adult FUSE).

Even at this age, the earliest age of assessment for this study, all of the five highest ranked phonemes for frequency at word initial position by the children rank within the top ten for adult FUSE indicating a fair degree of correlation between the two rank orders. Interestingly two vowel phonemes /ɑ/ and /o/ ranked high for frequency of usage by the children but did not appear in the top ten adult FUSE rankings being replaced by /j/ and /h/ instead.

At age 3 the following phonemes were ranked the highest for word initial frequency and usage by the children;

/k/ (ranked 4 for adult FUSE),
/t/ (ranked 3 for adult FUSE),
/p/ (ranked 8 for adult FUSE),
/m/ (ranked 2 for adult FUSE),
/s/ (ranked 1 for adult FUSE),

Again, all five of these rank within the top 10 for adult FUSE.

Followed in decreasing order of rank frequency by;

/n/ (ranked 5 for adult FUSE), /v/ (ranked 6 for adult FUSE),
/l/ (ranked 10 for adult FUSE), /o/ (ranked 11 for adult FUSE) and
/h/ (ranked 9 for adult FUSE).

At age 3, as can be seen, all apart from one phoneme that ranked in the children's most frequent word initial phonemes was not within the top adult FUSE rankings. This phoneme, /o/, ranked at position 11 for adult FUSE, once again as at age 2, replaced the phoneme /j/.

At age 5 the following phonemes were ranked the highest for children's word initial frequency usage;

/t/ (ranked 3 for adult FUSE),

/k/ (ranked 4 for adult FUSE),

/p/ (ranked 8 for adult FUSE),

/s/ (ranked 1 for adult FUSE),

/m/ (ranked 2 for adult FUSE),

Again, all five of these rank within the top 10 for adult FUSE.

Followed in decreasing order of rank frequency by;

/l/ (ranked 10 for adult FUSE), /v/ (ranked 6 for adult FUSE),

/o/ (ranked 11 for adult FUSE), /h/ (ranked 9 for adult FUSE) and

/ɑ/ (ranked 15 for adult FUSE).

Interestingly, at age 5, the two vowel phonemes /ɑ/ and /o/ which also ranked high for frequency of usage by the children at age 2 once again occurred within the top ten frequency rankings. Neither of these vowel phonemes occurred within the top ten adult FUSE rankings being replaced by /j/ and /n/ instead. The phoneme /j/ was consistently used less frequently than other phonemes which ranked higher for adult FUSE.

Whilst just a preliminary overview at this stage these results do seem to suggest that the Finnish children as a group do seem to utilise phonemes that have been found to rank the highest for FUSE.

Appendix 8.2 shows how these combined word initial child phoneme frequency rankings correlate with the adult FUSE rankings. In summary at age 2 the correlation between the 2 ranks (FUSE and overall frequency) was 0.73, at age 3 0.77 and at age 5 0.74. Comparing the Finnish adult FUSE rankings with the child group indicates a very close correlation at all three of the ages under assessment and a slight move towards the adult FUSE as the children develop over the three ages.

The children at age 2 are already demonstrating a preference (in terms of the overall frequency of word initial phonemes) for those that ranked highest for FUSE as seen by the correlation of 0.73. This close correlation remains over the three ages under investigation, increasing to .77 at age 3 and reducing slightly to .74 at age 5. The average correlation here is .74 the same as the average English correlation finding when using the child frequency to adult FUSE findings method of assessment.

8.2 Correlation of Adult FUSE Rankings to Group Child FUSE Rankings

The second correlation test compared the adult FUSE ranking with the group child FUSE rankings for the three ages. Both of these rank findings were provided in Chapter 7.

8.2.1 English Adult FUSE to Group Child FUSE

The FUSE rankings for adult English are provided in Section 7.1 above. In summary, the ten phonemes ranked the highest for FUSE i.e. judged to be the most important phonemes for the adult English system to differentiate between words, were found to be;

/s, l, f, w, t, b, m, k, r, h/

The phonemes ranked the highest for FUSE for the English children at age 2 were;

/b, t, h, m, k, s, ð, g/

This initial assessment shows that 8 out of the top ten from both adult and child FUSE are common to both top ten ranks. The child data, however, shows a more important role for the phonemes /ð/ and /r/ than the phonemes /l/ and /f/ which the adult data ranks in the top ten.

At age 3 the highest ranked phonemes were /w, b, h, m, ð, d, tr, t, k, s, g, f/. Again, as at age 2, the children continue to rank the phoneme /ð/ in the top ten ranked phonemes. Whereas at this age the phoneme /f/ is included in the top ten for the child FUSE ranking, the phoneme /l/ once again does not get included. Additionally the children rank the phoneme /d/ as well as the consonant cluster /tr/ whereas the adult data shows a higher ranking for /r/.

At age 5 the highest ranked phonemes were /b, m, t, h, k, w, r, s, ð, d/. The children at age 5 continue to substitute the adult top ten ranked phoneme /r/ and /l/ with the phonemes /d/ and /ð/ however it is interesting to observe that both of these phonemes have moved down in the ranks considerably over the three ages.

Appendix 8.3 provides the detailed correlation assessments for the adult English FUSE ranking when compared with the child English FUSE ranking showing the correlation for the children as a group at each age.

In summary the correlation findings moved from 0.69 at age 2, to 0.71 at age 3 and then up to 0.73 at age 5. Whilst there is only a small amount of change over the three ages these results do show a gradual increase in correlation over the three ages perhaps indicating that as the children's phonemic systems develop and the lexical base expands the children's FUSE rankings do more closely match the adult FUSE rankings. The average correlation over the three ages is .71.

Interestingly, these findings show a lesser similarity at age 2 than the adult FUSE to child frequency findings (0.69 as compared to 0.74) and a lower average correlation. However, the direction towards the adult language by the children is more clearly

observed using the FUSE method of assessment than with the purely frequency based assessment.

8.2.2 Finnish Adult FUSE to Group Child FUSE

The FUSE rankings for adult Finnish are provided in Section 7.2 above. In summary, the ten phonemes ranked the highest for FUSE i.e. judged to be the most important phonemes for the adult Finnish system to differentiate between words, were found to be;

/s, m, t, k, n, v, j, p, h, l/

The phonemes ranked the highest for FUSE for the English children at age 2 were;

/s, m, t, p, n, k, l, v, j/

Interestingly, there appears to be a very close correlation between the two most frequent rankings of children at age 2 and the adult Finnish FUSE rankings. All of the children's highest FUSE ranked phonemes also appear in the adult top ten FUSE ranking and only one of the adult top ten is missing, the phoneme /h/.

At age 3 the phonemes ranked the highest for FUSE for the Finnish children were found to be;

/n, m, s, t, o, p, ɑi, ei, ɪ, j, r, u:, v, ui, e/.

Initially, looking at the two rankings compared it appears that there is less of a correlation between the Finnish children at age 2 and the adult FUSE than at age 3. The child rankings show a less significance of the phonemes /k/ and /l/ than the earlier age and a number of vowels and diphthongs have now appeared in the top ten ranking.

At age 5 the phonemes ranked the highest were /s, m, t, n, v, j, k, l, o, p/. At this age the Finnish children appear to have returned to the earlier close match of top ten FUSE phonemes although once again the highly ranked adult FUSE phoneme /h/ is now replaced with the vowel phoneme /o/.

Appendix 8.4 provides the detailed correlation assessments for the adult Finnish FUSE ranking when compared with the child Finnish FUSE ranking showing the correlation for the children as a group at each age.

In summary the correlation findings moved from 0.70 at age 2, to 0.64 at age 3 and then up again to 0.78 at age 5. At age 3 there appears to be a change in direction of the correlation, however at age 5 there is a higher correlation than seen so far with any of the correlation tests. The average correlation over the three ages is .73 which is slightly higher than the English average FUSE correlation (0.70) and the same as the average Finnish adult FUSE to child frequency correlation findings.

8.3 Correlation of Word Initial Phoneme Frequency Rankings

As a direct comparison with the new FUSE findings the adult word initial phoneme frequency findings were compared with the child word initial phoneme frequency findings. Both of these have already been presented in Chapter 7.

8.3.1 English Adult to Child Word Initial Phoneme Frequency

Appendix 8.5 shows how the combined group child word initial phoneme frequency rankings (provided in 7.3.1) correlate with the adult word initial phoneme frequency rankings (provided in 7.1.1).

In summary at age 2 the correlation between the 2 ranks (adult and child word initial frequency rankings) was 0.81, at age 3 0.84 and at age 5 0.81. This shows a slightly closer correlation than with the FUSE findings above but as with the FUSE to frequency findings (see 8.2.1 above) does not seem to reflect the move of the children

over the three ages towards the adult system as the FUSE based method of assessment does.

These findings seem to support previous frequency based studies (Section 3) and demonstrate a significant level of correlation between the child's usage of word initial phonemes and the adult language data sample. However, few studies have used such correlation methods when considering the closeness and few have looked specifically at particular word positions for their assessments as this study has. Also, previous frequency assessments have often been based around word tokens rather than types.

The average correlation measure across the three ages is 0.82 and this will be used for comparison with later correlation findings.

It is accepted that with this group child frequency ranking, which has been produced from a group word type basis, the word initial phonemes that have been counted from this set, have not made allowance for the idiosyncratic use of phonemes by individual children. As discussed in Chapter 7 some phonemes are used infrequently, and so would be ranked low, but are used by all the children in the group under assessment whilst others which are used more frequently overall but by less children would be ranked higher. With this approach a greater range of phonemes was measured and compared to the larger adult phoneme list than was necessarily used by many of the children.

As a further comparison the individual frequency rankings for the different children were also compared with the overall adult frequency rankings in order to assess whether certain children individually demonstrate stronger or weaker relationships. These findings are given in Chapter 9.

8.3.2 Finnish Adult to Child Word Initial Phoneme Frequency

Appendix 8.6 shows how the combined child frequency rankings correlate with the adult frequency rankings.

In summary at age 2 the correlation between the ranks (adult and child word initial frequency rankings) was 0.85, at age 3 0.84 and at age 5 0.83. The average correlation measure across the three ages of 0.84 is very similar to the English average of .82. Whilst the individual correlation findings are high, and are indeed higher than any of the previous findings for Finnish, they do not seem to reflect a move towards the adult system that you would expect the children to be displaying.

It is interesting to observe that using the child phoneme frequencies combined in this way seems to indicate that the child frequencies less closely resemble the adult frequencies the older the children are. Whilst it is accepted that word samples under investigation are small (5 children) and the amount of difference is low these findings appear to contrast both with previous child language frequency findings quoted in Section 3 (which have also used individual case study or small groups of children) and the English frequency findings (of the same number of children) which appeared to show English children developing their word base and increasingly utilising the phonemes that rank high for frequency in the adult language.

As has already been stated much of the frequency correlation work has not looked specifically at non Indo-European languages and few studies have used correlation methods within word position for their assessments. Variations are always expected in real data and the movement is only slight however using frequency based assessments for Finnish does not seem to offer much of an insight into the path of the developing phonological system moving towards adult frequencies.

8.4 Summary of Correlation Results

Table 8.1 shows the English correlation results obtained by comparing the three different sets of rankings (i.e. the adult FUSE ranking to child word initial phoneme frequency ranking, adult FUSE to child FUSE ranking and the adult word initial phoneme frequency to child word initial phoneme frequency ranking) for the three ages of child data assessment. Table 8.2 shows the same for the Finnish correlation results.

Table 8.1 – Summary of English Correlation Results

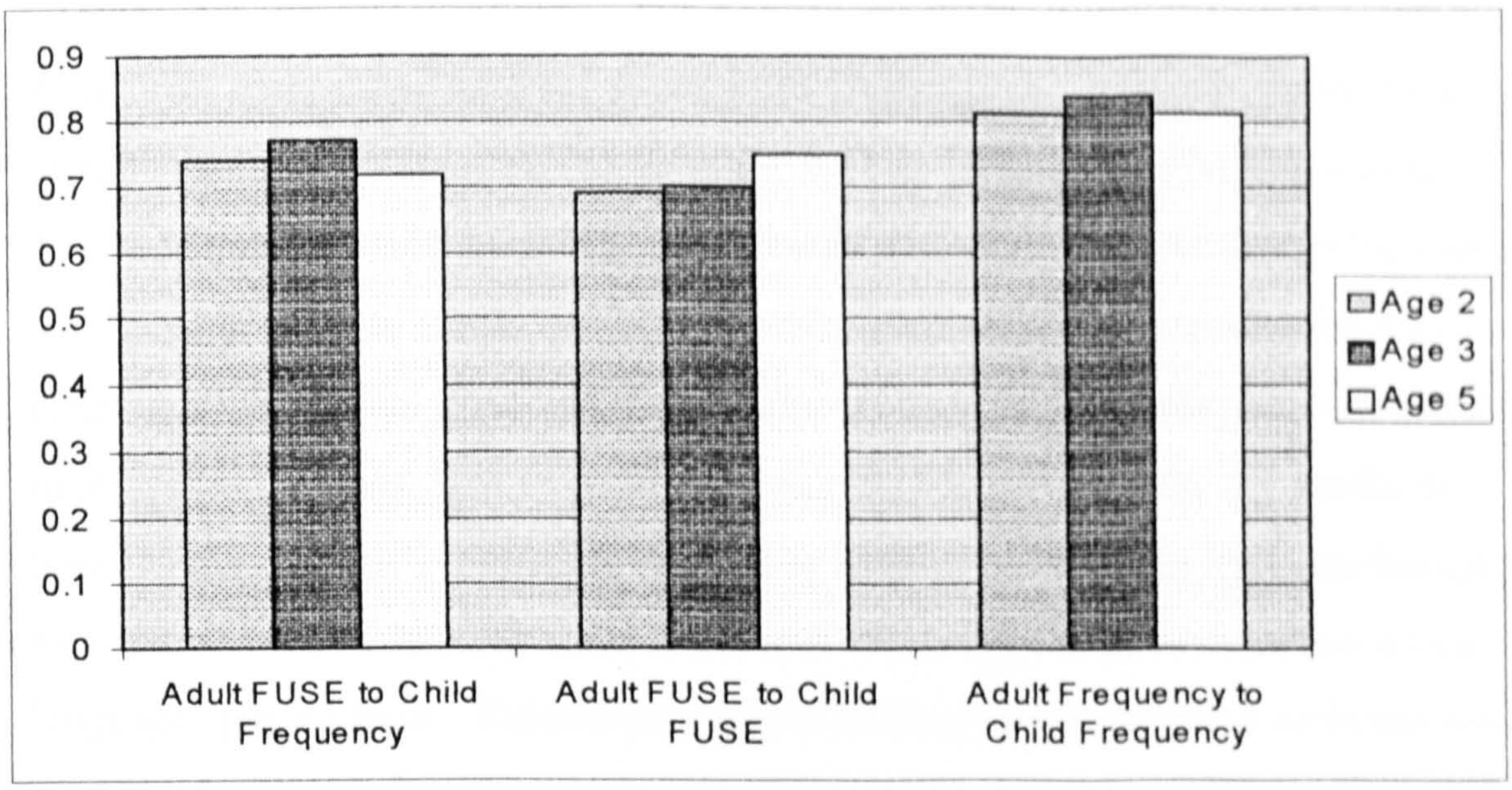
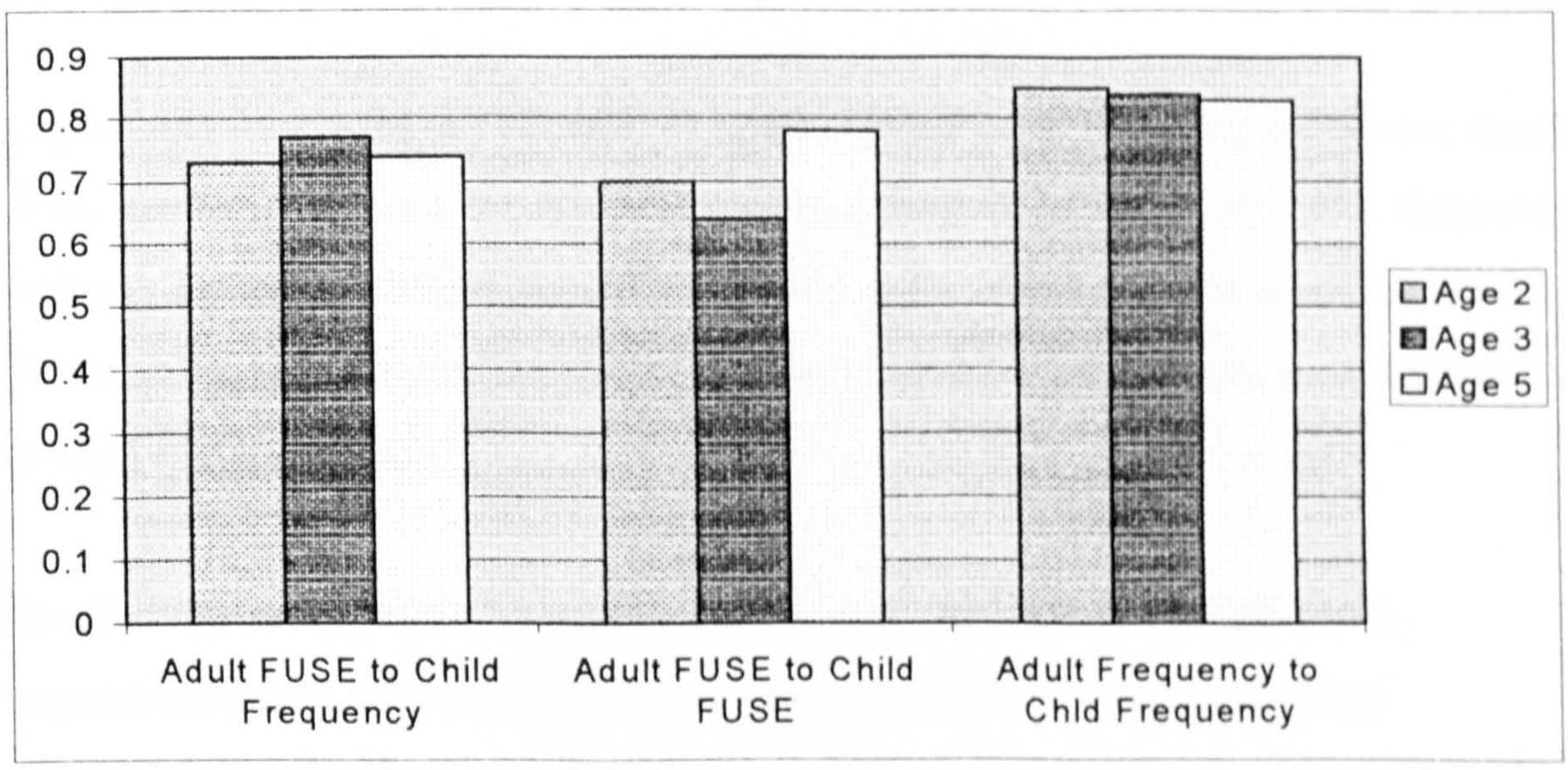


Table 8.2 – Summary of Finnish Correlation Results



Both the movement of the correlation results over the three ages as well as the level of correlation at the three ages and with the different correlation assessments can be observed.

The closer frequency based correlation for the two languages over the three ages can be clearly seen as can the movement of correlation by the child results towards the adult FUSE ranking when using the FUSE method of assessment.

8.5 Cross-Linguistic Assessments

As a direct comparison with the above intra-language correlation findings it was thought useful to compare the findings for the two languages cross-linguistically. The two adult languages word initial phoneme frequency and FUSE rankings were correlated in order to assess the level of underlying similarity between the two languages word initial phoneme usage. The group child phoneme frequency and FUSE rankings for the three ages were also correlated to assess not only the similarity between the children's phonemic systems at each of the ages but also whether either frequency or FUSE ranking assist at all with observing the movement towards a language specific basis. Finally the adult to child opposite language rankings were correlated.

8.5.1 English Adult Frequency to Finnish Adult Frequency

English and Finnish have different phonemic systems and therefore a different number of phonemes for which a word initial frequency ranking would apply. The frequency rankings for the two languages differ not only in the actual rank for a particular phoneme but also the number of ranks (relating to the number of phonemes occurring in word initial position).

The findings in Appendix 8.7 show a much lower level of correlation cross-linguistically than with any of the other intra-language correlation findings. The correlation findings on this basis show only a 0.26 correlation. A total of 137 phonemes from the two languages were involved in word initial frequency rankings (79 from the English findings and 58 from the Finnish findings). Out of this total 100 were found to exist for only one language or the other and only 37 were common to both languages.

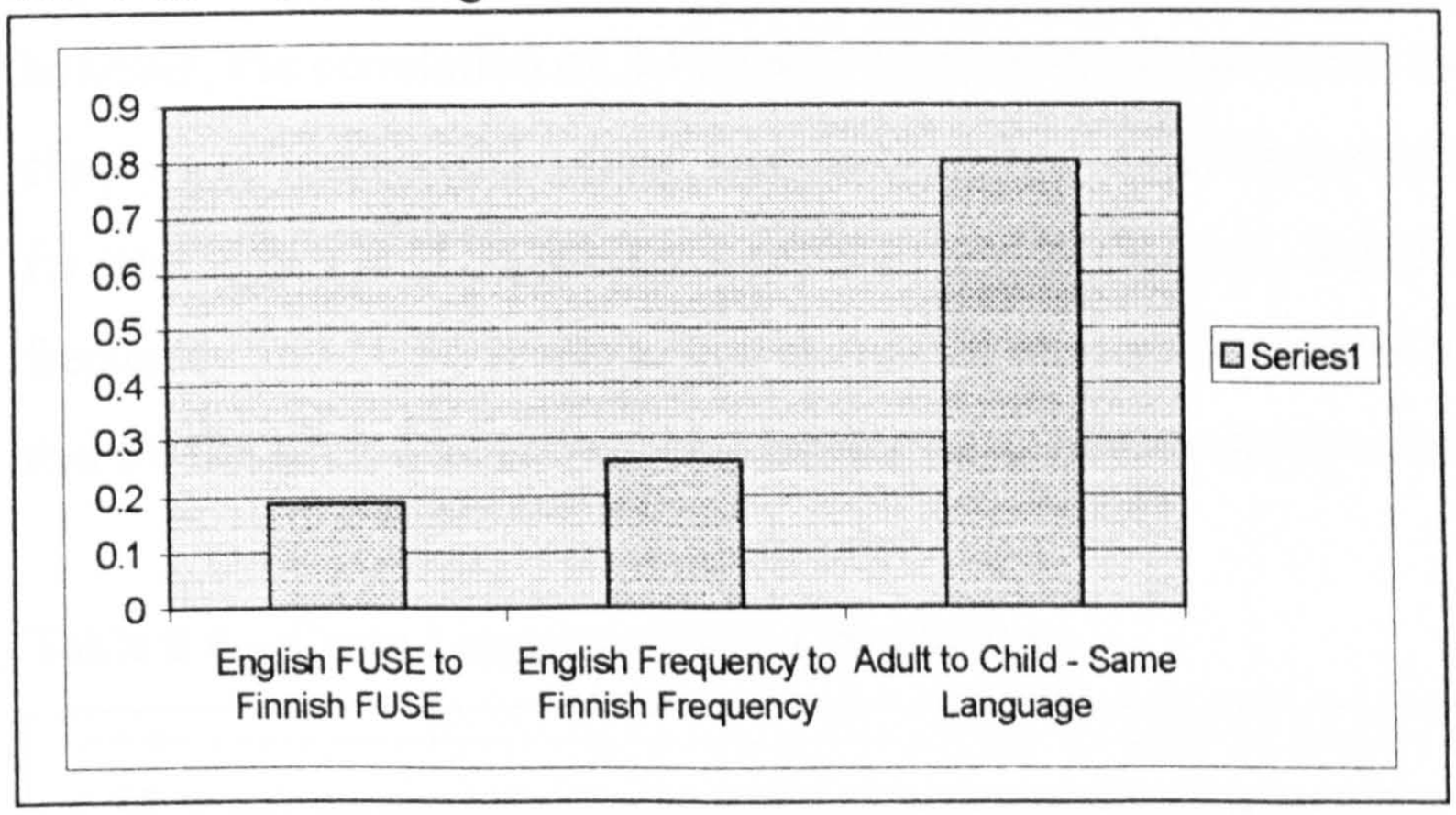
8.5.2 English Adult FUSE to Finnish Adult FUSE

In order to test whether inherent similarity or underlying commonality exists between the two target language systems, perhaps indicating some universal features of the

languages under review, the correlation between the adult FUSE rankings for the two languages were also compared. English and Finnish have different phonemic systems and therefore a different number of phonemes for which FUSE could apply. The FUSE rankings for the two languages differ not only in the rankings but also the number of ranks (relating to the number of phonemes involved as contrasts).

All the phonemes with a rank for FUSE for both languages were listed with their FUSE rankings. This gave a total of 120 phonemes with ranks (46 for Finnish and 74 for English), only 30 of which are common to both languages. The Finnish and English rankings were correlated to see whether any underlying similarity was being detected. Appendix 8.8 provides the results. In summary only a .19 correlation was found which is considerably lower than the comparisons between like language results provided above. Table 8.3 shows the frequency and FUSE levels of correlation between the two adult language samples and enables the lower correlation to be assessed against a sample of the same language findings presented above.

Table 8.3 – Cross-linguistic Adult Correlations



8.5.3 English Child Frequency to Finnish Child Frequency

The combined frequency rankings for the three ages of age 2, 3 and 5 present another set of figures which can be correlated. Here any correlation might be indicating some underlying universality in the children's usage of word initial phonemes across the two languages.Appendix 8.9 provides the findings. In summary at age 2 only a .29 correlation was found on this basis between the children's word initial phoneme

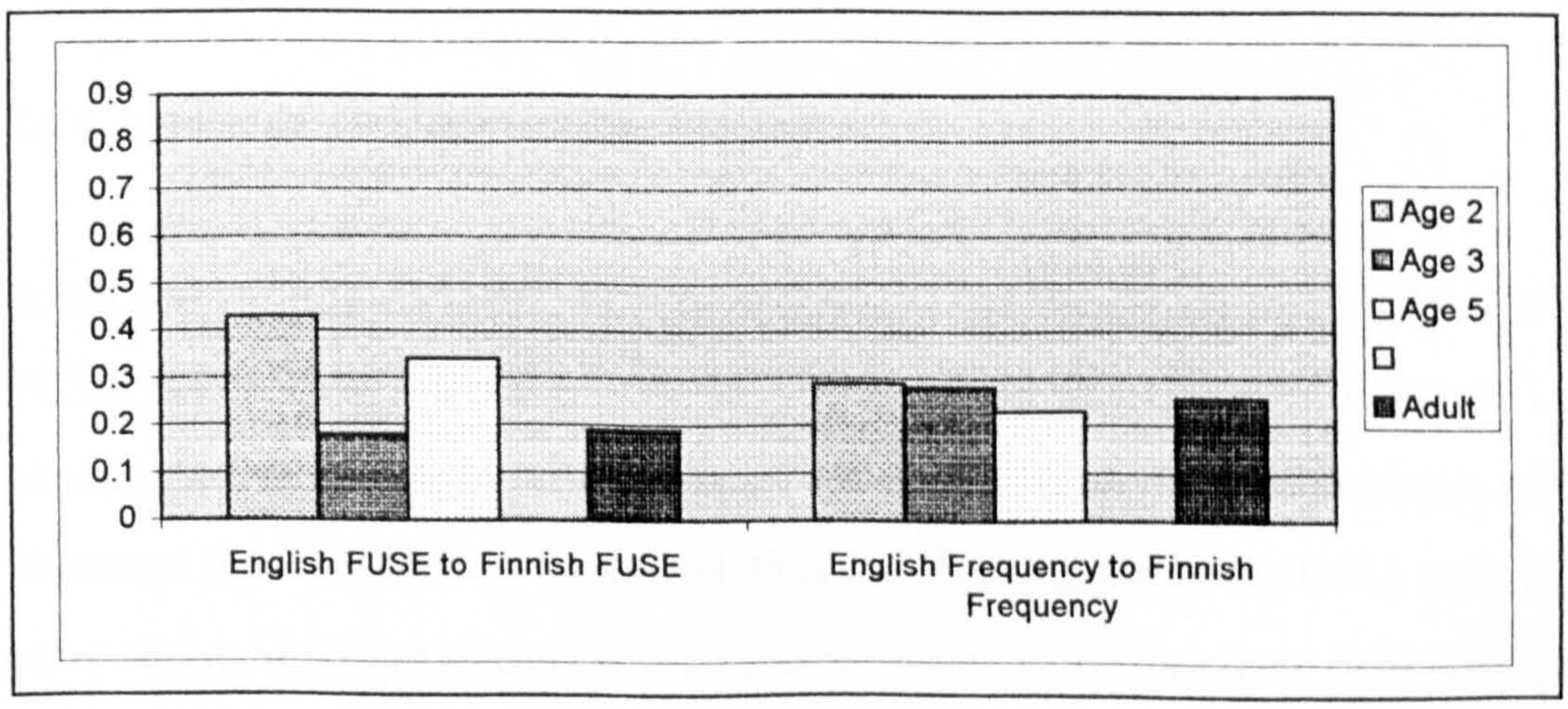
frequency, at age 3 a 0.28 correlation and at age 5 a 0.23 correlation. Whilst once again only small samples are being tested here and only for three ages these results do seem to indicate that as the children develop their phonological systems they start to use more language specific phonemes and move away from the more universal basis reflected in the slightly higher correlation at the earlier ages.

8.5.4 English Child FUSE to Finnish Child FUSE

The two sets of FUSE rankings for the two languages and the three ages of age 2, 3 and 5 present another set of figures which can be correlated. Here any correlation might be indicating some underlying universality in the children's contrastive usage of word initial phonemes across the two languages. Appendix 8.10 provides the findings.

In summary at age 2 a .43 correlation was found which is higher than the age 2 cross-linguistic frequency correlation and could be indicating a closeness in the contrastive phonemes selected by the children at this age. At age 3 a 0.18 correlation might suggest a move away from the more common base evidenced at age 2. At age 5, however, the correlation of .34 could be indicating a movement back again or might simply a reflection of the small size samples that are being tested here. Further analysis for later ages would be needed in order to further test this. Table 8.4 shows the frequency and FUSE levels of correlation for the two languages at the three child ages and for the adult data and enables the direction of movement to be observed.

Table 8.4 – Cross-Linguistic Child Correlations



8.5.5 English Adult to Finnish Child

The English adult FUSE and frequency rankings were correlated with the Finnish child FUSE and frequency rankings for each of the three ages under assessment. Appendix 8.11 provides the FUSE correlation findings and Appendix 8.12 the word initial frequency correlation findings .

In summary at age 2 a .52 FUSE and a .44 frequency correlation was found, at age 3 a .36 FUSE and a .39 frequency correlation was found and at age 5 a .42 FUSE and .27 frequency correlation was found. All of these are significantly lower than the adult to child same language findings. They also all show a movement away from the opposite adult language as the children develop their phonological systems towards the language specific system of the language of acquisition.

8.5.6 Finnish Adult to English Child

Appendix 8.13 provides the FUSE correlation findings and Appendix 8.14 the word initial frequency correlation findings for Finnish adult to English child . In summary at age 2 a correlation of .22 for FUSE and .20 for frequency was found reducing at age 3 to correlations.04 and .12 and at age 5 to .15 and .14. Again these findings indicate a significantly lower correlation than the adult to child same language correlation findings. They represent the lowest correlation findings over all the correlation tests but still indicate a movement away from the opposite adult language as the children develop their phonological systems.

8.6 Comments on Correlation Testing

So far the various correlations between FUSE and frequency and FUSE and child FUSE have looked specifically at the phonemes involved in minimal pairs i.e. those phonemes that appear in minimal pairs will have a rank. The determinate of whether a phoneme was included in the correlation was that the phoneme had a rank in one or other of the ranking tables (i.e. it appeared in one of the two sets being compared). Those phonemes that existed in the child frequency rankings (where frequency but not

contrast was taken) that didn't have an adult rank (no contrast in minimal pairs) were added as a 0 frequency at the lowest ranked position in the adult ranks for the correlation. Those phonemes in the adult ranks that did not occur in the children's frequency ranks were added to the children's frequency ranks at the lowest level. For the FUSE to FUSE correlations those child FUSE rankings that did not have phonemes presented in the adult FUSE rankings had the phonemes added at the lowest level below FUSE frequency one phonemes. Another approach would be to look at the whole possible phonemic system and include all the possible phonemes in both rankings thereby giving the phonemes that did not occur in any of the systems a status in the correlation processing.

Another approach again would be to look specifically at the phonemes that occurred in word initial position but did not appear in the FUSE rankings (e.g. some of the diphthongs in English that were allowed for as they existed in word initial position but did not occur as contrasts in minimal pairs). It would be expected that both of these approaches would in fact increase the level of confidence and as there is already a significant level of confidence these alternative approaches have not been completed for this study.

The cross-linguistic correlations of two different phonemic systems gave additional problems of comparing from a different set of phonemes. The problem of comparing systems with different numbers of phonemes could be mitigated by considering reciprocal ranks, rather than plain ranks, e.g. rank 10 would be mapped onto 1/10. With this approach all the ranks would lie between 0 and 1 and the amount of difference in the ranks would be reduced. However, this further refinement of the method of assessment has been left for further work.

Having provided the results for the adults and the children as a group at each age in Chapter 7 and having correlated these findings for frequency and FUSE in Chapter 8 the next section, Chapter 9, will similarly present the results for the children as individuals at the three ages. The correlation of these individual child results will also be undertaken using both the adult FUSE and the adult frequency of word initial phonemes.

Chapter 9 : Individual Child Results

In line with usual practice in linguistic research each of the five English and five Finnish child data results are now presented individually over the three ages under assessment. This will enable a non-parametric assessment of individual development patterns to be made, as is the practice with language acquisition case study approaches (e.g. Kunnari 2000 and Savinainen-Makkonen 2000).

Individual frequency rankings and FUSE rankings will be presented for each of the children and for each age individually. The fifteen frequency rank structures for each language (five children and three ages) will then be correlated with the adult FUSE rank structures for both languages. The fifteen FUSE rank structures will likewise be compared with the adult FUSE rank structures for both languages. The average correlation for the children at each age will then be compared with the group child to adult correlation findings given in Chapter 8.

9.1 English Child Results

9.1.1 General Findings

9.1.1.1 Word Types

Each child demonstrated a different set of word types and also a different frequency of word tokens. For example, Jason at age 5 demonstrated the highest number of word tokens with 1083 words uttered and 304 different word types. At age 2, however, he demonstrated the lowest number of word types with a total of only 45 different words out of 104 word tokens.

Table 9.1 provides the numbers of tokens and types for each child at the 3 ages being assessed in this study. The type/token ratio, whilst not of direct use to this research, is of interest in that it demonstrates a fair degree of consistency overall between the total number of words spoken and the number of different words spoken. The word type

and token figures are reported directly as according to the Childes CLAN programming results.

Table 9.1 – English Child Type/Token Ratios

Child	Age	Tokens	Types	Type/Token Ratio
Benjamin	2	345	125	0.38
Benjamin	3	505	166	0.33
Benjamin	5	558	222	0.41
Betty	2	333	120	0.36
Betty	3	469	165	0.35
Betty	5	836	266	0.31
Elspeth	2	117	53	0.45
Elspeth	3	318	136	0.43
Elspeth	5	807	252	0.31
Geoffrey	2	118	54	0.46
Geoffrey	3	342	125	0.37
Geoffrey	5	721	240	0.33
Jason	2	104	45	0.39
Jason	3	259	98	0.38
Jason	5	1083	304	0.28

As discussed in Chapter 6, the adult English orthographic word forms as represented in the full MRC database were used as a look up table to provide the phonemic transcriptions for the words used by each child at each age. Pronunciation variants were added, interjections were deleted and word lists were expanded to include root forms of vowels and singular nouns. This resulted in a list of different word types phonemically represented for each child at each age.

Table 9.2 shows the number of actual word types processed for each child at each age.

The totals presented in table 9.2 include duplicate word types which have been produced by different children and therefore are different than the totals of word types presented in Chapter 7 for the children as a group.

The total of 2360 word types used by the children will be utilised for comparison with the Finnish child number of word type findings.

Table 9.2 – English Child Word Types

	Age 2	**Age 3**	**Age 5**
Benjamin	129	167	231
Betty	101	164	271
Elspeth	53	134	260
Geoffrey	45	121	247
Jason	41	85	310
Total	369	671	1319

A summary of the most frequent words used individually by each child and at each age is given in Appendix 9.1.

As has already been noted in Chapter 7 it is interesting to observe that the words most frequently used tend to be used by more than one child and also match the most frequently spoken adult words.

9.1.1.2 Word Initial Phonemes

In order to ascertain the usefulness of the earlier proposed acquisition theories some general statistics about the phonemes seen in word initial position for the five children individually were gathered.

Appendix 9.2 provides a summary of the word initial phoneme frequency counts for all the children individually at the three ages.

Table 9.3 shows the children's overall frequency (all children's frequency totals combined without analysis of duplicate word types) for the most frequent word initial phonemes together with the number of children, out of the five, that actually utilised the phoneme in word initial position.

Table 9.3 – English Child Word Initial Phoneme Frequency

Phoneme	Frequency	Number of Children Using Phoneme
Age 2		
b	30	5
m	29	5
h	24	5
k	21	5
d	18	5
w	18	4
t	16	5
D	15	5
n	15	5
j	13	5
Age 3		
w	63	5
h	45	5
m	40	5
d	38	5
D	37	5
b	36	5
t	28	5
k	27	5
g	24	5
p	23	5
Age 5		
w	101	5
h	82	5
b	81	5
m	73	5
s	65	5
D	60	5
l	59	5
d	55	5
k	54	5
t	48	5

In summary;

- At age 2 a total of 45 different phonemes were observed in word initial position.

  The phonemes most frequently observed in word initial position were /b, m, h, k and d/ which were used by all the five children.

- At age 3 a total of 55 different phonemes were observed in word initial position.

  The phonemes most frequently observed in word initial position were /w, h, m, d, dʒ, b, t, k, g, p, ɪ, n, s, r, f, l, j /.

- At age 5 a total of 65 different phonemes were observed in word initial position.

  The phonemes most frequently observed in word initial position were /w. h, b, m, s, ð, l, d, k, t, r, p, n, f, g, ɪ, j/.

It is also interesting to observe the range of different word initial phonemes utilised by different children. The largest range of different word initial phonemes was observed for Elspeth and Jason at age 5 where 51 out of the possible 79 different word initial phonemes and phonemic units were utilised.

Table 9.4 demonstrates the range of word initial phonemes utilised by the children at the three ages.

9.1.1.3 Minimal Pair Findings

Each of the five children presents their own groups of minimal pair groupings based upon their own use of the phonemic system and the words they produce. Each child

therefore has their own FUSE count of phonemes involved in these word initial minimal pair groups.

Table 9.4 – Range of English Child Word Initial Phonemes

	Number of Phonemes Used in WI Position	Not Used
Age 2		
Ben	37	42
Betty	39	40
Elspeth	23	56
Geoffrey	28	51
Jason	23	56
Age 3		
Ben	40	39
Betty	40	39
Elspeth	37	42
Geoffrey	32	47
Jason	28	51
Age 5		
Ben	48	31
Betty	46	33
Elspeth	51	28
Geoffrey	48	31
Jason	51	28

A total of fifteen files, with one for each of the five children at each of the three ages has been created and for each child at each age the phonemes have been ranked for FUSE. Each individual child reflects their own pattern of FUSE rankings and these will be compared with the adult FUSE rankings in Section 9.1. below.

Appendix 9.3 provides a summary of these fifteen FUSE totals relating to English child phonemic usage together with a total FUSE count for each phoneme. As discussed in Section 7.3.2.1, the individual children's FUSE totals for a particular phoneme (based upon the range of words in use by a particular child) were added together to provide another way of viewing word initial phoneme contrast usage.

As can be seen in Appendix 9.3 with this approach, Child FUSE count A, the phonemes /ð, ɒ, b, g, h, j, k/ ranked the highest at age 2, the phonemes /ɒ, ɪə, h, ɜ, eə, b, j/ at age 3 and the phonemes /ð, ɒ, ɪə, eə, b, ɜ, eɪ/ at age 5. These results are very different to the group word initial FUSE total rankings presented in Chapter 7.

9.1.2 Correlation of Adult Rankings to Individual English Child Rankings

9.1.2.1 Adult FUSE to Individual Child Frequency

Each of the five children had their individual word initial phoneme frequency rankings compared with the adult FUSE rankings to see whether different children do in fact demonstrate different trends.

In summary;
Ben demonstrated a correlation of .63 at age 2, .71 at age 3 and .69 at age 5.
Betty demonstrated a correlation of .61 at age 2, .64 at age 3 and .68 at age 5.
Elspeth demonstrated a correlation of .58 at age 2, .58 at age 3 and .66 at age 5.
Geoffrey demonstrated a correlation of .61 at age 2, .66 at age 3 and .71 at age 5.
Jason demonstrated a correlation of .52 at age 2, .52 at age 3 and .61 at age 5.

All of the children demonstrated a higher correlation at age 5 than at age 2 perhaps once again indicating a move towards the adult phonemic system as the children's phonological systems developed. The range of movement for Ben is .06, for Betty is .07, for Elspeth is .08, for Geoffrey is .10 and for Jason is .09.

Each child's individual correlation at the three ages enables individual differences in the children's phonemic usage to be observed. Ben's average correlation over the three ages is .67, Betty's is .64, Elspeth is .60, Geoffrey's is .66 and Jason's is .55.

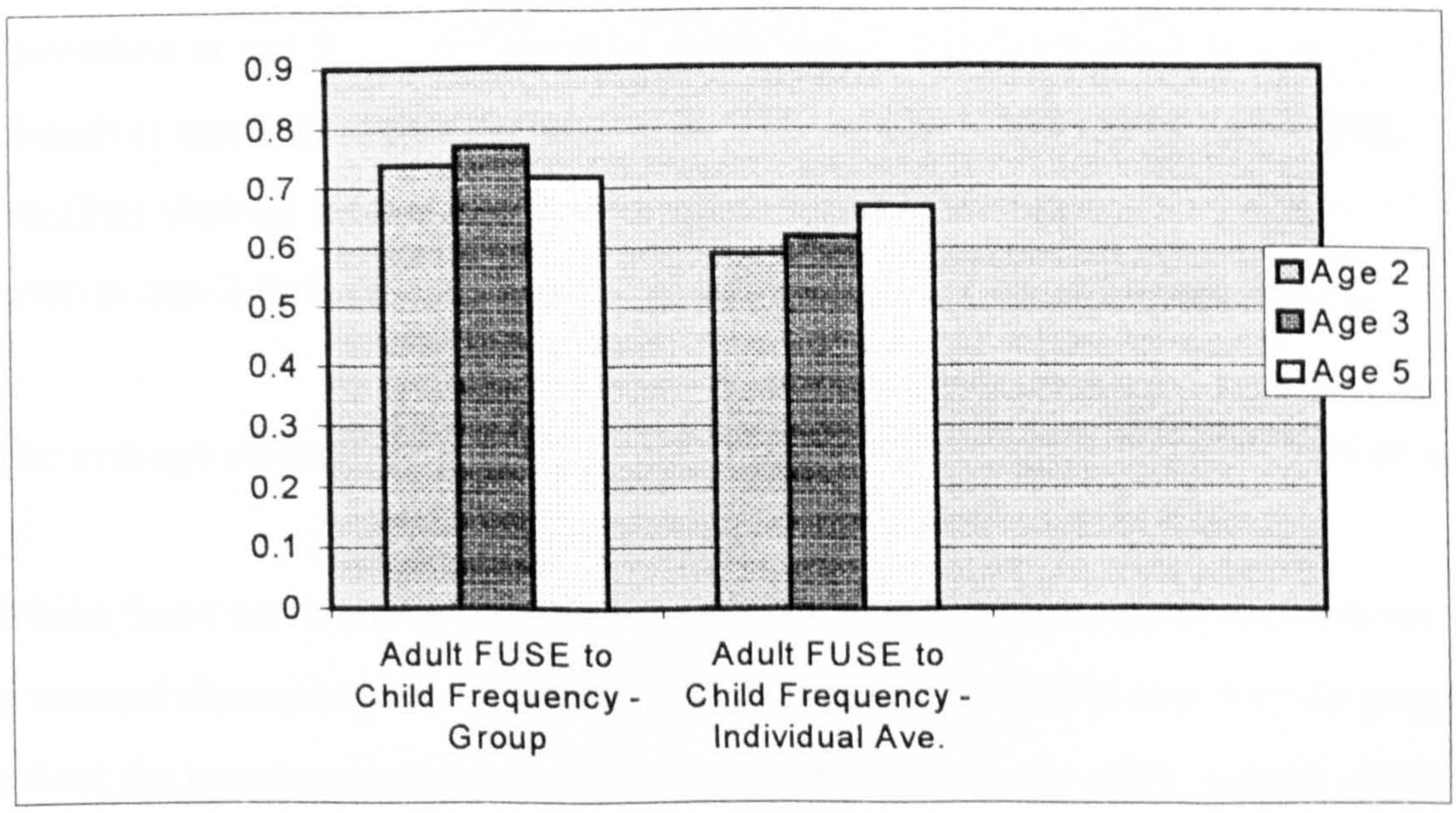
The average correlation for the five children overall at age 2 is .59, at age 3 is .62 and at age 5 is .67.

The group frequency findings (provided in Chapter 7) compared to the adult FUSE rankings show a higher correlation than the average of the individual children's frequencies compared to adult FUSE at all three ages. This is probably due to the fact that utilising the frequency rankings for all children's frequencies combined means that ranks for a greater range of phonemes than that in reality utilised by individual children is given. Also, with the group child approach the frequency of word initial phonemes takes as its starting point word types and therefore duplicate word types would have been deleted from the calculation of word initial phoneme frequency.

The individual child rankings are taken as a more complete reflection of individual children's systems and as can be seen below the individual child correlation findings do appear to indicate a direction of movement towards the adult FUSE that the group findings do not. Interestingly with the children observed as a group the correlation results do not show the direction of movement observed when utilising the average of the individual child findings

Table 9.5 shows the relative correlation findings. The group child frequency to adult FUSE and the average of the individual child frequencies to adult FUSE at the three ages.

Table 9.5 –Adult FUSE to Group and Individual Child Frequency Correlations



9.1.2.2 Adult FUSE to Individual Child FUSE

As shown above each of the five children presents their own FUSE ranking at each age based upon their own use of the phonemic system and words.

All children had their limited contrast phonemes ranked and compared with the 74 adult FUSE rankings. Only one child, Betty, utilised a phoneme that hadn't registered as a contrast with the adult data. This phoneme, a vowel triphthong /aɪə/ was used at age 5 as the contrast in a minimal pair and it meant that for Betty's correlation calculations the number of phonemes being compared was 75 instead of 74(the adult FUSE ranking for this extra phoneme was added as the lowest rank).

Detailed correlation assessments were completed showing the correlation between the English adult FUSE ranking and the child FUSE rankings for each of the children at each age.

In summary;

Ben aged 2 showed a 0.54 correlation, at age 3 a 0.61 correlation and at age 5 a 0.58 correlation.
Betty showed a 0.54 correlation at age 2, a 0.62 correlation at age 3 and a 0.68 correlation at age 5.
Elspeth at age 2 showed s correlation of 0.56, at age 3 0.58 and at age 5 0.66.
Geoffrey showed 0.60 at age 2, 0.65 at age 3 and 0.70 at age 5.
Jason at age 2 showed a correlation of 0.60, at age 3 0.56 and at age 5 0.71.

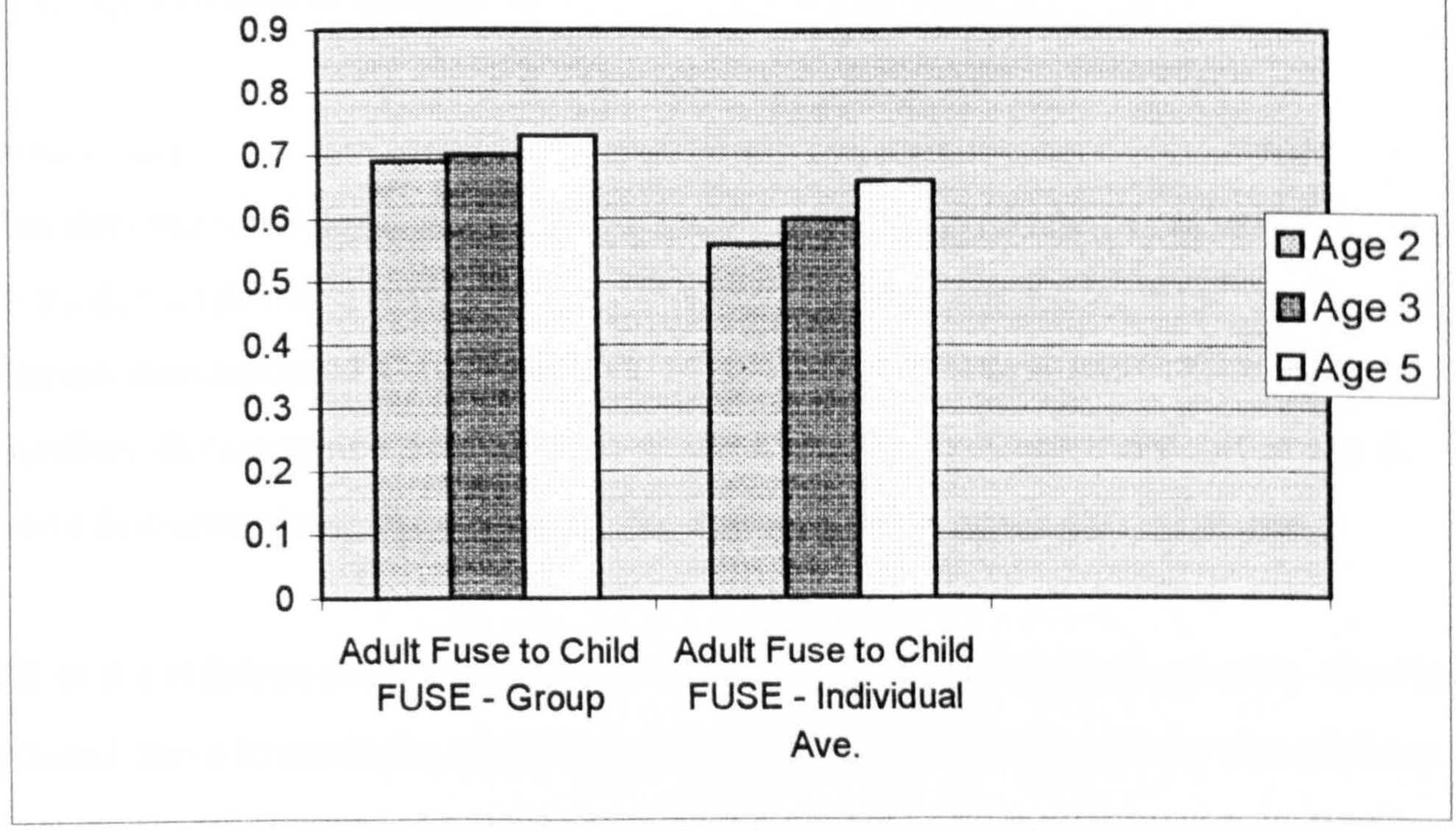
The average correlation at age 2 is 0.56 rising to 0.6 at age 3 and then 0.66 at age 5.

Whilst these correlations seem lower than the group child English correlations presented above they are all within a confidence level of 95% and they do properly reflect the movement of the children's systems towards the adult system. All of the children show more correlation with the adult FUSE rankings at age 5 than at age 2.

The movement for the various children was; .04 for Ben, .14 for Betty, .10 for Elspeth, .10 for Geoffrey and .11 for Jason.

Table 9.6 shows the correlation findings for the group child FUSE to adult FUSE and for the individual child FUSE average and adult FUSE.

Table 9.6–Group and Individual FUSE to Adult FUSE Correlations



The correlation findings for the English adult FUSE based rankings to frequency of phonemes in individual children's speech (see above) were not much closer than these FUSE to FUSE rankings which is surprising considering that more phonemes are actually used by the children than are seen acting as contrasts within minimal pairs. All phonemes utilised by the children are included in the child frequency rankings and this gives a much wider range of phonemes to compare with the 74 phonemes of the adult FUSE rankings. The small word counts for children's minimal pair groupings observed in the children's data reduces the range of phonemes observed as contrasts. For example, at age 2 Ben only uses 18 different phonemes as contrasts, rising to just 31 even at age 5. With this approach the majority of phonemes ranked for FUSE for the adult data do not even appear in the child FUSE findings and despite them having the lowest ranking this high majority does seem to affect the correlation results.

Once again, as with the findings in Section 9.1.2.3, utilising the average of the individual FUSE findings seems to more clearly show a direction of movement by the children towards the adult FUSE rankings despite the actual correlation being lower over the three ages.

9.1.2.3 Adult to Individual Child Word Initial Frequency

The individual child frequency rankings were also correlated with the adult frequency rankings provided in Chapter 7.

In summary;

Ben demonstrated a correlation of .71 at age 2, .78 at age 3 and .83 at age 5.

Betty demonstrated a correlation of .73 at age 2, .74 at age 3 and .80 at age 5.

Elspeth demonstrated a correlation of .66 at age 2, .74 at age 3 and .80 at age 5.

Geoffrey demonstrated a correlation of .70 at age 2, .73 at age 3 and .77 at age 5.

Jason demonstrated a correlation of .57 at age 2, .59 at age 3 and .71 at age 5.

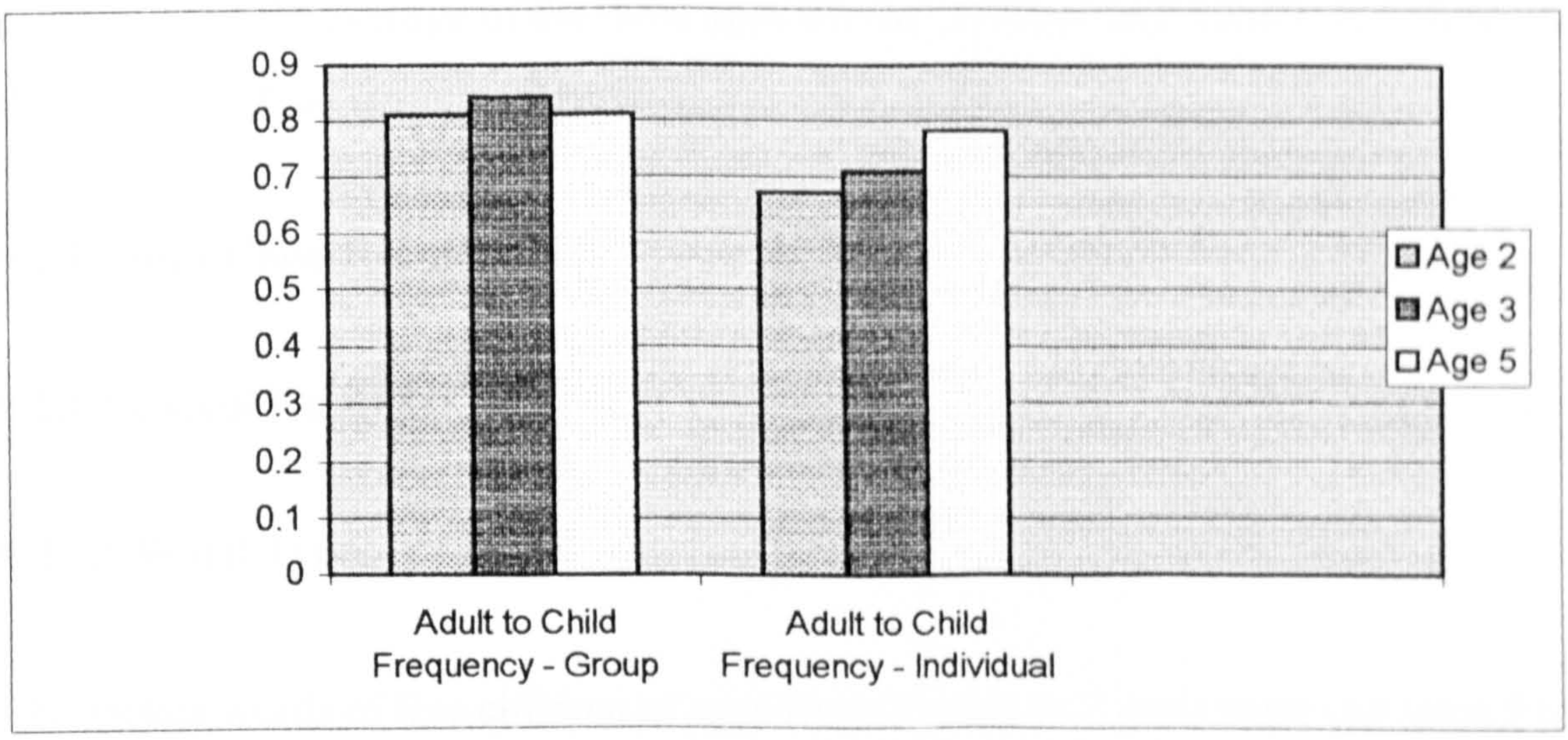
All of the children show a higher correlation at age 5 than at age 2 possibly showing a gradual move towards the adult frequency rankings. The range of movement from age 2 to age 5 varied, as can be seen, from .07 for Betty and Geoffrey up to.14 for Elspeth and Jason.

To give an indication of how particular children's frequency rank correlations differed each child has an average correlation calculated. Ben's average correlation over the three ages is .77, Betty's is .75, Elspeth is .73, Geoffrey's is .73 and Jason's is .62.

The average correlation for the five children overall(taking all correlations at each age) at age 2 is .67, at age 3 is .71 and at age 5 is .78. Whilst these findings based upon individual children's performance show a slightly lower correlation than for overall child frequency given above (0.81 at age 2, 0.84 at age 3 and 0.81 at age 5) they are perhaps a better indication of actual frequency and usage.

Table 9.7 shows the group and individual child frequency to adult frequency findings.

Table 9.7 –Adult Frequency to Group and Individual Child Frequency Correlations

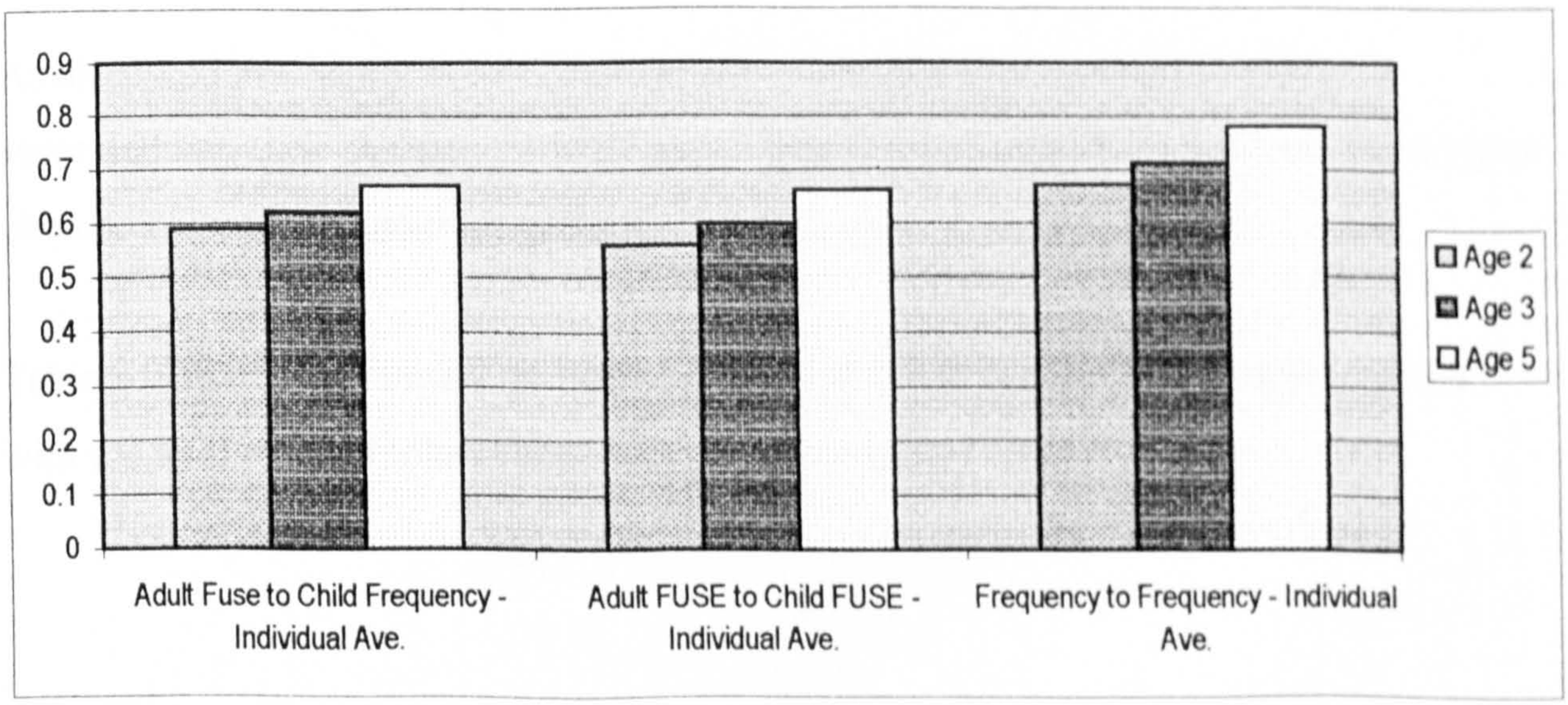


It is interesting to observe that once again utilising the individual child frequency ranking compared with the adult frequency ranking seems to indicate a direction of movement towards the adult language which the group findings do not do despite the correlations starting higher and remaining closer over the three ages.

9.1.3 Comparison of English Findings

The correlation findings for the three approaches applied to the English data can be see more clearly in table 9.8.

Table 9.8 – English Findings for Individual Children



All three methods of assessment, frequency and FUSE based, not only correlate closely across the average of the three ages for the children individually but all seem be showing an overall movement towards the adult system

9.2 Finnish Child Results

9.2.1 General Findings

9.2.1.1 Word Types

The spoken words of five children at ages from 2 years to 7 years were extracted from the Oulu corpora. The children varied in their use of language in term of the word forms and phonemes utilised. They also varied in their volume of words spoken at each age, as would be expected. As this study will only be comparing a child's phonemic usage within their own productive system of use (i.e. only compared to their own full output lexicon) this will not pose a problem.

Whilst the timing of the speech samples was kept constant by the original transcribers the amount of speech recorded, in terms of both tokens and types of words, varied for each child and at each age. For example, Virpi demonstrated the highest number of word types of all the children at age 2 with a total of 229 different word types out of the 543 word tokens, Teppo demonstrated the highest number at age 3 (241 types) and Riikka at age 5 (328 types).

As discussed in Chapter 6, interjections were deleted and long word forms were separated into their component word parts. This resulted in a list of different word types phonemically represented for each child at each age.

Table 9.9 shows the number of word types found for each child at each age together with the total of word tokens and the type/token ratio

Table 9.9 – Type/Token Ratio for Finnish Child Data

Child	Age	Tokens	Types	Type/Token Ratio
Harri	2	228	96	0.42
Harri	3	199	94	0.47
Harri	5	262	131	0.5
Riikka	2	201	109	0.54
Riikka	3	244	121	0.5
Riikka	5	638	328	0.51
Sami	2	555	138	0.25
Sami	3	463	176	0.38
Sami	5	464	229	0.49
Teppo	2	362	182	0.5
Teppo	3	643	251	0.39
Teppo	4	538	292	0.54
Virpi	2	543	229	0.42
Virpi	3	197	118	0.6
Virpi	5	122	79	0.64
		5659	2573	

The totals above include duplicate word types which have been produced by different children and therefore are different than the totals of word types presented in Chapter 7. The total of 2573 word types used by the children compares with the 2360 word types produces by the English children.

It is interesting to compare the words most frequently spoken by the children with the words most frequently spoken by adult Finnish speakers and represented in the larger adult Finnish corpora and word listings. It is also interesting to compare and contrast the phonemic contents of the words uttered by Finnish as compared to English children.

A summary of the most frequent words used by each child and at each of the three ages under analysis is given in Appendix 9.4

9.2.1.2 Word Initial Phonemes Frequency

In order to ascertain the usefulness of the earlier proposed theories, some general statistics about the phonemes seen in word initial position for the five children under analysis were gathered. In order to compare the adult FUSE rankings with the frequency findings each individual child's frequency rankings at each age will need to be assessed.

Each of the five children presents their frequency ranking based upon their own use of the phonemic system and the words they produce. Each child will therefore have their own phoneme frequency rankings for each age of assessment. Appendix 9.5 provides the word initial phoneme frequencies for the five Finnish children at the three ages.

A frequency count of the most frequent phonemes seen at word initial position for each word type, counting each type only once (without its actual frequency count) and summing all the child frequencies together is given in Table 9.10. This combined frequency count is different than that used for the group child frequency findings where duplicate words utilised by more than one child were deleted before the word initial phonemes were counted.

In summary;

At age 2 the phonemes most frequently observed in word initial position were /t, p, k, m and s/ which were used by all the children.

At age 3 the phonemes most frequently observed in word initial position were /t, k, m, p, s, n, o, v, l, h, j, e, r, ɑ, i, y, au /.

At age 5 the phonemes the most frequently observed in word initial position were /t, k, m, s, p, o, l, v, n, j, j, ɑ, i, e, r, au, ei/.

Table 9.10 – Finnish Child Word Initial Phoneme Frequency

Frequency	Phoneme	No. Children
Age 2		
114	t	5
106	p	5
92	k	5
68	m	5
64	s	5
43	o	5
41	n	5
40	l	5
28	v	5
27	a	5
27	j	5
23	I	5
Age 3		
131	t	5
106	k	5
76	m	5
72	p	5
65	s	5
46	n	5
41	o	5
32	v	4
25	l	5
22	h	5
22	j	5
21	e	5
Age 5		
144	t	5
133	k	5
107	m	5
99	s	5
91	p	5
63	o	5
62	l	5
60	v	5
49	n	5
40	j	5
35	h	5
30	a	5

9.2.1.3 Minimal Pair Findings

Each of the five children presents their own groups of minimal pair groupings based upon their own use of the phonemic system and the words they produce. Each child therefore has their own FUSE count of phonemes involved in these word initial minimal pair groups. A total of fifteen files, with one for each of the five children at each of the three ages has been created and for each child at each age the phonemes have been ranked for FUSE. Each individual child reflects their own pattern of FUSE rankings and these will be compared with the adult FUSE rankings in Section 9.2. below.

Appendix 9.6 provides a summary of these fifteen FUSE totals relating to Finnish child phonemic usage together with a total FUSE count for each phoneme.

9.2.2 Correlation of Adult Rankings to Individual Finnish Child Rankings

9.2.2.1 Adult FUSE to Individual Child Frequency

Each child's individual frequency was compared with the adult FUSE ranking to see whether different children demonstrate different trends..

In summary;

Harri demonstrated a correlation of .74 at age 2, .74 at age 3 and .81 at age 5.
Riikka demonstrated a correlation of .80 at age 2, .75 at age 3 and .66 at age 5.
Sami demonstrated a correlation of .78 at age 2, .77 at age 3 and .76 at age 5.
Teppo demonstrated a correlation of .75 at age 2, .58 at age 3 and .75 at age 5.
Virpi demonstrated a correlation of .78 at age 2, .76 at age 3 and .76 at age 5.

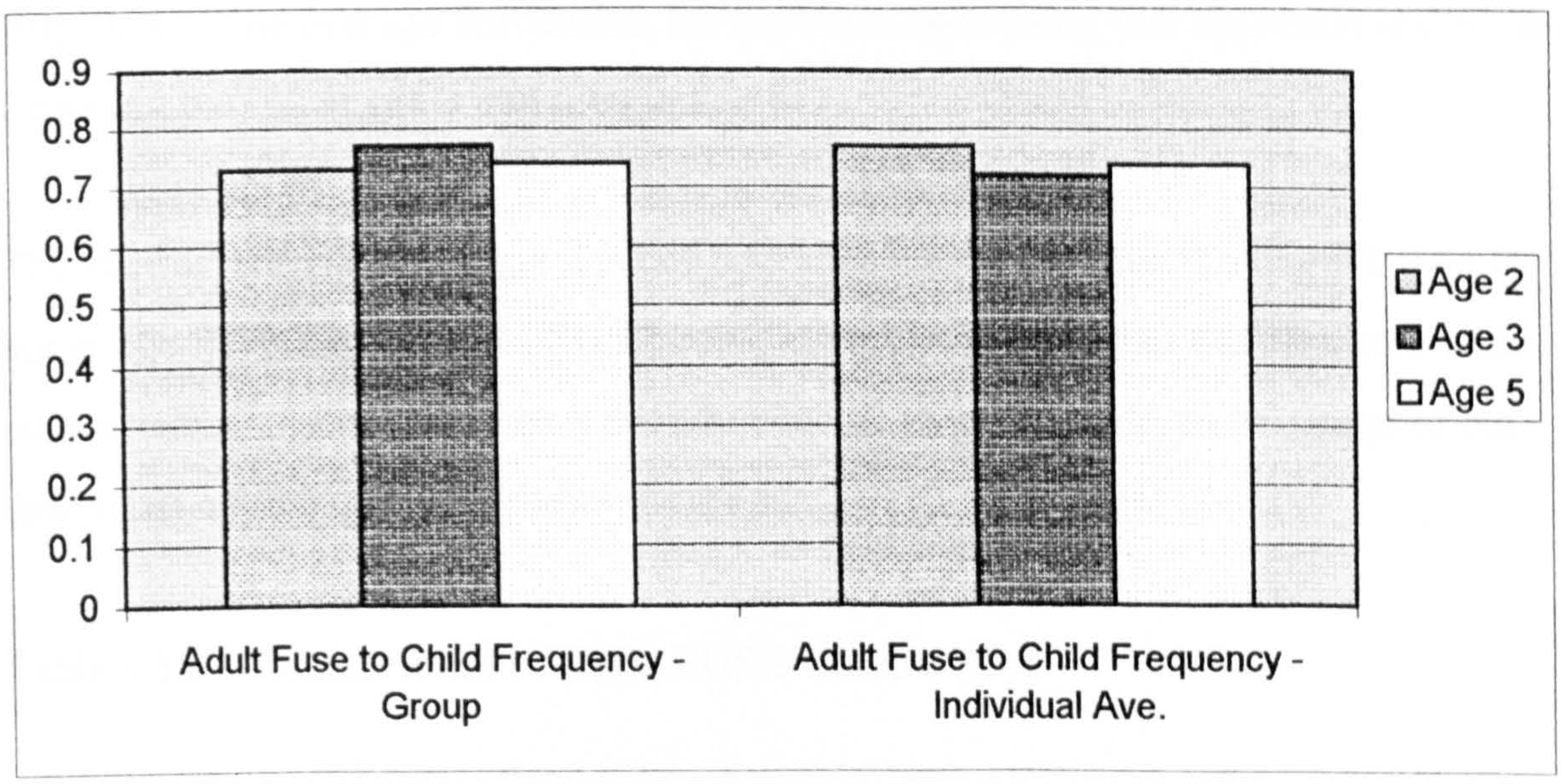
Only one of the children, Harri, demonstrated a higher correlation at age 5 than at age 2. The other children's word initial frequency findings seem to suggest a move away from the FUSE rankings as the children get older.

The range of movement from age 2 to age 5 is .07 for Harri, .-14 for Riikka, -.02, for Sami, .0 for Teppo, and -.02 for Virpi. Harri's average correlation over the three ages is .76, Riikka's is .73, Sami's is .77, Teppo's is .69 and Virpi's is .76.

The average correlation for the five children overall at age 2 is .77, at age 3 is .72 and at age 5 is .74.

Table 9.11, below, more clearly shows the individual and group Finnish child word initial phoneme frequency to adult FUSE frequency correlations.

Table 9.11 Finnish Adult FUSE to Group and Individual Child Frequency Correlations



9.2.2.2 Adult FUSE to Individual Child FUSE

Each of the five children presents their own FUSE ranking at each age based upon their own use of the phonemic system and words. Correlation between the Finnish adult FUSE ranking and the child FUSE rankings for each of the children at each age can be summarised as follows;

Harri moved from 0.51 at age 2 to 0.60 at age 3 and then back to 0.50 at age 5.

Riikka aged 2 showed a 0.62 correlation, at age 3 a 0.64 correlation and at age 5 a 0.68 correlation.

Sami showed a 0.63 correlation at age 2, 0.62 at age 3 and 0.66 at age 5.

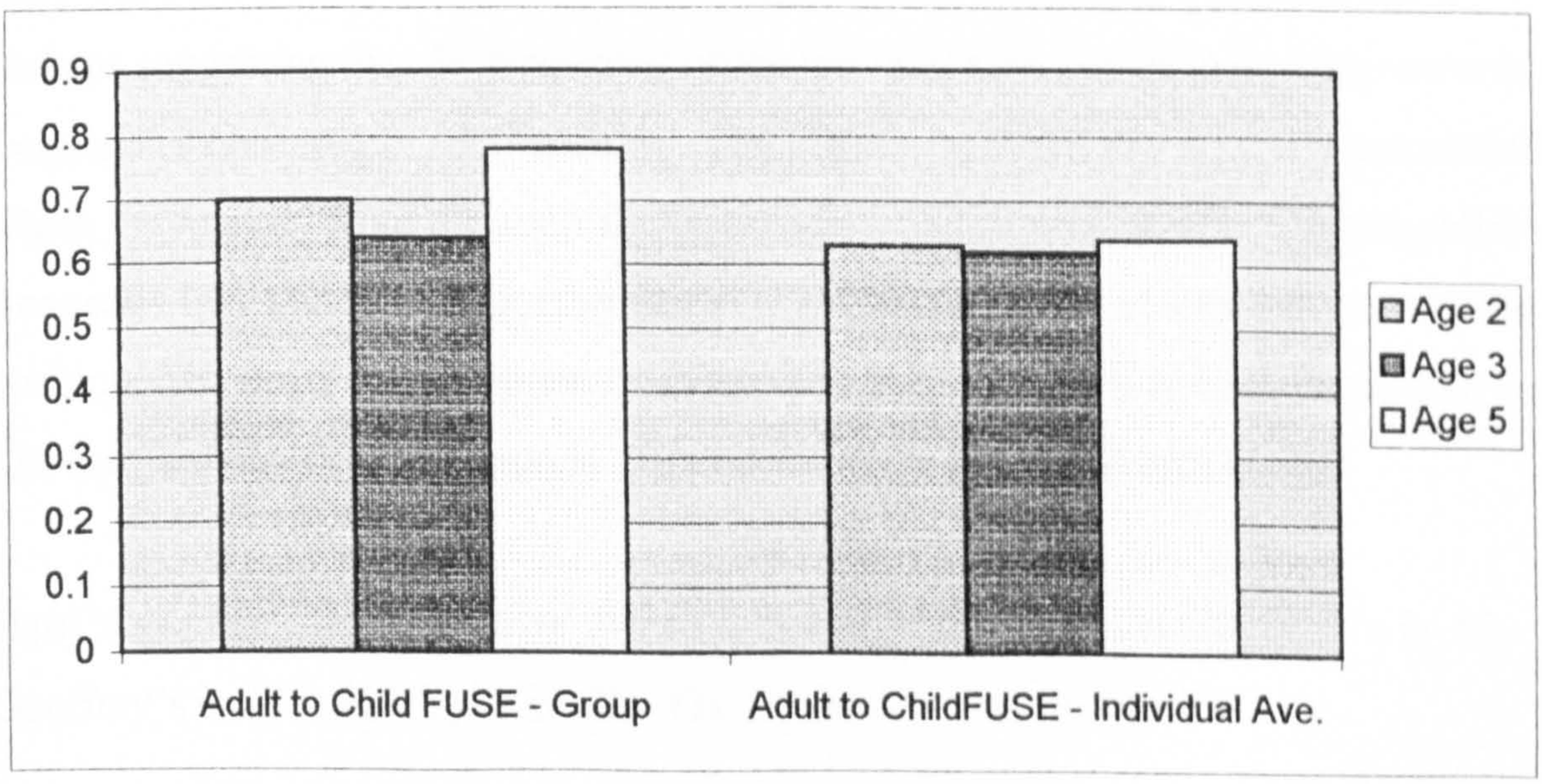
Teppo showed a 0.71 correlation at age 2, a 0.63 correlation at age 3 and a 0.76 correlation at age 5.
Virpi showed a 0.72 correlation at age 2, a 0.64 correlation at age 3 and 0.64 correlation at age 5.

Three of the children showed more correlation at age 5 than at age 2 perhaps indicating that their path's of development were towards the adult system. The ranges here were .06 for Riikka, .03 for Sami and .05 for Teppo. Harri and Virpi seemed to show a move away from the adult language FUSE as they developed (-.01 and -.08 respectively).

Harri's average correlation was .53, Riikka's was .64, Sami's .63, Teppo's .7 and Virpi's .66. The average correlation for the three ages using this approach is 0.63 at age 2 moving to 0.62 at age 3 and 0.64 at age 5.

The starting correlation at age 2 is considerably higher than for the English data (0.56) but at age 5 the English child FUSE correlate more closely than the Finnish FUSE comparisons. Table 9.12 shows the adult FUSE to group child and average of the individual child FUSE correlation findings.

Table 9.12 – Finnish Adult to Child FUSE Comparisons



9.2.2.3 Adult to Individual Child Word Initial Frequency

As with the English findings the individual frequency rankings for the different children were also compared with the overall adult frequency rankings in order to assess whether certain children individually demonstrate stronger or weaker relationships. It was thought that the above approach might suggest more similarity as the children developed and increased their overall phonemic usage.

The individual child frequency rankings over the 3 ages were correlated with the adult frequency rankings.

In summary;

Harri demonstrated a correlation of .75 at age 2, .73 at age 3 and .82 at age 5.
Riikka demonstrated a correlation of .82 at age 2, .77 at age 3 and .76 at age 5.
Sami demonstrated a correlation of .71 at age 2, .67 at age 3 and .66 at age 5.
Teppo demonstrated a correlation of .83 at age 2, .67 at age 3 and .82 at age 5.
Virpi demonstrated a correlation of .81 at age 2, .81 at age 3 and .81 at age 5.

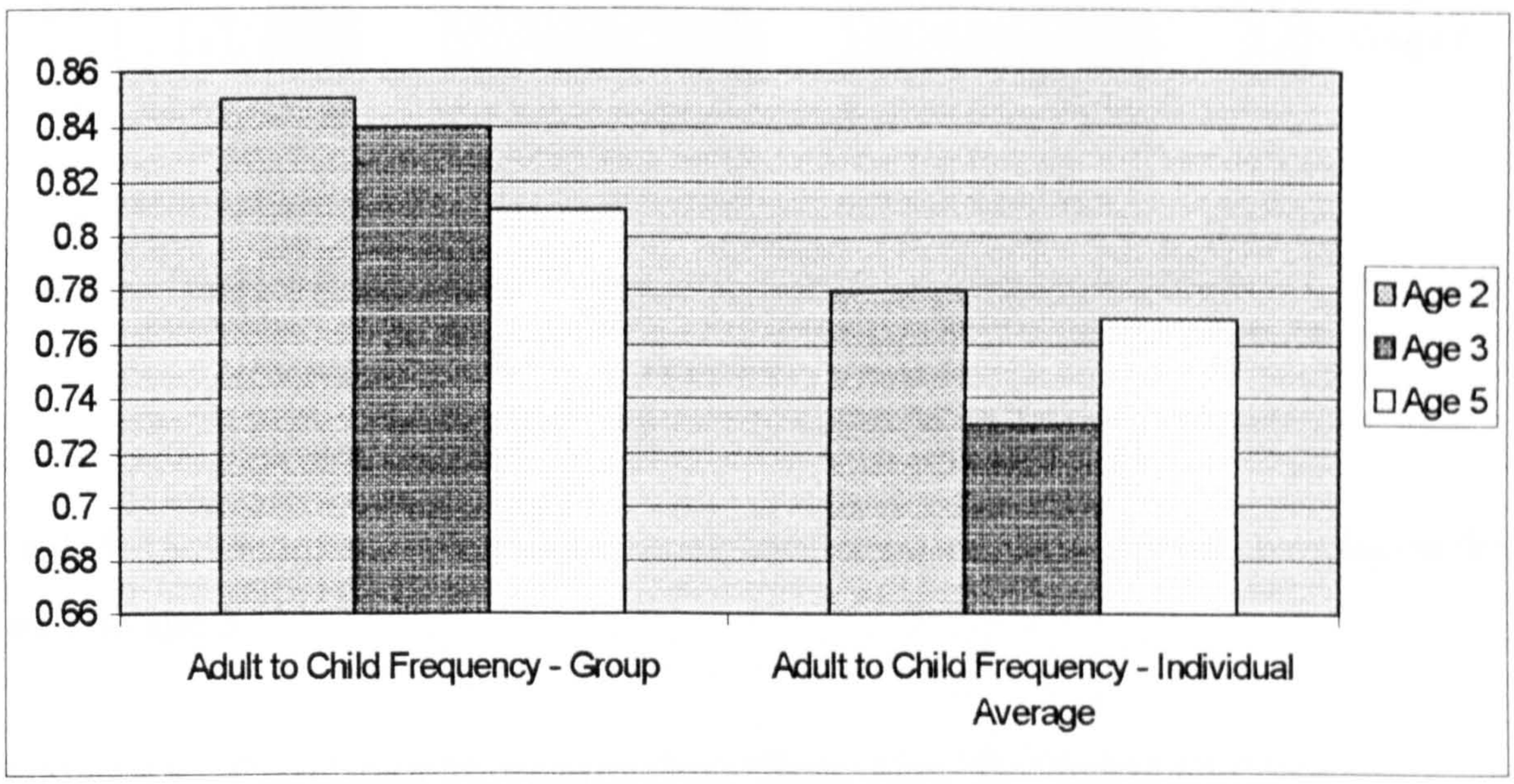
Only one of the children, Harri, shows a higher correlation at age 5 than at age 2 perhaps suggesting that the frequency of the use of his phonemes is moving towards the frequencies observed in adult speech. The remaining 4 children all show correlations which move away from the adult findings perhaps suggesting that for Finnish children frequency is not such an important factor in phonemic usage, at least for word initial position. The range of movement from age 2 to age 5 for Harri is +.07, for Riikka is -.06, for Sami is -.05, for Teppo is -.01 and for Virpi is .0 or no movement.

Harri's average correlation over the three ages is .76, Riikka's is .78, Sami's is .68, Geoffrey's is .77 and Virpi's is .81. The average of these being .76.

The average correlation for the five children at age 2 is .78, at age 3 is .73 and at age 5 is .77.

Table 9.13 shows the correlation of both the group and individual child frequency findings when compared with the adult frequency findings.

Table 9.13 – Finnish Adult to Child Frequency Correlations



As can be seen, frequency does not seem to be a useful indicator of the movement of acquisition for either the group combined Finnish child data nor for the individual child findings.

9.2.3 Comparison of Finnish Findings

The correlation findings for the three approaches applied to the Finnish data can be see more clearly in table 9.14.

Whilst the adult FUSE to child FUSE correlations are seen to be slightly lower than two other types of assessment they are still within a confidence level of more than 90%. They seem to offer one advantage in showing an overall movement towards the adult system. Both the other methods show erratic findings with the final correlation at age 5 actually showing less closeness than the original system at age 2.

Table 9.14 – Finnish Findings

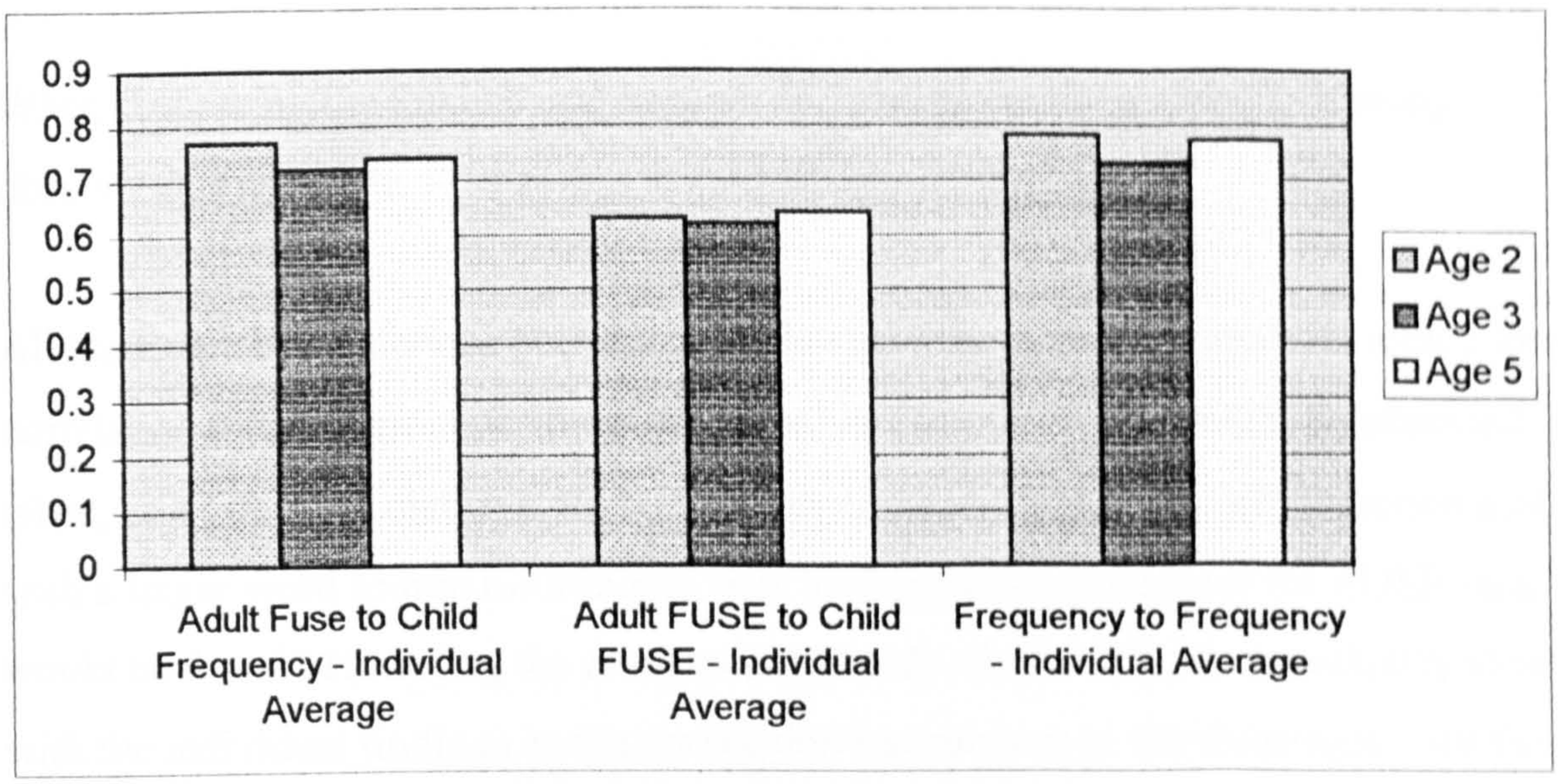
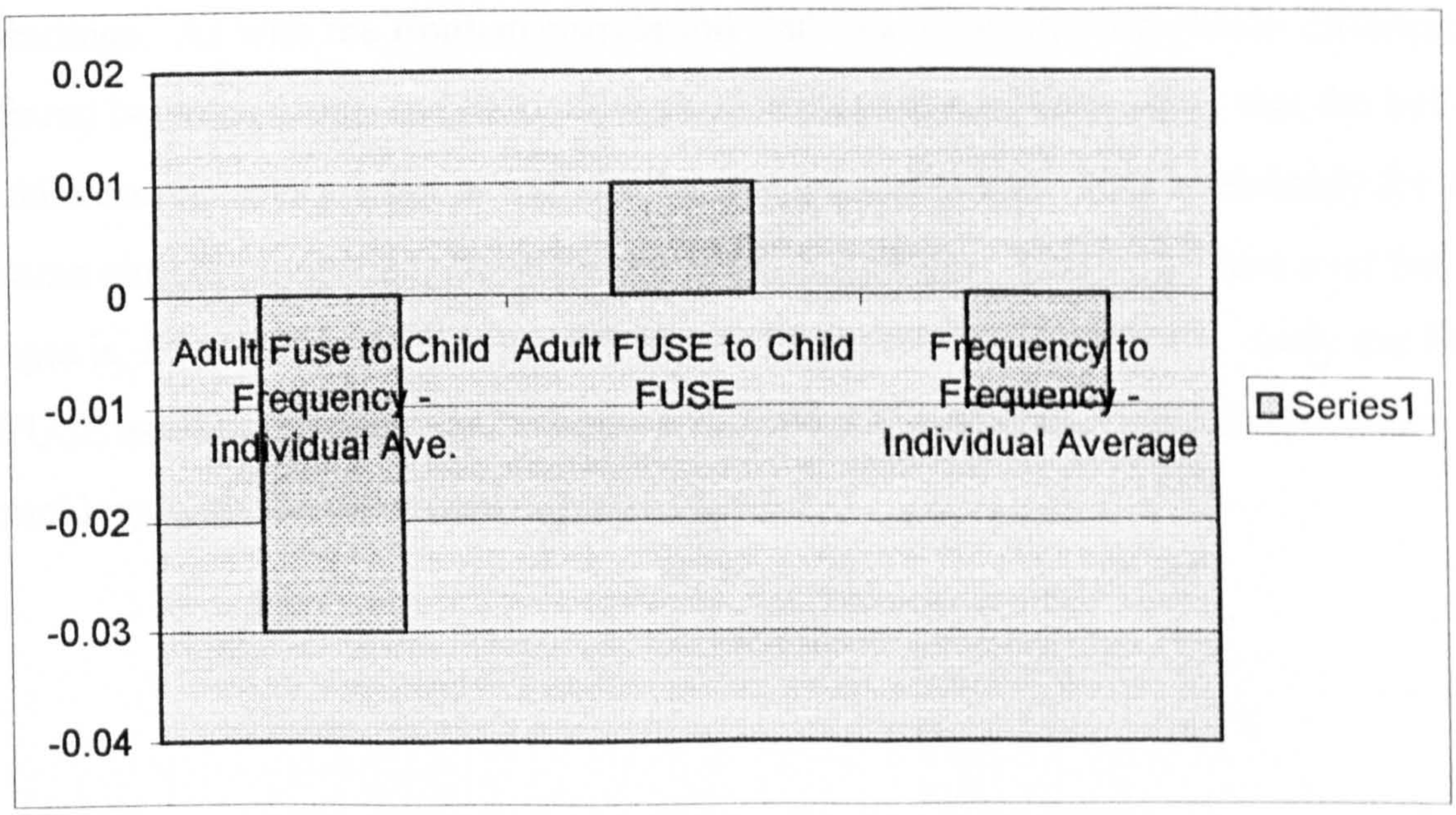


Table 9.15 demonstrates this more clearly and gives the movement of correlation from age 2 to age 5.

Table 9.15 – Correlation Movement Over Three Ages for Finnish Children



For Finnish the best method of measurement would therefore appear to be the adult FUSE to child FUSE as it does show a positive movement towards the adult data. It is recognised that this movement of only .1 is however only small and further analysis would need to be done on more and larger data samples to be sure that these results were not simply an indication of the very small data source.

9.3 Summary of Correlation Findings

When comparing the Finnish and English correlation findings some interesting observations are made.

All the English correlations presented above show the individual analysis with a lower correlation than the group findings presented in Chapter 8. This might be expected when, as a group the children are more likely to use a wider range of phonemes and with a larger word source for minimal pair findings more phonemes for FUSE rankings would be found. However, the movement towards adult FUSE is more clearly shown with the individual findings as the correlation increases over the three ages. All three methods of assessment show this movement with the English child individual data whereas only the FUSE to FUSE method shows the movement for the English child group findings.

The Finnish results presented here however suggest a different pattern from the English findings. As with the English correlation findings the actual correlation differences found between group and individual child Finnish correlations show that the individual child correlations are not as high as the group correlations. This is probably for the same reasons suggested above for the English findings. The movement over the three ages is, however, less obvious with the children treated individually. Only the FUSE to FUSE assessment once again shows a direction of movement towards the adult rankings with the individual correlations.

Chapter 10 - Conclusions

This chapter summarises the main areas of work completed for this study and summarises the results presented in Chapters 7 to 9. The main purpose of this chapter is to draw some conclusions from the results and to suggest ways forward for future research.

10.1 Summary of Findings

This study set out to explore whether a relationship between ambient language and phonological acquisition could be detected using an empirical corpus based approach and a method of assessment, developed for this study, known as FUSE.

The phonemic systems of the two historically unrelated languages of English and Finnish were assessed and phonemic usage data was compiled using adult spoken language corpora for these languages as a representation of ambient language. Child data samples for five children over three ages of development were likewise assessed for phonemic usage so that the similarity in usage between adult and child samples could then be compared. This assessment provided much new data about the usage of words, the word and syllable structures used by adult and child speakers of these languages, as well as information about phonemic usage.

One objective of this study was to develop a method for assessing the relationship between phonological acquisition and ambient language. The method of assessment, known as FUSE, was applied to the adult and child spoken language data samples of Finnish and English such that correlation between the adult data samples and children's developing phonological systems could be viewed in a new way. The method of approach adopted for this initial application had to take into account the phonemic inventories and phonotactic rules of these two languages and aimed additionally to incorporate a measure of speaker usage. This was achieved by using the functional load approach (as a basis for a systemic approach to the analysis) and then utilising only the most frequently spoken words from corpora of the two languages as a measure of speaker usage. The FUSE method was designed in such a way that it could, in the

future, be applied to any language (to provide consistency in cross-linguistic studies) and to both adult and child languages (so that the effect of the ambient language could be explored in terms of the child's developing phonemic system).

The process for achieving this measure has resulted in a greater knowledge of English and Finnish word types, new statistics on the frequency of phonemes (both overall and by particular word position) and word structures and information about the use of contrastive phonemes for Finnish and English has been gained. Also, more is now known about the differences in the typical structures of the two languages (i.e. that Finnish words tend to be longer and with more syllables). Interestingly, with the words translated, the same most frequently spoken words appeared in both the English and Finnish adult data files however the English and Finnish children seemed to adopt a different approach to utilising these most frequent words with the Finnish children demonstrating more individual preferences.

The frequency information, based upon word types rather than the more traditional word token approach, has provided a new set of frequency figures and a different basis from which to compare the child data.

Despite this difference the ten most frequent phonemes over all word positions for English /ɪ, t, n, ə, s, l, k, d, r, z/ matched very closely with previous phoneme frequency findings where word tokens had been assessed. Differences with other studies were accounted for by a number of factors such as the non-recognition of consonant clusters as phonemic units and other phonemic system differences. Out of a total of 97 possible phonemes and phonemic units identified for this study only 87 were observed being utilised in the most frequent word types of the spoken data corpora used for this study. Six word initial two consonant cluster sequences /ʃr, gw, θw, dw, sj, sf/, three word initial three consonant cluster sequences /spj, skl, skj/ and the vowel triphthong /eɪə/ did not occur at all in the word forms selected for processing. The phonemes /z/and /ɪ/ were found to be used the most extensively across word positions.

A total of seventy nine different phonemes/phonemic units were observed in word initial position (out of an expected 85) and the ten most frequently observed WI phonemes were found to be /k, s, ɪ, r, d, f, m, p, ə, b/. Word structures ranging from one to fifteen phonemes in length were found with the most frequent length of word type being words of four phonemes.

A total of 313 different word structures were identified with the most frequent word structure being CVC. A total of 1621 word types were involved in groups of minimal pairs where the contrastive phoneme occurred at word initial position and the phonemes ranked highest for FUSE were /s, l, f,w, t, b, m, k, r, h/.

For a direct comparison between two different size word files, the processing was also completed for a larger English word file. This file of words extracted again from the MRC Psycholinguistic Database contained an additional 4685 phonemic transcriptions for hapax legomena word and as would be expected from the larger word basis there were more minimal pairs observed and the overall frequencies of the individual phonemes involved in minimal pair strings were therefore higher. Despite the larger range of words from which to calculate FUSE the FUSE findings were found to be similar and it is interesting to observe that regardless of the size of the dataset the same phonemes consistently appear within the top ten FUSE rankings confirming their important role within the English language.

The findings for the most frequent phonemes observed in the Finnish adult data /t, i, ɑ, s, l, e, n, k, o, m/ also concur closely with the findings of other Finnish language researchers (e.g. Pääkkönen 1973, Vainio1996 and Pajunen & Palomäki 1984).

A total of 61 (out of a possible 62) different phonemes/phonemic units were observed in word initial position and the ten most frequently observed WI phonemes for Finnish were /k, t, s, m, p, v, l, n, h, j/. Words of one to 21 phonemes were found with the most frequent word length being words of six phonemes. The interesting thing here is that all the most frequent word lengths are longer than those found in the English adult

data. The most frequent word structure was 'CVCCVC' and the phonemes /ɑ/ and /t/ were found to be the most extensively used phonemes across all word positions.

A total of 1278 Finnish words were involved in groups of minimal pairs where the contrast phoneme occurred at word initial position. This compares with 1610 words for English. Despite the adult English word file containing less actual word types than the Finnish word file there were more minimal pairs where the initial phoneme provided the contrast found in the adult English data. This may be indicating something about the different nature of minimal pairs in English and Finnish (i.e. that English tends to have more word initial contrasts whereas perhaps Finnish utilises another word position such as word final position).

Words of up to 10 phonemes in length were involved in the Finnish adult minimal pair groupings and a total of 47 different phonemes were found to be involved as the word initial contrast in minimal pairs. The highest ranked Finnish phonemes according to the FUSE assessment were /s, m, t, k, n, v, j, p, h, l/. Vowel phonemes ranked considerably higher than with English perhaps an indication of more open first syllable usage (i.e. nucleus without a preceding consonant).

The spoken words of the five English speaking children at the three ages under investigation were analysed both as a group and individually. The most frequent words for the children as a group at each age matched closely to the most frequently spoken adult words. Interestingly, at age 2 all of the most frequently spoken consonant phonemes over all word positions were also included in the top ten most frequently used adult phonemes perhaps suggesting a closeness in the frequency of phoneme usage between the children and the adult language even at this young age. However, at age 2 with word position taken into account the group results show a very different top ten ranking to the adult word initial top ten phoneme frequency findings. A range of only 45 different phonemes (compared to 79 in the adult data) were observed in the English child data at word initial position at age 2, a range of 54 different phonemes were observed at age 3 and this increased to 64 at age 5, indicating a development of phoneme usage. Word structures ranging from one to 7 phonemes in length at age 2,

from one to 7 at age 3 and from one to 9 phonemes at age 5 were found perhaps indicating that as the children developed the length of the words they tended to utilise got slightly longer. The most frequent word type length was 3 phonemes at ages 2 and 3 and 4 phonemes at age 5 indicating again that as they developed the children more frequently utilised longer forms. At all three ages the most frequently used word structure was thus found to be CVC. This matches the most frequent adult English word structure observed in the data.

Unlike the English child data where the children were found to demonstrate the same most frequent word types the Finnish children demonstrated individual preferences for certain words at each stage of development and also demonstrated word usage that was very different than the adult data. The children used a range of 36 different phonemes at age 2 increasing to 43 (compared to an adult total of 62) different phonemes at age 5. Only one phoneme /u/ that occurred in the top ten child at age 2 usage does not occur in the adult top ten phoneme usage ranking perhaps suggesting that already the frequency of usage of phonemes by the children, even at age 2, is close to that of adult usage. Only the phoneme /o/, which occurs in the top ten word initial phoneme frequency rankings for all three ages under assessment, does not occur at all in the adult word initial findings.

With all the children's word initial phoneme frequencies combined the range of word initial phonemes used by the children again, as with the English child data findings, increased over the three ages from 28 different word initial phonemes at age 2 to 33 at age 5 (compared to 61 different word initial adult phonemes). With the Finnish child data word structures ranging from one to 14 phonemes in length at ages 2 and 3 and from one to 17 phonemes in length at age 5 were identified and the most frequent word type length was 6 phonemes for all the ages. Across all the ages the most frequent word structure was found to be CVCCV, in all instances a two syllable word structure. As with the Finnish adult data there is more use of multi-syllable word structures than with the English data. A range of 15 phonemes were seen as the word initial contrast at ages 2 and 3 and 16 at age 4. The phonemes ranked the highest for the Finnish children age 2 show an identical match to the top nine Finnish adult FUSE ranked phonemes.

The FUSE assessment has enabled an observation of the phonemes that are involved in the most minimal pair groups and therefore provide the most amount of contrastive ability. An examination of whether these highly contrastive phonemes, which might be predicted to be acquired the earliest by children acquiring the language, are in fact used the most frequently by children has indicated that children do indeed tend to use these phonemes more frequently.

Interestingly, the resultant FUSE rankings are different to both the overall frequency and specific word position frequency rankings. For example, using frequency of occurrence in ambient language alone without any regard to word position the phonemes that are ranked the highest (and might be predicted to be acquired and utilised the earliest) are /ɪ, t, n, ə, s, l, k, d, r, z/. Using word position and frequency as an indicator then a word initial frequency ranking of /k, s, ɪ, r, d, f, m, p, ə, b/ is predicted for English whereas the FUSE rankings suggest that it would be important for the English child to acquire the phonemes /s, l, f,w, t, b, m, k, r, h/ first as these phonemes provide the most lexical differentiation in the word initial position.

Similarly for Finnish, the phonemes /t, i, ɑ, s, l, e, n, k, o, m/ which are used the most frequently over all word positions might be expected earliest in child data.
The Finnish word initial phoneme frequency findings show the phonemes /k, t, s, m, p, v, l, n, h, j/ being utilised the most frequently in word initial position whilst the phonemes /s, m, t, k, n, v, j, p, h and l/ are ranked the highest for FUSE.

With regard to the relationship between the ambient language (as reflected in the adult data rankings) and the child data findings, the correlations all indicate significant closeness between same language adult and child data samples both in terms of frequency and FUSE. These findings are consistent across all three ages which would tend to suggest that already at age 2 the children are utilising phonemes that more closely represent their language of acquisition thus providing support for the role of language specific features in the phonological acquisition process. The high FUSE correlations observed for both languages may be indicating that phonemic contrasts do have a role to play in the acquisition of both Finnish and English phonological systems.

They perhaps indicate a universal approach to acquisition based upon the individual language's usage of contrasts rather than frequency of use.

The cross-linguistic assessments, on the other hand, show that there is little correlation between the two languages either for the adult or child data samples. Neither the FUSE nor frequency based assessment could detect a significant level of universality between the child data samples even at the earlier ages of assessment. However, the child data does appear to correlate slightly more closely cross-linguistically at the earlier ages of assessment. This may be indicating a more universal basis at an even earlier age than studied here from which the children move towards the language specific features of their language of acquisition. Similarly, the adult to child opposite language correlations seemed to indicate a movement towards the language of acquisition. Further research is required in order to investigate this more thoroughly.

The actual movement towards adult language over the three child stages of development, although only slight, is better shown with the FUSE method of assessment for both languages than with the frequency based assessments. The typical frequency based assessments do not seem to offer much indication of development, particularly for Finnish, as the correlations, which would be expected to increase as the children develop their phonological usage and lexical base, in fact show lesser relationship with the adult data the older the children are. Again, this is an area for future research.

10.2 Future Work

Having applied the method to two small adult data samples representing two unrelated languages the next stage would be to apply the measure similarly to other languages. As the number of readily available corpora is increasing all that time this not only means that it should be possible to test the methodology on other languages but also on larger data samples. The testing of the method on other child ages may also throw more light on the movement towards adult language that the FUSE method seemed to detect in the Finnish and English samples. Perhaps testing FUSE with more closely related languages may show the abilities of the FUSE measure to recognise this

closeness between the languages in the early child data samples as well as showing the movement towards the adult language.

Further applications could involve the development of the methodology itself, perhaps, for example to give a more important role to frequency by using frequency of word types in the assessment of FUSE. Similarly, the overall frequency of phonemes could be used as part of the FUSE count by utilising word tokens rather than word types as done for this first application.

One point to be highlighted by the results of the minimal pair groupings is the difference in number of minimal pairs with a word initial contrast found for the two adult languages. The adult English word data, although containing fewer word types than the Finnish data, demonstrated more minimal pairs with a word initial phoneme contrast. Although not reported here word final phonemes were found to mark significantly more contrasts in the Finnish language and could therefore be predicted to be of more importance in marking lexical differentiation than word initial phonemes. Future work might therefore look at other word positions and perhaps compare the usage of contrastive phonemes from a word position basis.

Assessing other word positions than word initial would also enable the importance of particular word positions for different languages to be better observed cross-linguistically. Different languages may use particular inflectional morphemes, for example, that have an effect on minimal pairs as does the positioning of affixes. The supra-segmental element of stress might also be built into the FUSE method such that the stressed syllables of words are assessed.

The observation of Child Directed Speech would be another interesting area to analyse for FUSE in order to assess whether a particular child's speech more closely reflects his individual learning environment than the ambient adult language overall.

In summary, this study has set out to explore the relationship between Finnish and English ambient language and phonological acquisition.

A new method of assessment, incorporating features of the phonological system and the usage of particular phonemes within the system, has been developed thereby providing a new way of viewing the interaction of language specific and universal components of the acquisition process. From this process a range of new findings about the usage of the phonemic system of the two languages has been produced.

The new information about phonemic usage and the structures of Finnish and English words and syllables, both in adult and child speech, that is now available is anticipated to be of future use to linguists working in the fields of child language and speech processing, particularly in Finland. The data provided in this study can be also seen to have relevance for assessment and remediation of phonological disorders in children as the question of which contrasts should be introduced into a phonological system before others can be partly assumed by FUSE findings provided in this study.

The correlation findings tend to support the view that children, already at age 2, are using words that more closely reflect the phonemic usage in their ambient surrounding language than any universal basis. This may be because already at age 2 the children are quite a way along their developmental paths. Both frequency and FUSE based methods of assessment similarly show this intra-linguistic closeness, suggesting a role for language specific rather than universal theories of acquisition.

Whilst more work would need to be done to similarly assess other languages, larger data samples, other word positions and for a greater range of child ages in order to fully support FUSE as an observational tool the systemic based FUSE method does appear to enable a movement over the three child ages assessed towards the adult system to be better observed than with the purely frequency based methods. That a movement towards the ambient same language (and away from the other ambient language) was detected with FUSE may be indicating that children do in fact adopt a universal approach to the acqusition of their phonological system and that the principle driving this process is based upon contrast usage within the phonemic system they are acquiring, rather than frequency.

Bibliography

Aitchison, J. (1994) *Words in the Mind; An Introduction to the Mental Lexicon*, (2nd edition), Oxford : Blackwell.

Amsler, R.A. (1984) *Machine-Readable Dictionaries.* In Williams M.E. ed. *Annual Review of Information Science and Technology*, 19, 161-209. American Society for Information Science(ASIS); Knowledge Industry Publications, Inc.

Anderson, R. and Smith B. (1987) Phonological Development of Two-Year Old Monolingual Puerto Rican Spanish-speaking Children. *Journal of Child Language*, 14, pp. 57-78.

Atwell, E. & Souter, C. (1993) *Corpus-Based Computational Linguistics*, Rodopi, Leeds.

Baayen, H. (1993) Statistical Models for Word Frequency Distributions : A Linguistic Evaluation. *Computers and the Humanities, Volume 26,* Netherlands:Kluwer.

Barrett, M. (1995) Early Lexical Development. In Fletcher, P. and MacWhinney, B. (eds.) *The Handbook of Child Language*. Oxford:Blackwell.

Berndt, R.S., D'Autrechy, C.L. & Reggia, J.A. (1994) Functional Pronunciation Units in English Words. *Journal of Experimental Psychology;Learning Memory and Cognition,* Volume 20, No. 4, APA Inc.

Boysson-Bardies, B. de, Halle, P., Sagart, L., and Durand, C. (1989) A Crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language*, 16, 1-17.

Boysson-Bardies, B. de and Vihman, M. M. (1991) Adaptation to language; Evidence from babbling and first words in four languages. *Language*, 67.

Boysson-Bardies, B. de., Vihman, M. M., Roug-Hellichius, L., Durand, C., Landberg. I., and Arao, F. (1992) Material evidence of infant selection from target language: a crosslinguistic study. In Ferguson, C. A., Menn, L., and Stoel-Gammon, C. (eds.), *Phonological Development: models, research, implications*, 553-62. Timonium, MD: York Press.

Brown, G.D.A. (1984) A frequency count of 190,000 words in the London-Lund Corpus of English Conversation, *Behavioural Research Methods Instrumentation and Computaters*, 16(6), 502-532.

Catford, J.C. (1988) Functional Load and Diachronic Phonology. In *The Prague School and its Legacy*, Tobin, Y ed. Amsterdam:John Benjamins.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.

Chomsky, N. & Halle, M. (1968) *The Sound Pattern of English*. New York:Harper & Row.

Church, K. W. (1993) Introduction to the Special Issue of Computational Linguistics Using Large Corpora, *Computational Linguistics* 19:1.

Clark, J. & Yallop, C. (1995) *An Introduction to Phonetics and Phonology*. 2nd ed. Oxford, Blackwell.

Coltheart, M. (1981) *The MRC Psycholinguistic Database*, Quarterly Journal of Experimental Psychology 33A, 497-505.

Coolican, H. (1994) *Research Methods and Statistics in Psychology* (2nd Edition). London:Hodder & Stoughton.

Creaghead, N. A. (1989) Phonological Development : Assessment and Remediation of Articulatory and Phonological Disorders. London : Merrill.

Crystal, D. (1987) *The Cambridge Encyclopaedia of Language*, CUP.

Dewey, G. (1923) *Relative Frequency of English Speech Sounds*, Cambridge:Harvard University Press.

Dobrich, W. & Scarborough, H. S. (1992) Phonological Characteristics of Words Young Children Try to Say. *Journal of Child Language*, No. 19.

Eckman, R. R. (1977) Markedness and the Contrastive Analysis Hypotheses. *Language Learning* Volume 27, No 2, pp.315-330.

Edwards, M. L. (1974) Perception and Production in Child Phonology; The Testing of Four Hypotheses. *Journal of Speech and Hearing Research*, No. 20, pp.766-780.

Ferguson, C. A. & Farwell, C. B. (1977) Words and Sounds in Early Language Acquisition; English Initial Consonants in the First Fifty Words. In Wang (ed.) *The Lexicon in Phonological Change*, The Hague:Mouton.

Ferguson C. A., Menn, L. & Stoel-Gammon, C. eds., (1992) *Phonological Development: Models, Research, Implications*, Maryland:York Press.

Fletcher P. & MacWhinney B. eds. (1995) *The Handbook of Child Language*, Oxford :Blackwell.

French, N. R., Carter, C. W. and Koenig, W. (1930) The Words and Sounds of Telephone Conversation. In *Bell Systems Technical Journal*, No. 9, pp.290-324.

Garside, R, Leech, G. & Sampson, G. (1987) *The Computational Analysis of English : A Corpus Based Approach,* Harlow:Longman.

Gibbon, D, Moore, R & Winski, R. (1997*) Handbook of Standards and Resources for Spoken Language Systems,* CUP.

Gimson, A. C. (1994) *An Introduction to the Pronunciation of English*, 5th edition. London:Arnold.

Goldsmith, J. A. (1990), *Autosegmental and Metrical Phonology*. Cambridge, Mass: Blackwell.

Greenberg, J. H. (1959), *A Method of Measuring Functional Yield as Applied to Tone in African Languages*. Monograph Series on Language and Linguistics 12:7-16

Grunwell, P. (1985) Phonological Assessment of Child Speech (PACS). Windsor: NFER-Nelson.

Grunwell, P. (1980) Developmental Language Disorders at the Phonological Level. In Jones, F. M. (ed.) *Language Disability in Children* pp. 129-158, Lancaster:MTP Press.

Hakulinen, Auli & Leino, Pentti (1983), *Nykysuomen rakenne ja kehitys*(The Structure of Modern Finnish and Its Development), Suomalaisen Kirjallisuuden Seura:Helsinki.

Harrikari, Heli (1998), *At-will spoonerisms and vowel length in Finnish*, Helsinki University Press.

Hockett. C.F. (1955) *A Manual of Phonology*, Indiana University Publications in Antrhopology and Linguistics; Memory II of the International Journal of American Linguistics, Baltimore, Waverly Press.

Hua, Z & Dodd, B. (2000) *The Phonological Acquisition of Putonghua,* Journal of Child Language, 27.

Häkkinen, Kaisa (1983), *Suomen Kielen Äänerakenteen Ominaispiirteistä,* 'On the Characteristics of Sound Structure in Finnish', in Hakulinen & Leino (1983)

Iivonen, A. (1986) Lapsen fonologisen kehityksen tutkimusmetodiikka. In Lehtihalmes, M. & Klipp, A. (eds.), *Logopedis-fonitrinen tutkimus Suomessa.* Soumen logopedis-foniatrisen yhdistyksen julkaisuja 19, 17-58.

Iivonen, A. (1995) Lapsen fonologis-foneettinen kehitys: 3. Lapsi leksikon kynnyksellä. Suomen logopedis-foniatrinen aikakauslehti, 15, 1-15.

Iivonen, A. (1998), *Aspects of the Phonotactical Acquisition in Children,* Proceedings of the Seventh Nordic Child Language Symposium, No 13, Oulu

Iivonen, A. (1998), Intonation in Finnish. In Hirst, D. & Di Cristo, A. (eds.) (1998) *Intonation Systems : A Survey of Twenty Languages,* CUP.

Ingram, D. (1988) The Acquisition of Word Initial [v]. *Language and Speech* 31:77-85.

Ingram, D. (1989) *First Language Acquisition: Method, Description, and Explanation,* CUP.

Ingram, D. (1989) *Phonological Disability in Children,* Whirr.

Ingram, D. (1999) *Phonological Acquisition* in Barrett, M (ed.) The Development of Language, Psychology Press.

Jakobson, R. (1968) *Child Language, Aphasia, and Phonological Universals* (A.R. Keiler, trans.) The Hague:Mouton.

Jimenez, B. C. (1987) *Acquisition of Spanish Consonants in Children Aged Three to Five Years Seven Months.* Language Speech and Hearing Services in Schools 18:357-363).

Johansson, S. (1995) *ICAME – Quo Vadis? Reflections on the Use of Computer Corpora in Linguistics*, Computers and the Humanities 28:243-252.

Karlsson, F. (1983), *Suomen Kielen, äänne- ja muotorakenne*, WSOY:Juva.

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*, Longman

Kent, R. D. (1992) The Biology of Phonological Development. In Ferguson, C. A., Menn, L. and Stoel-Gammon, C. (eds.) *Phonological Development;Models Research Implications* (p65-90), Timonium MD:York Press.

King, R. D. (1967) *A Measure for Functional Load.* Studio Linguistica, 21, pp.1-14.

Knowles, G. (1987) *Patterns of Spoken English;An Introduction to English Phonetics.* Essex:Longman.

Kunnari, S. (1999) *Word Length in Syllables in Children's Early Words.* Paper presented at the 20th Annual Phonology Conference, Bangor.

Kunnari, S. (2000) *Characteristics of Early Lexical and Phonological Development in Children Acquiring Finnish*, Ph.D. thesis, Acta Universitatis Ouluensis; Oulu.

Ladefoged, P & Maddieson, I (1996) The Sounds of the World's Languages, Oxford:Blackwell.

Lass, R. (1984) *Phonology; An Introduction to Basic Concepts.* CUP.

Leinonen-Davis, E. (1987), *Assessing the Functional Adequacy of Children's Phonological Systems,* Ph.D Thesis, Clinical Linguistics & Phonetics, No.4, 257-270

Leinonen, E. (1990) Functional Motivation in the Development of Phonological Contrasts. In The Proceedings of the Conference on Child Language Disorders. University of Trondheim, Norway.

Leopold, W. (1947) Speech Development of a Bilingual Child; A Linguists Record (Volume 2) Sound Learning in the First Two Years. North Western University Press : Illinois.

Locke, J. L. (1993) The Child's Path to Spoken Language. Harvard University Press.

Macken, M. (1978) Permitted Complexity in Phonological Development; One Child's Acquisition of Spanish Consonants. Lingua 44. 219-53.

Macken, M.A. and Ferguson, C.A (1983*) Cognitive Aspects of Phonological Development: Model, Evidence and Issues* in K.E. Nelson(ed.) Children's Language, Vol.4, pp. 256-282. Hillsdale, NJ: Erlbaum.

Maddieson, I. (1984), Patterns of Sounds, CUP.

McEnery, T. & Wilson, A. (1996) *Corpus Linguistics.* Edinburgh University Press.

MacWhinney, B. (1995) The CHILDES Project; Tools for Analyzing Talk (2nd. Ed.) Carnegie Mellon University : Lawrence Erlbaum Associates.

Menn, L. & Stoel-Gammon, C. (1995) Phonological Development. In Fletcher, P. & MacWhinney, B. (eds.), *The Handbook of Child Language*, Oxford:Blackwell.

Menyuk, P. (1968) The Role of Distinctive Features in Children's Acquisition of Phonology. *Journal of Speech and Hearing Research*, Vol. 11, pp138-46.

Meyerstein, R. B. (1970) *Functional Load.* The Hague : Mouton.

Mitton, R. (1986), *A Description of a Computer-Usable Dictionary File Based on The Oxford Advanced Learner's Dictionary of Current English*, Oxford Text Archive Web Site, UK.

Miller, G.A. (1963), *Language and Communication*, London:McGraw-Hill.

Mines, M.A., Hanson B.F. & Shoup J.E. (1978) *Language and Speech*, Vol 21 part 3.

Oller, D. K. (1973) Regularities in Abnormal Child Phonology. Journal of Speech and Hearing Disorders, 38, pp.36-47.

Olmsted, D. (1966) A theory of the child's learning of phonology, *Language* 42 :531-5.

Olmsted, D. (1971) *Out of the Mouth of Babes*, The Hague:Mouton.

Pajunen, A. & Palomäki, U. (1984), *Tilastotietoja Suomen Kielen Rakenteesta* (Frequency Analysis of Spoken and Written Discourse in Finnish), Kotimaisten Kielten Tutkimuskeskus:Helsinki.

Paunonen, H. (1994), *The Finnish Language in Helsinki in Nordber*, B(Ed.) The Socialinguistics of Urbanization:The Case of the Nordic Countries, Berlin:Walter de Gruyter.

Paunonen, H. (1995), *Suomen Kieli Helsingissä, Huomioita Helsingin puhekielen historiallisesta taustasty ja nykyvariaatiosta*, Helsingin yliopiston suomen kielen laitos, Hakapaino Oy, Helsinki.

Pye, C., Ingram, D. and List, H. (1987) A Comparison of Initial Consonant Acquisition in English & Quiché. In K. E. Nelson & A. Van Kleeck (eds.) *Children's Language*, vol. 6:175-90 Hillsdale, NJ:Erlbaum.

Pääkkonen, Matti (1973) *Grafeemit Ja Konteksti* (Graphemes and Context), Suomalaisen Kirjallisuuden Seura:Helsinki

Roach, P. (1991) *English Phonetics and Phonology ; A Practical Course* (2nd Edition). CUP.

Rumelhart, D. E. & McClelland, J. L. eds. (1986) *Parallel Distributed Processing; Explorations in the Microstructure of Cognition*, Volume 2, Psychological and Biological Models, pp216-71. Cambridge, Mass:MIT.

Sampson, G. (2001) *Empirical Linguistics*. Continuum International.

Sander, E. (1972) *When Are Speech Sounds Learned?* Journal of Speech and Hearing Disorders 37:55-63.

Savinainen-Makkonen, T. (2000) *Suomalianen Lapsi Fonologiaa Omaksumassa*, Publications of The Dept. of Phonetics *42*, University of Helsinki;Finland.

Schwartz, R. G. (1988) Phonological Factors in Early Acquisition. In Locke J. L. *The Emergent Lexicon : The Child's Development of a Linguistic Vocabulary*. Academic Press.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*, Oxford University Press

Skinner, B.F. (1957) *Verbal Behaviour*. Englewood Cliffs, NJ:Prentice-Hall

Slobin, D. I.. (1985) *The Crosslinguistic Study of Language Acquisition, Volume 2, Theoretical Issues*, Hillsdale, NJ:Erlbaum Ass.

Slobin, D. I.. (1992) *The Crosslinguistic Study of Language Acquisition, Volume 3,* Hillsdale, NJ:Erlbaum Ass.

Slobin, D. I. (ed.) (1997) *The Crosslinguistic Study of Language Acquisition, Volume 4,* London:Erlbaum Ass.

Smith, N. V. (1973) *The Acquisition of Phonology;A Case Study.* Cambridge:CUP.

Stampe, D. (1969) *The Acquisition of Phonetic Representation.* Papers from the Fifth Regional Meeting of the Chicago Linguistic Society, Chicago, IL. Reprinted in D.

Stampe, D. (1979), A Dissertation on Natural Phonology. New York:Garland.

Stoel-Gammon, C. and Dunn, C. (1985) *Normal and Disordered Phonology in Children,* Austin, Texas:Pro-Ed.

Stoel-Gammon, C. & Cooper, J.A. (1984) Patterns of Early Lexical and Phonological Development, *Journal of Child Language,* Volume 11 (247-271).

Stoel-Gammon, C. and Herrington, P. B. (1990) Vowel Systems of Normally Developing and Phonologically Disordered Children. *Clinical Linguistics and Phonetics*, No. 4. pp145-160.

Stoel – Gammon, C. (1998) Sounds and words in early language acquisition: the relationship between lexical and phonological development. In Paul R. (eds.) *Exploring the Speech Language Connection.* Baltimore: Brookes Publishing.

Sulkala, H. and Karjalainen, M.. (1992), *Finnish*, London:Routledge

Svartvik, J. (1990) *The London-Lund Corpus of Spoken English:Description and Research*, Lund University Press.

Templin, M. (1957) Certain Language Skills in Children; Their Development and Inter-relationships. The Institute of Child Welfare Monograph. Minneapolis: Minnesota Press.

Tobin, Y. (1997) *Phonology as Human Behaviour, Theoretical Implications & Clinical Applications*, Duke University:Durham.

Toivainen, J. (1990) *Acquisition of Finnish as a First Language ; General and Particular Things*, Dept. of Finnish and General Linguistics, Turku University.

Toivainen, J. (1997) The Acquisition of Finnish. In Slobin, D. I. (ed.) *The Crosslinguistic Study of Language Acquisition,* London:Erlbaum Ass.

Vainio, M. (1996) *Phoneme Frequencies in Finnish Text and Speech* in Iivonen & Klippi (eds.) Studies in Logopedic and Phonetics 5, Helsinki University Press.

Vihman, M. M. (1981) *Phonology and the Development of the Lexicon; Evidence from Childrens' Errors*, Journal of Child Language, Vol. 8, pp. 239-264.

Vihman, M. M. (1996) *Phonological Development ; The Origins of Language in the Child.* Oxford:Blackwell.

Wang, W. & Crawford, J. (1960) *Frequency Studies of English Consonants.* Language and Speech 3:131-9.

Wang, W. (1967) *The Measurement of Functional Load.* Phonetica 16:36-54

Wells, J.W. (1986) *A Standardised machine-readable phonetic notation.* In Proceedings of the IEE conference on speech input/output techniques and applications. London,

Wiik, K. (1977), *Suomen Tavuista*, Virittäjä 81:265-278.

Wilson, M. (1987), MRC Psycholinguistic Database:Machine Usable Dictionary Version 2.00. Science and Engineering Research Council, Oxon.

Zipf, G.K. (1935) *The Psycho-Biology of Language.* Cambridge:MIT Press.

Corpora

Helsingin Puhekielen Aineisto, University of Helsinki Multilingual Data Bank, Dept.of General Linguistics, Helsinki University, Finland.

Oulun yliopisto, Suomen kieli, Lapsenkielen materiaali 'Oulu University, Finnish Child Language Material', Oulu University, Finland.

The CHILDES Project 2000, [CD-ROM]. Carnegie Mellon University, USA.

MRC Psycholinguistic Database, Version 2.Oxford Text Archive, Online Access Web site 2000/2001, //www.psy.uwa.edu.au/MRCDataBase/mrc2.