Australian Journal of Educational & Developmental Psychology. Vol 4, 2004, pp 13-28

Issues Related to the Construction of a Purpose-Built Domain-Specific Word Corpus

Lisa Thomas[1], H. Peter Pfister[1], and Peter Peterson[2]
[1]School of Behavioural Sciences, [2]School of Language and Media
University of Newcastle
NSW Australia

There is growing interest in the use of semantic collections in order to identify and analyse domain knowledge. This paper describes some technical issues to consider when contemplating research which incorporates small-to-medium domain-specific word sets. The purpose of the corpus construction described was to provide an external word collection which could be transformed to a numeric frequency scale which could take the place of an "expert" in order to evaluate the lexical content of aircraft Visual Landing Approach concept maps. Although this paper is based on research in the field of aviation education, the underlying principles are more widely applicable.

General Corpora Definitions and Uses

The study of naturally occurring word frequencies has been a focus of computational linguistics, and in particular of the field of corpus linguistics. A word collection known as a corpus is constructed from some set of texts in order to determine what is characteristic of that text set through the identification of vocabulary patterns that either differ or conform to a norm (Ide & Walker, 1993). In contrast to Chomskyan generative linguistics, which focuses on internal knowledge of language structures (Chomsky, 1957), empirical corpus linguistics seeks to describe language as it actually occurs, including the number and range of words which are appropriate in a defined context and which may not conform to the constraints of correct grammaticality and syntactic rules (Sampson, 1987; Kennedy, 1992). The underlying premise or assumption of an empirical approach is that semantic content can be meaningfully related to quantifiable word patterns in source texts drawn from constrained natural language sub-sets.

Corpus linguistics uses this representative sample of spoken and/or written words in order to provide an authoritative body of linguistic evidence, one which can support generalizations and against which hypotheses can be tested (Trask, 1993; Coulthard, 1994). For example, it is reasonable to assume that novices and experienced individuals in a specific field may use either different domain vocabularies, that is, different words, or different relative distributions of the same words (Solomon, 1990). Word frequency distribution may be used in uni-textual analysis or for parallel text comparisons (Francis & Kucera, 1982; Sinclair, 1991; Caudery, 1992; Coulthard, 1994; Davis, Dunning & Ogden, 1995; McEnery & Wilson, 1996). One drawback however is that the simple word counts and values derived from those counts are not generally sufficient for word sense disambiguation (Toglia *et al*, 1978; Manning & Schutze, 1999).

Studies of word frequencies and frequency distributions are not restricted to corpus linguistics. Thematic content analysis, which is based in the social sciences (de Sola Pool, 1959; Saporta &

ISSN 1446-5442 Web site: http://www.newcastle.edu.au/journal/ajedp/

Sebeok, 1959; Iker, 1974; Beardsworth, 1980), and literary analysis (Burrows, 1987; Ide, 1989; Ide & Walker, 1993) are also concerned with quantifying and interpreting word occurrences, and both approaches have benefited in recent years from computerisation (Roller, Mathes & Eckert, 1995).

An example of this type of word-based thematic analysis is found in the integrative cognitive complexity scale described by Baker-Brown, Ballard, Bluck, de Vries, Suedfeld, and Tetlock (1992) in which the presence of text elements gives evidence of progressive differentiation and/or integration. This coding system focuses on structure rather than content, and its devisors suggest it can be used with any connected verbal discourse.

Corpus linguistics can complement these other language-based analysis methods (Biber, 1993). However corpus linguistics has one specific advantage in that it may require relatively less qualitative contextual knowledge for meaningful application (McEnery & Wilson, 1996). As a minimum, corpus analysis requires only that any sample of words from a defined domain be a domain representative sample (Baayen, 1993).

Word Frequency Divisions

A word frequency analysis typically involves raw word counts, ranks, and weights, and the comparison of these between different sources. A corpus which has been constructed from a representative selection of texts is more likely to demonstrate a range of word frequencies than one which has been constructed with bias (McEnery & Wilson, 1996). It follows that when a corpus has been derived from naturally occurring texts it may be partitioned into frequency divisions which indicate functional differences within its overall word frequency spectrum.

A high-frequency group typically includes several functional/structural words (e.g., to, of, in, at, and, for, than) which are indicative of English language structure. High frequency words tend to have more diverse meanings than do lower frequency words, which implies a correlation between frequency and semantic complexity (Kilgarriff & Rosenzweig, 2000). A relatively high word frequency does not imply conceptual validity for an individual word or for the passages in which that word participates. For example, individually a high frequency word is just that- a word which occurs frequently. The presence of two high frequency words in some sort of relationship says nothing about the correctness of that relationship, just that there *may be* a likelihood of those words co-occurring.

The medium-frequency range of words denotes words of lesser generality but also of repeat frequencies (Herdan, 1964). Within this group, given a typical distribution not skewed by underlying functions, may be found a class of commonly used content words.

Low-frequency words tend to bear greater informational value than words which occur more frequently (Herdan, 1964). The percentage of rare words as a representative feature of a text represents the richness, or diversity, of the text (Liiv, 1997). The size of the group of words which occur only once, denoted by the term "hapax legomena", is a measure of vocabulary richness, and grows with an increase in vocabulary, which is an indication of word learning (Holmes, 1994). Sichel (1986, p. 53) noted that:

> the number and proportion of hapax legomena have been used to measure vocabulary richness in general. A person with a large proportion of hapax legomena is considered to command a richer vocabulary than one with a low proportion.

A related rare word category is the hapax dislegomena, or the collection of text words which are used twice. A practical difficulty with compiling word frequency counts from published, “sanitized” texts, and from small corpora is that the sample range may not produce reliable counts of these low frequency words (Burgess & Livesay, 1998).

In summary, when a corpus has been derived from naturally occurring texts, functional differences may translate into differences in word frequencies. On the other hand, individual words with similar frequencies may have different functions. It may be noted when referring to word frequencies, whether as raw frequency counts, or as frequency related weights, that similar values do not necessarily imply semantic collocations, collocations being characteristic co-occurring patterns of words. It is more likely that words of similar frequencies and similar functions or meanings will not in fact occur together (Haskel, 1971; Berry-Rogge, 1974; Delcourt, 1992), and two words that have similar meanings, individually may occur with similar frequencies, but *not together* unless for emphasis.

Corpus Size

Many of the statistics regarding words and their frequency distributions derive from works on corpora that involve large numbers (>1,000,000) of words which are in general use (e.g., the Brown Corpus of contemporary American English and the British National Corpus). Large corpora are generally constructed from a range of diverse text categories, in order to make possible cross-category comparisons (Hofland & Johansson, 1982; Johansson, 1985). Smaller purpose-built corpora deriving from more restricted domains are increasingly common with the availability of software (Ide & Walker, 1993; Hickey, 1994). Baayen (1993) noted that a relatively small size corpus may be useful if it is used to study only that limited range of topics actually contained in the corpus. Therefore, a purpose-built corpus need not approach the size of the large general purpose corpora. However, the provision of a workable scale may necessitate the incorporation of a much greater word range than the potential word range from the topic under study itself.

Source Selection Issues

The goal of document location and selection is to create a text collection that would be maximally representative of available aviation texts. For a text sample to be of use, it must be typical, representative and unbiased, and include, where appropriate, samples of a broad range of authors and genres (Biber, 1988, 1993; Althiede, 1996).

Selection of Corpus Inputs

The particular corpus described in this paper was constructed from three sources, the desired outcome being a representative range of Visual Landing Approach related words from both written and spoken sources. The largest source in terms of word number comprises 92 texts from aviation documents selected for their relevance to the Visual Landing Approach. Together these 92 texts are identified as Input A.

These documents which comprise Input A were products of particular historical and cultural influences and were chosen to recreate a broad historical aviation scope, with the emphasis on the major themes and sources relating to Visual Landing Approach flight instruction. Each was specifically selected in order to mirror the scope of information which might be either provided to or be otherwise available to Australian flight instructors and flight students. Aviation museums and libraries in Australia, the United Kingdom, Canada, and the United States were contacted. These organisations provided roughly half of the documents evaluated for corpus inclusion.

Other major sources included active flight training organisations in those countries mentioned and private individuals.

One hundred and sixteen discrete texts were originally evaluated for content and relevance. Of these, twenty-four were eliminated because they were deemed to be too technical in their outlook, or they would have overly biased the geographical or chronological text distribution. Additionally, both the number and diversity of the corpus Input A source texts were intended to lessen the possibility that the overall corpus word range would be overly constrictive in vocabulary, as both Inputs B and C, described below, were derived solely from contemporary Australian sources.

Of the 92 texts finally selected for Input A, 27 came from the United Kingdom, 47 from the United States of America, 11 from Australia, 3 from Canada, 3 from New Zealand, and 1 from Norway (in English). The texts were drawn from a 90 year range and included both military and civil aviation sources, and official and popular genres. These 92 texts did not necessarily comprise whole documents. When identifiable Visual Landing Approach themes were embedded in a larger document, the relevant sections were selected out of that document.

These 92 texts may be described as “examples”, as opposed to “samples” chosen by a random selection, of Visual Landing Approach writing. As a proper sample suitable for full linguistic analysis, this “exampling” method of document selection would have been inappropriate. Neither have the document excerpts been chosen through spread sampling, which Yule (1944) suggested may be even more closely representative than a random sample of the same size, but criticized as being prone to word “clumping” if not done properly.

Input B consists of the transcriptions of five oral interviews with Australian flight instructors. All these interviews were conducted by the researchers during the period 1998 through 2000. Input C is composed of the 77 responses provided by Australian General Aviation pilots to an open-ended survey questionnaire on Visual Landing Approach expertise conducted by the researchers in 1999.

Therefore the Visual Landing Approach corpus comprised three separate input types, two of which derive from written sources (82% of total words) and one from oral sources (18% of total words). As a comparison, the 100,000,000 word British National Corpus contains approximately 10% from spoken data. The general semantic differences between written and spoken genres are highlighted in Biber (1988, 1993), Chafe (1986), Hayes (1988), and Halliday (1994).

Corpus Word Weighting

The principal purpose of the weighting procedure used in this study was to correct for the over- or under- representation of words due to relative size differences in their input sources. A secondary purpose was to minimize subjectivity effects arising from document selection and excerption. There are at minimum five different ways of weighting text excerpts, ranging from no weighting at all, through intermediate procedures including variations on set block-length averaging, to a fully normalized procedure. (J. Burrows, personal communication, March 2000). All sampling and weighting procedures are in spirit ultimately based upon the works of Yule (1944) and Zipf (1949) although they themselves diverged on how to list different forms of the same word, Zipf choosing to include every variation as a separate entry while Yule preferred to group them under one heading.

Constituent Length Adjustments within Inputs A, B and C

Prior to applying the word weighting procedure, the constituent units within the corpus were adjusted to compensate for large disparities in text length. The 92 texts in Input A had a total word count of 46,180. These texts were of varying lengths ranging from 108 to 1128 words (M = 501.13, SD 263.887). For the weighting procedure, each text was considered as a component.

Input B originally consisted of five interviews with a total word count of 10,469. One interview contained 8,328 words with the next largest containing 1,053 words. The remaining three interviews totaled 1,088 words among them. These three interviews differed from the other two in length and from the longest interview also in purpose. Unlike the longest interview, which was free-flowing, the three short interviews were essentially single-question probes, as was the second longest interview. Based on these considerations, the three short interviews were combined into a single unit. Input B therefore consists of three components.

The 77 survey responses which make up Input C have a total word count of 2,178. The responses ranged in length from one to 103 words (M = 28.987, SD 16.509). The relative brevity of the responses when compared to the constituent elements of inputs A and B argued against maintaining each of the 77 as a separate component for the purpose of word weighting. Therefore Input C consists of one component.

Following these adjustments, which left intact the overall input file word length, the components in inputs A, B and C by which the subsequent word weights were derived were as follows:

Table 1: Number of Corpus Input Words and Components

Input	*Number of Words*	*Number of Components*
A	46,180	92
B	10,469	3
C	2,178	1
Totals	58,827	96

Internal Weighting Factors

The inputs to this Visual Landing Approach semantic data base consist of text excerpts of widely differing lengths. Therefore within each of Input A, Input B, and Input C a simple word count would result in relatively greater contributions from the longer extracts. Similarly, because Input A, Input B, and Input C contain a differing total number of words, a simple word count would result in a greater relative contribution from the longer Input A than from the shorter inputs B and C. To compensate for different corpus component lengths, inputs were normalized by an internal and an external word weighting procedure.

Within each of Input A and Input B the total number of words was divided by the number of components to obtain a mean word count. For each component this mean word count was divided by the number of words in the component to provide an Internal Weighting Index. This procedure was not applied to Input C due to the general brevity of the individual components of Input C.

The mathematical expressions are as follows:

The mean number of words in Input A is $\tilde{n}_A$ where $\tilde{n}_A$ is given by:

$$\tilde{n}_A = \frac{1}{N_A} \sum_{i=1}^{i=N_A} n_{Ai} \qquad (1)$$

where N_A = number of components in Input A
n_{Ai} = number of words in component i of Input A

The Internal Weighting Index for component i of Input A is t_{Ai} where t_{Ai} is given by:

$$t_{Ai} = \frac{\tilde{n}_A}{n_{Ai}} \qquad (2)$$

Similar equations apply to Inputs B and C, i.e.,

$$\tilde{n}_B = \frac{1}{N_B} \sum_{i=1}^{i=N_B} n_{Bi} \quad , t_{Bi} = \frac{\tilde{n}_B}{n_{Bi}} \qquad (3)$$

and

$$\tilde{n}_c = \frac{1}{N_C} \sum_{i=1}^{i=N_C} n_{Ci} \quad , t_{Ci} = \frac{\tilde{n}_C}{n_{Ci}} \qquad (4)$$

The derivation of the Internal Weighting Index is illustrated by its application to the interview Input B. After adjustments as discussed above, Input B consisted of three components, each with a different Internal Weighting Index:

Table .2: Internal Weighting Index for Input B

Components	*Number of Words*	*Internal Weighting Index = Mean/Number of Words*
1	8,328	0.419
2	1,053	3.314
3	1,088	3.207
Total	10,469	
Mean	3,490	

That is, each individual word in Input B component 1 is multiplied by 0.419, each word in component 2 is multiplied by 3.314, and each word in component 3 is multiplied by 3.207.

A like procedure was applied to the document Input A to obtain Internal Word Weighting Indices for each of the 92 components. As a length-based Internal Weighting Index was not calculated for Input C, the Input C Internal Weighting Index was 1.0. That is, every word in Input C counted as one Internal Weight unit.

External Weighting Factors

A similar procedure was used to produce an External Weighting Index for each total Input A, Input B, and Input C. In this case the total number of words in the three inputs was divided by three to obtain a mean word count. For each input A, B, and C this mean word count was divided by the number of words in the input to provide an External Weighting Index.

The mean number of words in the three inputs, A, B and C is $\tilde{n}_{ABC}$ where $\tilde{n}_{ABC}$ is given by:

$$\tilde{n}_{ABC} = \frac{\left[\sum_{i=1}^{i=N_A} n_{Ai} + \sum_{i=1}^{i=N_B} n_{Bi} + \sum_{i=1}^{i=N_C} n_{Ci} \right]}{3} \qquad (5)$$

The external weighting factor for Input A is T_A and is given by:

$$T_A = \frac{\tilde{n}_{ABC}}{\sum_{i=1}^{i=N_A} n_{Ai}} \qquad (6)$$

Similarly the external weighting factors for inputs B and C are given by:

$$T_B = \frac{\tilde{n}_{ABC}}{\sum_{i=1}^{i=N_B} n_{Bi}} \qquad (7)$$

and

$$T_C = \frac{\tilde{n}_{ABC}}{\sum_{i=1}^{i=N_C} n_{Ci}} \qquad (8)$$

The External Weighting Indices for the three inputs A, B and C are illustrated in Table 3:

Table 3: External Weighting Indices for All Inputs

Input	*Number of Words*	*External Weighting Index = Mean/Number of Words*
A	46,180	0.4246
B	10,469	1.873
C	2,178	9.003
Total	58,827	
Mean	19,609	

The overall weighting factor for any word in a document is given by the product of the internal and external weighting factors. e.g., the overall weighting factor for any word in Document i in Input A is given by $t_{Ai}\ T_A$.

Lexical Issues in Relation to Word Distributions

Lexical analysis requires a definition of a "word", and of the forms that a word may take. A graphic word may be defined as a sequence of alphanumeric characters surrounded by spaces, and may contain punctuation marks (Hofland & Johansson, 1982). A lexical word is one or more grammatical words which form a lexical unit. That is, a lexical word fills a single grammatical position and has a generally consistent meaning (Francis & Kucera, 1982). A lexical word may also be referred to as a lexeme (Trask, 1993). For example, "cat" and "cats" are both particular forms of the lexeme "cat".

A lexemic word and a graphic word are not necessarily commensurate, and information based on word counts may differ slightly between analysis stages and across software. This is because the definition of a "word" differs slightly across software, due to the character strings that a given software program will recognise as a meaning-conveying string. This can give rise to a class of ambiguous (non-interpretable) character strings, as distinguished from the set of unambiguous decipherable character strings. Landini (2000) argues for the desirability of deleting ambiguous character strings prior to statistical analysis. In the current analysis, a limited group of ambiguous character strings were included in word counts because of their minimal effect on overall distributions.

This analysis necessitated additional tests of file characteristics derived from linguistic and literary fields in order to demonstrate the properties that held within and across the corpus documentary inputs. The research tests employed in the characterisation of textual features were two standard Zipf analyses, the Yule's K word characteristic, and the token/text ratio. A computer based concordance program was used in conjunction with Microsoft® Excel and Microsoft® Word in the combination of lexemes and computation of lexeme (word-stem) weights.

Zipf Distribution

Zipf (1949) described two general properties of natural languages, since accorded the status of "laws", these being the rank-frequency law and the number-frequency law. The rank-frequency law says the plot of log (frequency) (y-axis) versus log (rank) (x-axis) approximates a straight line of slope –1. Figure 1 shows such a plot for all words in the Visual Landing Approach corpus. The line shown has a slope of –1.

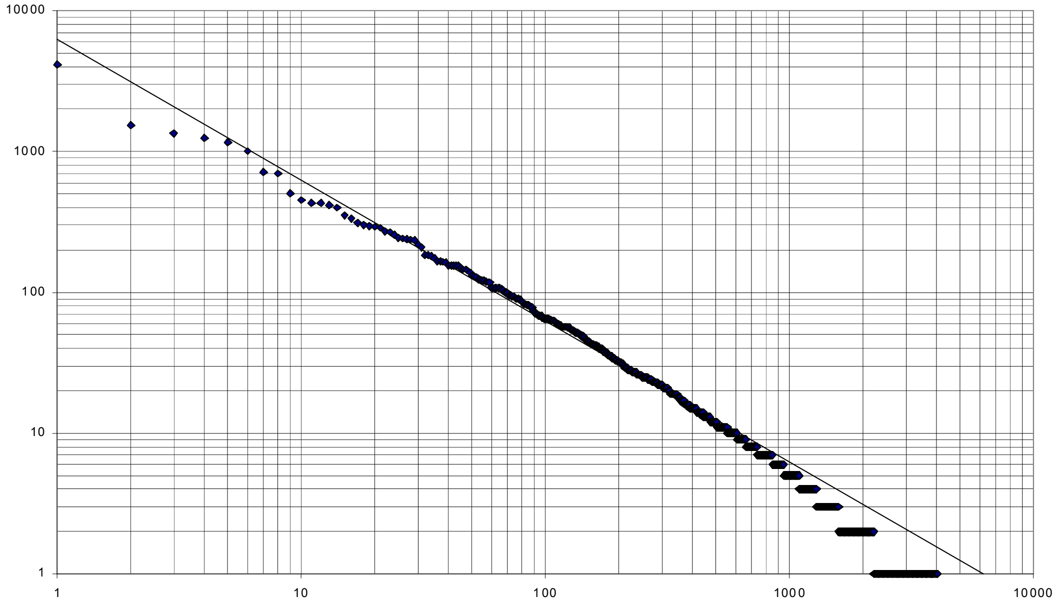

Figure 1. The plot of the Visual Landing Approach corpus according to Zipf's rank-frequency law.

The number-frequency law says that, n being a word's frequency, the plot of log (n) (y axis) versus log (number of words with frequency n) (x axis) approximates a straight line of slope –0.5. Figure 2 shows such a plot for all words in the Visual Landing Approach corpus. The line shown is of slope –0.5.

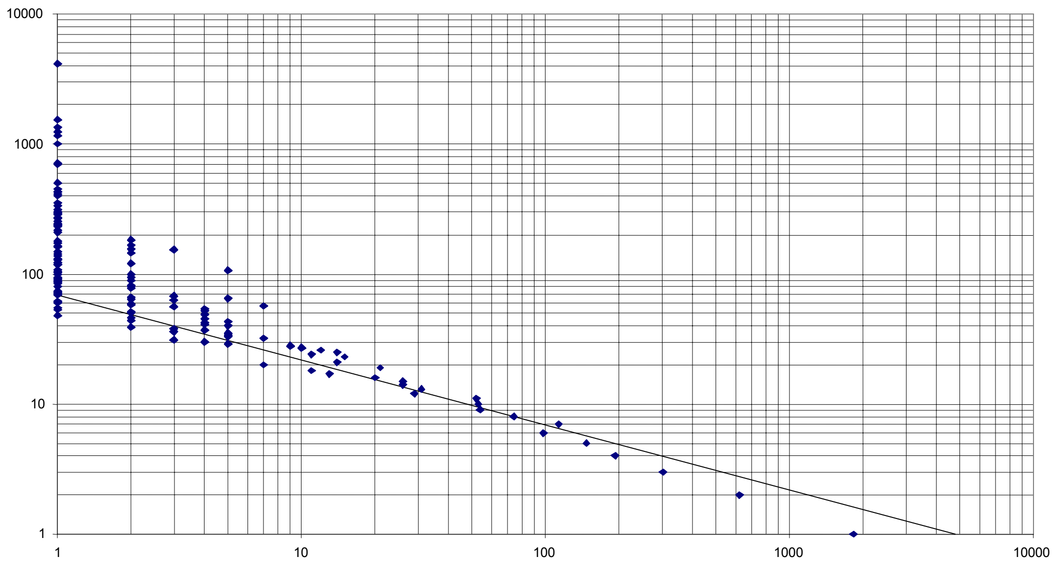

Figure 2. The plot of the Visual Landing Approach corpus according to Zipf's number-frequency law.

Landini (2000) noted that the rank-frequency law tends to be most clearly observed with high frequency words, and the number-frequency law with low frequency words (see also Turner, 1997). These properties hold for a variety of alphabetic, syllabic, and logographic natural languages. They also appear to hold for both written and spoken discourses (Balasubrahmanyan & Naranan, 1996), with constructed language types (Chen, 1991; Li, 1992), and with small word sets (Ridley & Gonzales, 1994). Analysis of semantic distances from the ideal slope may used to differentiate between different texts, and may also be used to indicate when a text has been selected or edited with bias.

Figure 1 indicates that, apart from the three most frequent words, the corpus words do generally follow a slope of –1, only deviating for words of frequency less than about ten. Furthermore, Figure 2 indicates that the low frequency corpus words tend to follow a slope of –0.5. The corpus word distribution therefore follows the Zipf "laws", with the qualification noted by Landini (2000).

Yule's K

Yule's K index of uniformity is a derived characteristic of word distributions which reflects the concentration of high frequency words (Yule, 1944) and is a standard index available in the Simple Concordance Program 4.0.4 . As a measure of vocabulary richness it is based on the assumption that the occurrence of any given word is based on chance and may be regarded as a Poisson distribution. A comparatively large K implies that the author's vocabulary is highly concentrated on repeated words, whereas a comparatively small K indicates that the author's vocabulary is less concentrated. Holmes (1994) noted that Yule's K is constant with respect to the size of the sample text, which is a favorable feature in text comparisons.

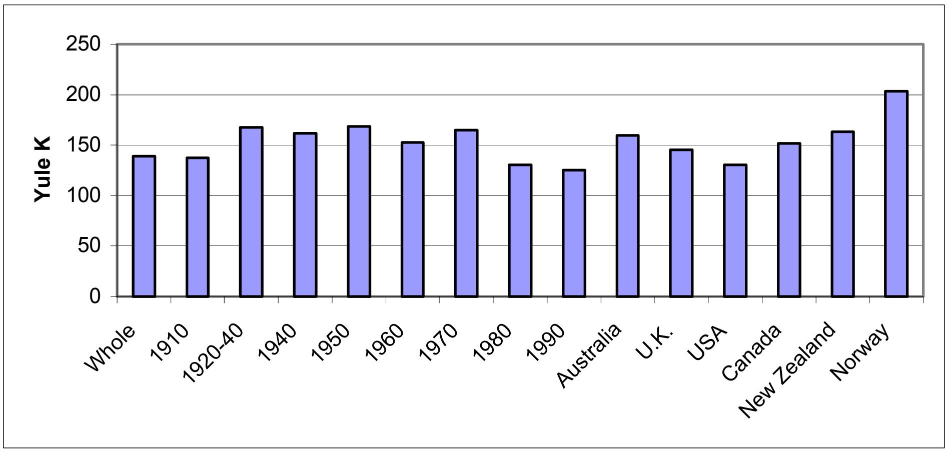

Figure 3. Chart of Yule's K characteristic for Visual Landing Approach corpus Input A and Input A corpus sub-divisions by decade and country of origin.

In Figure 3, the sub-divisions represent different methods of apportioning the texts which made up Input A, when the age and provenance of those texts were considered. Yule's K for the total word set in Input A was 138.92. For comparison, the Yule's K value for Input B was 89.19, and for Input C was 101.77. Therefore the words in the document Input A were more concentrated on repeated words than were the words from the survey responses (Input C), which in turn were more concentrated on repeated words than the words from the flight instructor interviews in Input B. These results were to be expected due to the degree of "focus" of each genre. The relative magnitude of these differences is not as great, however, as the differences in Yule's K values for the overlapping sub-sets in Figure 3. The Yule's K value for the corpus overall, that is Input A, Input B, and Input C, was 123.92. That is, when all three inputs were considered together, the high frequency concentration within the largest input (Input A) was diluted. This was one of the desired outcomes from the combination of different text genres into the Visual Landing Approach corpus.

The Text/Token Ratio

The concordance program provided the text/token ratios for the word sub-sets in this study. The text/token ratio is an indication of vocabulary diversity, for it compares the number of unique word types to the overall size of the document(s) from which they are drawn. In Holmes' (1994) review of quantitative stylistic metrics, he noted that the text/token ratio is sensitive to variations in document length, and is only useful for document comparison with texts of similar length. The effect of text size on the text/token ratio is evident in the values in Figure 4.

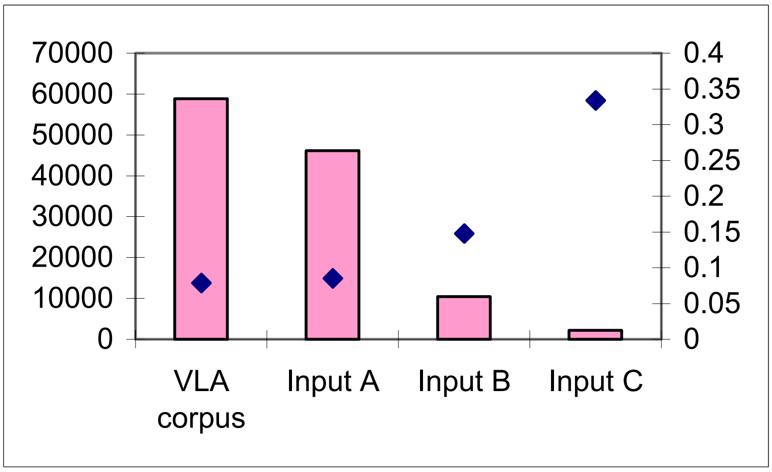

Figure 4. Word counts (columns, left scale) and associated text/token ratios (right scale) for Visual Landing Approach corpus and corpus sub-divisions.

Although Inputs A, B, and C together comprise the totality of the Visual Landing Approach corpus, individually the input text/token ratios vary among themselves and also when compared against the overall corpus value. These ratio differences are noted as confirmation of Holmes (1994) but they do not in themselves invalidate the use of the corpus text/token value for this research, which is as a general indication of the available number of unique corpus text words against which the concept map words could be matched, and not for document comparison *per se*. According to the concordance program the corpus contained 58827 token words with a text/token ratio calculated at .0788, indicating a total of 4635 available distinct orthographic words available for concept map word comparison.

Lemmatisation of Corpus Words

"Lemmatisation" is the term given to the grouping of grammatical words with the same stem or the same meaning and which belong to the same major word class, and which differ only in inflection or spelling. (Francis & Kucera, 1982). Grouping related words into their respective word stems, or lemmas, results in a more representative distribution of word types within a defined text. For example, grouping the forms “fly”, “flew”, “flies”, and “flying” into a lemma of “fly”, and summing associated values for those forms into that lemma provides a more realistic indication of the importance of the concept of “fly” within a word set than allowing the separate forms to stand on their own. Therefore lemmatisation of the corpus words was necessary in order to reflect the relative importance of those lemmas within the overall Visual Landing Approach vocabulary.

Abbreviations and their related forms were combined when the abbreviation could be interpreted unequivocally. Mathematical formulas and symbolic relationships generally were not decomposed into their constituent elements. Spelling variants were considered to have semantic equivalence. Despite the increasing homogeneity of British, American, and Australian English (see Ramson, 1966; Collins & Blair, 1989; Cornelius, 1989; Taylor, 1989), numerous words demonstrated both a British and an American orthography (e.g., centre, center; aeroplane, airplane). All recognizable variants, including mis-spellings, were grouped with their appropriate word stems.

Various pseudo-words, constructed by the use of a hyphen or slash to form a compound from two otherwise unlinked constituents, such as “five-year”, appeared in the documents. In order to eliminate these pseudo-words, when the constituent elements were themselves meaningful words the hyphen or slash was disregarded, allowing each de-hyphenated element to be treated as a token in its own right.

Individual examples were checked against their entries in Webster’s Dictionary of the English Language (McKechnie, 1987), and if there were substantial differences in meaning in derivative forms, those forms were not combined. Noun plurals and verb tenses were added to their root forms to produce one word. Although most adjectival and adverbial forms were combined and their weights added, this was not obligatory. Nominalised verb forms were generally not coupled with their verb stems.

A marked form is a form or construction which differs from another with which it stands in a paradigmatic relationship. For example, the lexical items “hostess” and “inconsistent” are marked with respect to “host” and “consistent” (Trask, 1993). Generally, marked forms were not lemmatised, although individual examples were considered for lemmatisation on a case-by-case basis.

Where only a variant or variants of a noun or a verb were present in the corpus but not the root form, one of the variants was selected to bear the weight. Assumptions of equal or similar meanings were avoided. Neither were rare usages grouped into superordinate categories in order to achieve a match. For example, the mention of a particular aircraft model, such as the C152, was not combined into a more general aircraft type, e.g., Cessna.

The resultant word set may be described as a non-tagged corpus rather than a tagged or a labeled corpus. That is, no grammatical tags or morpho-syntactic descriptions were attached to the word tokens, such as those used by various corpus encoding Text Encoding Initiative conformant software. Tagging would have preserved the grammatical structure of the de-contextualized corpus content by identifying the specific context source of each word token, thus making syntactic comparisons possible between different inputs. However, the concept map content which drove the development of this corpus was not necessarily expressed in a developed syntactic form, making cross-document syntactic analysis extraneous for this research.

Whenever words were combined into word stems, the occurrences (counts) of those words were added, and their associated weights were also added arithmetically. This involved the addition of those corpus word weights which had resulted from the Internal Weighting Index and External Weighting Index procedure. An overall summed weight was produced for each word stem based on the total number of word tokens attaching to that particular word-stem. Grouping on word stems and related forms did not affect the overall total weights.

Domain Specific Terms

With the exception of a small set of words developed specifically to describe aviation phenomena, or borrowed from non-English languages and the use of which is almost exclusively restricted to aviation (e.g., ailerons), it was assumed that there are few unambiguously aviation words, and that aviation terms are identified through their appearance in an aviation context. Even in an aviation context, a word may be used as a non-aviation homograph. For these reasons, there was no *a priori* categorization of aviation and non-aviation specific words.

Summary

In any domain, some content will be more central or salient than another. Through the provision of a weighted word frequency scale, the use of more domain central words can be highlighted. It is also reasonable to assume that novices and experienced individuals in a specific field may use either different domain vocabularies, that is, different words, or different relative distributions of the same words. Systematic and significant differences in word uses can be identified through a quantified scale drawn from an appropriately constructed word collection.

The corpus described in this paper was constructed on the highly restricted theme of the Visual Landing Approach. By confining the topic so narrowly, and incorporating a range of constituent texts, a relatively small but useful word set was created. Clearly a corpus of this size is unable to fulfill the same functions as do large word sets such as the British National Corpus. However, when a restricted research topic such as the Visual Landing Approach can be similarly isolated in other fields, and texts are available for incorporation, the construction of a practical purpose-built corpus arguably is within the technical ability of most researchers.

REFERENCES

Altheide, D. L. (1996). *Qualitative Media Analysis*. Thousand Oaks CA: Sage.

Baayen, H. (1993). Statistical models for word frequency distributions: A linguistic evaluation. *Computers and the Humanities, 26*, 347-363.

Baker-Brown, G., Ballard, E. J., Bluck, S., de Vries, B., Suedfeld, P., & Tetlock, P. (1992). The conceptual/integrative complexity scoring manual. In C. P. Smith (Ed.), *Motivation and Personality: Handbook of thematic content analysis* (pp. 401-418). Cambridge: Cambridge University Press.

Balasubrahmanyan, V., & Naranan, S. (1996). Qualitative linguistics and complex system studies. *Journal of Qualitative Linguistics, 3*(3), 177-228.

Beardsworth, C. (1980). Analysing press content. In H. Christian (Ed.), *The Sociology of Journalism and the Press* (pp. 371-395). Staffordshire: University of Keele.

Berry-Rogghe, G. (1974). The computation of collocations and their relevance in lexical studies. In A. Aitken, R. Bailey & N. Hamilton-Smith (Eds.), *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge England: Cambridge University Press.

Biber, D. (1993). The multi-dimensional approach to linguistic analysis of genre variation: An overview of methodology and findings. *Computers and the Humanities, 26*, 331-345.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods, Instruments, & Computers, 30*(2), 272-277.

Burrows, J. F. (1987b). Word patterns and story shapes: The statistical analysis of narrative style. *Journal of the Association for Literary and Linguistic Computing, 2*(2), 61-70.

Caudery, T. (1992). Review of Frequency Analysis of English Vocabulary and Grammar Based on the LOB Corpus in two volumes by Stig Johansson and Knut Hofland (1989). *English Studies, 73*(3), 276-279.

Chafe, W. (1986). Writing in the perspective of speaking. In C. Cooper & S. Greenbaum (Eds.), *Writing in the perspective of speaking* (pp. 12-39). Beverly Hills, CA: Sage.

Chen, Y. (1991). Zipf's law in natural languages, programming languages, and command languages: the Simon-Yule approach. *International Journal of Systems Science, 22*(11), 2299-2312.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Collins, P., & Blair, D. (Eds.). (1989). *Australian English: the Language of a New Society*. St. Lucia: The University of Queensland Press.

Cornelius, S. (1989). A comparison of magazine articles from "The Bulletin" and "Newsweek" magazines based on readers' perceptions of differences between American and Australian English. *Working Papers of the Speech, Hearing and Language Research Centre*, 71-110.

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *Forensic Linguistics, 1*(1), 27-43.

Davis, M., Dunning, T., & Ogden, B. (1995). *Text alignment in the real world: Improving alignments of noisy translations using common lexical features, string matching strategies and N-gram comparisons*.: European Association for Computation Linguistics.

de Sola Pool, I. (1959). Trends in content analysis today: A summary. In I. de Sola Pool (Ed.), *Trends in Content Analysis* (pp. 189-233). Urbana IL: University of Illinois Press.

Delcourt, C. (1992). About the statistical analysis of co-occurrence. *Computers and the Humanities, 26*, 21-29.

Francis, W. N., & Kucera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston MA: Houghton Mifflin.

Halliday, M. (1994). *An Introduction to Functional Grammar* (2nd ed.). London: Arnold.

Haskel, P. (1971). Collocations as a measure of stylistic variety. In R. Wisbey (Ed.), *The Computer in Literary and Linguistic Research* (pp. 159-168). Cambridge: Cambridge University Press.

Hayes, D. (1988). Speaking and writing: Distinct patterns of word choice. *Journal of Memory and Language, 27*(5), 572-585.

Herdan, G. (1964). *Quantitative Linguistics*. London: Butterworths.

Hickey, R. (1994). Applications of software in the compilation of corpora. In M. Kyto, M. Rissanen & S. Wright (Eds.), *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora* (pp. 165-186). Atlanta GA: Rodopi.

Hofland, K., & Johansson, S. (1982). *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities.

Holmes, D. (1994). Authorship attribution. *Computers and the Humanities, 28*(2), 87-106.

Ide, N. (1989). A statistical measure of theme and structure. *Computers and the Humanities, 23*, 277-283.

Ide, N., & Walker, D. (1993). Introduction: Common methodologies in humanities computing and computational linguistics. *Computers and the Humanities, 26*, 327-330.

Iker, H. (1974). An historical note on the use of word-frequency contiguities in content analysis. *Computers and the Humanities, 8*, 93-98.

Johansson, S. (1985). Word frequency and text type: Some observations based on the LOB corpus of British English texts. *Computers and the Humanities, 19*(1), 23-36.
Kennedy, G. (1992). Preferred ways of putting things. In J. Svartvik (Ed.), *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter.
Kilgarriff, A., & Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities, 34*, 15-48.
Landini, G. (2000). *Zipf's laws in the Voynich Manuscript*. Available: http://web.bham.ac.uk/G.Landini/evmt/zipf.htm [2000, 29 August].
Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory, 36*(6), 1842-1845.
Liiv, H. (1997). A method for singling out representative linguistic features. *Journal of Qualitative Linguistics, 4*(1-3), 139-142.
Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge MA: MIT Press.
McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University.
McKechnie, J. (Ed.). (1987). *Webster's New Universal Unabridged Dictionary* (2nd ed.). New York: Simon & Schuster.
Ramson, W. (1966). *Australian English: An historical study of the vocabulary, 1788-1898*. Canberra: Australian National University Press.
Ridley, D. R., & Gonzales, E. A. (1994). Zipf's law extended to small samples of adult speech. *Perceptual and Motor Skills, 79*, 53-154.
Roller, E., Mathes, R., & Eckert, T. (1995). Hermeneutic-classificatory Content Analysis. In U. Kelle (Ed.), *Computer-aided Qualitative Data Analysis* (pp. 167-176). London: Sage.
Sampson, G. (1987). Evidence against the "grammatical"/"ungrammatical" distinction. In W. Meijs (Ed.), *Corpus Linguistics and Beyond*. Amsterdam: Rodopi.
Saporta, S., & Sebeok, T. A. (1959). Linguistics and content analysis. In I. de Sola Pool (Ed.), *Trends in Content Analysis* (pp. 131-150). Urbana IL: University of Illinois Press.
Sichel, H. (1986). Word frequency disributions and type-token characteristics. *Mathematical Scientist, 11*, 45-72.
Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Solomon, G. (1990). Psychology of novice and expert wine talk. *American Journal of Psychology, 103*(4), 495-515.
Taylor, B. (1989). American, British and other foreign influences on Australian English since World War II. In P. Collins & D. Blair (Eds.), *Australian English: the Language of a New Society*. St. Lucia: University of Queensland Press.
Toglia, M. P., Battig, W. F. *et al*. (1978). *Handbook of Semantic Word Norms*. Hillsdale NJ: Lawrence Erlbaum Associates.
Trask, R. L. (1993). *A Dictionary of Grammatical Terms in Linguistics*. London: Routledge.
Turner, G. R. (1997, August 9 1997). *Relationship between vocabulary, text length and Zipf's law*. Available: http://www.btinternet.com/~g.r.turner/ZipfDoc.htm [1999, December 19].
Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. New York: Hafner.

Dr. Lisa Thomas BA (Franklin & Marshall); MAR MDiv (Lancaster Theological Seminary); BSc (Aviation) (Hon 1st Class) PhD (Newcastle) is currently researching developmental changes in pilots.

Associate Professor H. Peter Pfister PhD (Newcastle) has particular research interests in aviation psychology and human factors.

Associate Professor Peter Peterson PhD (Newcastle) has ongoing research interests in English syntax, second language acquisition, and syntactic theory.