HAMLET – a Multidimensional Scaling approach to Textual Analysis

User notes for Hamlet II

for Windows ©

Covers version for Windows™ NT/2000/XP
As amended: 4 January, 2006
Polonius: "What do you read, my lord?"

Hamlet: "Words, words, words."

Hamlet (II,ii,194)

These notes describe the latest version of the HAMLET programs for computer-assisted text analysis for use with Microsoft Windows NT/2000/XP. They still apply to some extent to the earlier versions for Windows 3.xx/95/98 and for MS-DOS. The corresponding on-line help files cover any differences in using these earlier versions.

I am grateful for the facilities provided for work on earlier versions by the University Computing Service, Southampton and the Hochschulrechenzentrum, Johann Wolfgang Goethe-Universität, Frankfurt am Main, and for past support from the Department of Politics, University of Southampton, UK. Visits to Dr.h.c.Ekkehard Mochmann and Bruno Hopp at the European Data Laboratory (EUROLAB) at the Zentralarchiv für empirische Sozialforschung in Cologne, Germany, provided an invaluable stimulus for extensions and improvements in the version described in these notes.

Special thanks are due to James Brier for his contribution to handling different language conventions, for his help in adapting the multidimensional scaling routines for PC use, and his continued invaluable advice throughout the development of the versions of HAMLET documented here.

Alan Brier

Southampton, UK
and Cologne, Germany
4 January, 2006
A Multidimensional Scaling approach to Textual Analysis

User notes for Hamlet II for Windows ©

---

Contents

1. General Principles
   4

2. Analysis of joint frequencies
   2.1 The vocabulary list
   2.2 Context units
   2.3 Text conventions
   2.4 Using HAMLET – counting joint frequencies
   2.5 Cluster Analysis
   2.5.1 Hierarchical clustering
   2.5.2 Non-hierarchical clustering
   2.6 Correspondence Analysis
   2.6.1 Multidimensional scaling of the matrix of joint frequencies
   2.7 Example
   5

3. Multidimensional Scaling
   3.1 MINISSA
   3.2 INDSCAL
   3.3 Selecting configurations from different texts for detailed comparison using
       Procrustean Individual Differences Scaling (PINDIS)
   3.4 PINDIS
   3.5 Example of using PINDIS
   22

4. Correspondence Analysis
   40

5. Basic utilities for looking at word usage
   5.1 Wordlist
   5.2 Comparing wordlists
   5.3 Key_Word_In_Context lists
   5.4 Text profile – lengths of words and sentences
   43

I. References and further reading
   49

II. Adapting to additional language conventions
   51

III. Technical details
   54
1. General Principles

On starting **HAMLET II for Windows**, the following main control and editor window is opened, from which all other procedures are called. Besides accessing the online help file, the **Help** menu also shows details of how to register your copy of **HAMLET II**.

![HAMLET II for Windows window](image)

The main analytical procedure, **HAMLET Joint Frequencies**, is intended to be used wherever there are good grounds for searching for inter-connections between a number of key words, or more generally, character strings, known to occur in the text. It produces matrices of joint frequencies of the items of a specified vocabulary list with respect to a suitably chosen unit of context. The vocabulary list applied can consist simply of individual words, but as each main entry can be associated with a larger number of ‘synonyms’ or related items, the list can be developed into a relatively complex, ‘dictionary’-like structured list of categories, where the main items can be category names, of theoretical significance for the investigation but not necessarily occurring directly in the text.

After calculating a matrix of joint frequencies of all pairs of main entries in the vocabulary list, **HAMLET II** offers as options a simple cluster analysis procedure and a correspondence analysis to explore word and category associations. If this suggests the existence of a structure which may be of interest, the non-metric multi-dimensional scaling procedure **MINISSA** can be applied to the matrix of joint word frequencies derived from an individual text. Configurations of word usage or co-occurrence of categories derived by applying the same vocabulary list to a number of different texts, which are thought to be related, can then be compared using **Individual Differences Scaling (INDSCAL)** and **Procrustean Individual Differences Scaling (PINDIS)**. The additional tools **KWIC**, **WORDLIST**, **COMPARE** and **PROFILE** may first be used to help to determine the broad characteristics of word usage in the text(s) of interest: **KWIC** offers Key-Word-In-Context listings for a given word or phrase.
It helps to have made an initial qualitative analysis of the domain to be covered, or to have an initial hypothesis about the terms of the discourse to be studied. It is also as well to begin with a limited and manageable set of main terms or categories to employ, and gradually refine or extend it in practice as seems appropriate. If you are altogether uncertain which words to specify as a vocabulary list, you can choose WORDLIST or the HAMLET II Vocabulary List Editor with its capacity to build a search list incrementally, to list words occurring in a given text. The point must be to try to include non-trivial terms which occur most frequently or which appear to be most distinctive for the content. COMPARE lists words common to pairs of texts, and is also useful in generating lists including synonyms for use in comparing a number of texts. PROFILE displays the distributions of word and sentence lengths. These are intended as simple free-standing tools for the exploration of word usage which can be applied to any text file with a minimum of fuss and can help to generate vocabulary lists to be applied in analyses of joint occurrences with HAMLET Joint Frequencies. At this stage, it is helpful also to think about the steps likely to be needed towards lemmatisation and disambiguation, within the scope of the present software.

The main HAMLET II interface provides a means of viewing, editing and annotating, the various files, including graphics, used or saved in the course of working with HAMLET II for Windows.

The texts in question should, for convenience, exist already in files stored in the Windows ANSI character set. If any files were created under MS-DOS, use the CONVERT utility first to convert them to Windows characters. To avoid any possible misunderstanding, it is strongly advised that it should be stored in plain text format, but the program may also be applied successfully to files in the format of widely-used word-processing programs, which should also be used to convert files to plain text format for analysis. The checking of a sample of text is nevertheless always advisable to be sure that special characters and hidden commands used by word-processing packages do not lead to unexpected results. This applies particularly to the generation of vocabulary lists and profiles of word and sentence lengths which can be strictly accurate only for plain text input.

Texts may be read in any of the languages recognised by the currently installed version of Windows. It is possible to amend the file HAMLET.WIN to take account of any special characters and lexical conventions applying to the language of the text to be considered should the default version not provide appropriate listings. Details of how to do this can be found at the end of the present document.

2. Analysis of Joint Frequencies

HAMLET Joint Frequencies generates basic statistics for individual and joint word frequencies and the corresponding frequencies expressed in a chosen unit of context.

Individual word frequencies \( f_i \) are counted together with joint frequencies \( f_{ij} \) for all possible pairs of words, and the corresponding standardised joint frequencies are calculated, by default

\[
\text{s}_{ij} = \frac{f_{ij}}{(f_i + f_j - f_{ij})},
\]
where \( f_{ij} \) and \( f_i, f_j \) refer respectively to joint and individual frequencies of words \( i \) and \( j \) in a given vocabulary list, expressed in units of context in each case.

This treats joint non-occurrences as irrelevant, which seems to be a suitable procedure in most textual analysis. It is also important to realise that it is indifferent to the order in which the words in each pair occur, and depends for its values on a sensible choice of context-unit being made in reading the text.

As a coefficient of similarity, this is known generally as Jaccard's coefficient for dichotomous data, which excludes consideration of occasions when neither property, in this case one of a pair of words, is present, and represents the probability of both of a pair of attributes being present in any pair of objects, when only those objects exhibiting one or the other are considered.

It has an expected value of

\[
E(s_{ij}) = \frac{f_i \cdot f_j}{t(f_i + f_j) - f_i \cdot f_j},
\]

since the expected value of \( f_{ij} \) is \( \frac{f_i \cdot f_j}{t} \),

where \( t \) is the total number of context-units counted in the text, and \( f_i, f_j \) and \( f_{ij} \) as before are the individual and joint frequencies of the words \( i \) and \( j \), expressed in context-units.

As an alternative, and for purposes of comparison, it is possible to employ instead Sokal's matching coefficient, in which the number of joint non-occurrences is included in the numerator and denominator of the calculation. In the terms already outlined, this coefficient is

\[
c_{ij} = \frac{(f_{ij} + t - (f_i + f_j - f_{ij}))}{t}
\]

and the term to be added to the numerator and denominator is \( t - (f_i + f_j - f_{ij}) \).

[The reader is referred to Coxon, (1982), chapter 2, Everitt and Rabe-Hesketh (1996), chapter 2, or to Sokal and Sneath (1963), for general treatments of measures of similarity between dichotomous variables.]

**Raw and standardised joint frequencies**, according to the coefficient selected, are displayed in lower-triangular format, suitably labelled with the corresponding vocabulary list entries. Either matrix can be regarded as a set of similarity measures between pairs of words, and can be saved for submission to further analyses using clustering or dimensional methods, to reveal characteristic word clusters or associations of symbols in the original text, as well as for comparison with the results of applying the same vocabulary list to other texts.

**HAMLET Joint Frequencies** optionally offers an exploratory single-linkage cluster analysis, and a correspondence analysis of the profiles of the context units selected, which can help to determine whether further analysis of the similarities matrix, for example by the multidimensional scaling techniques also offered by the program, is likely to be useful. The results of the cluster analysis can be displayed on the screen as a dendrogram, and the contents of clusters listed for selected levels of minimum similarity (\( 0.0 \leq s \leq 1.0 \)). The results of the
cluster analysis, if the data are suitable, are also displayed graphically in the first instance. Although clustering and dimensional analysis are not equivalent procedures, the absence of evidence of any substantial clustering at this stage normally indicates that there would be little point in continuing with a full dimensional scaling procedure. Non-metric scaling seems, for both theoretical and practical reasons, to be the best, or at least the most easily comprehensible, dimensional reduction procedure to apply to matrices of similarities based on joint occurrences of words in texts, although factor analysis, latent structure analysis and network analysis have also been applied in this context.

2.1 The vocabulary list

The vocabulary list to be used by HAMLET joint frequencies clearly has to be created in the first instance, and can then be saved in a named file for subsequent editing and re-use. Vocabulary lists can also be created incrementally, from the contents of sample texts intended for the analysis, and later edited directly, by calling the vocabulary editor from the main control window.

The terms supplied to the vocabulary or search list can contain the ‘wild card’ characters ‘@’ and ‘*’, with the following effect: when comparing words in the text with the vocabulary list, individual letters corresponding to the position of the character ‘@’, and all characters after, and including, the position of the character ‘*’, will be ignored. This, for example, provides a way of over-riding known transcription errors in the text, and of treating as equivalent words which differ only in their suffixes, that is, of basic lemmatisation. Disambiguation is more difficult, and it may be necessary to edit the source text to avoid serious errors in interpretation. Care is needed, however, that use of ‘wild-card’ characters does not create logical equivalences between two or more entries in the vocabulary list, which also may confuse the searching process.

In entering words to the vocabulary list, and in searching the text file, upper- and lower-case letters may be separately regarded or treated as equivalent. The latter option, of course, will normally regard words beginning sentences as different from the same words occurring later. Such words will have to be explicitly and separately specified in the vocabulary list if they are not to be missed. Hence the importance of knowing the basic vocabulary of the text before considering the application of HAMLET II. Care is needed here too, as an inadvertent choice e.g. of case-sensitive searching when the search list has been specified as case-insensitive can lead to target words which appear to be clearly present in the input file being reported as not being found! Whichever option is chosen remains in force for a given vocabulary list, including one which has previously been saved in a file for re-use, until the option applying to that list is explicitly changed using the editing option.
To simplify searching for groups of words of equivalent meaning, each main entry to a vocabulary list may have an accompanying list of words which will be counted as if they were its “synonyms”. These need, of course, not be literal synonyms, but can be any word strings found in the text which can conveniently be grouped together when calculating joint frequencies. In this way HAMLET II allows a vocabulary list to be developed into a kind of dictionary of categories by assigning to each of the main entries, as necessary, a set of related words to be counted as its equivalents. It is important to note that, unlike some other programs, for example TEXTPACK, HAMLET II requires particular words to be assigned logically to only one main entry or category at a time.

Main entries and their associated synonyms or related words can be saved for re-use, and edited as necessary to modify, delete or add new main entries or associated synonyms.

2.2 Context units

Several possibilities are offered for the definition of context-units when reading a text. These must be used with some care, to ensure that the units chosen are indeed capable of meaningful interpretation, and that they are not so large that almost all of the target words occur together in each unit, losing any discrimination in the analysis. In each case, joint occurrences are counted irrespective of the order in which the words in each pair actually occur.

- **Variable-length contexts**

are defined by the inclusion of a special character in the text to denote the end of each unit of context to be read. This may be a character (not normally used for punctuation, etc.) inserted in pre-editing as to mark the end of each context-unit as appropriate to the sense of the particular text, or it may be possible to take advantage of some relevant aspect of the orthography which can be used for this purpose. Some experimentation may be needed, to determine a stable and effective means of achieving the delimitation of context units which are meaningful for the text(s) considered.

- **Fixed-length contexts**

were used in the first work of this kind, and are specified as a kind of "sampling unit" consisting of a fixed number of words. The text is treated as a series of "blocks" of this fixed length, within each of which joint occurrences of words in the specified vocabulary list are counted. Careful consideration needs to be given to the choice of context length to be used, since different features are likely to emerge from a shorter context unit, more appropriate to analysis of phrases or sentences, than from a coarser level of context selection. It may, on the other hand, be possible to determine an appropriate context unit for a given text and vocabulary empirically, by varying the unit specified in a series of runs of HAMLET, and observing the effects on the numbers of co-occurrences counted.

- **Sentences**, with punctuation as specified by the user, may be chosen as the context unit; or

- the **Collocation** option counts joint occurrences within a given number, or span, of words, up to a maximum of 120. This will be generally slower in operation than the other context options, and is more suitable for smaller bodies of text or when specific word usage is of particular interest. There may, however, be reasons for preferring this procedure to the “classic” application of fixed-length context units, as described above, even in looking at large bodies of text. Some other “data
mining” procedures employ a similar technique to identify words occurring close to one another in a purely empirical way, without first considering the way in which these may be structured. This is not particularly sensible, as it can rapidly lead to confusion due to the sheer number of different collocations identified if the text(s) to be read are of any size or elaboration. Not that, if choosing collocations, the number of units of context displayed and the number of actual occurrences are recorded in HAMLET II as identical.

2.3 Text Conventions

HAMLET II distinguishes separate words in the input as continuous strings of letters, separated by punctuation, spaces, or the end of a line, unless there is a continuation character. The program contains extensive default definitions of recognised sets of letters and punctuation suitable for most purposes and most European languages. It should, however, be noted that it may be confused by numbers containing decimal points or commas into increasing the word count on each group of numerals: e.g. 60,000 may be read as two words, ‘60’ and ‘000’. Normally this is of no great consequence, but is mentioned here to emphasise the importance of prior knowledge of the nature of a file of text, and the need for pre-editing in cases where the above features are considered to be of real significance to the meaning of the text.

If any word in the original text requires continuation from one line of input to the next, a hyphen (‘-’) should always be used as the last character of the line to be continued, to indicate that the characters from the beginning of the next line form part of the current word. Otherwise, the end of a line (a normally invisible character) will automatically be regarded as marking the end of the current word on input, with the possibility that some words may become inadvertently divided.

Special characters

It may be that the text contains characters which should not be considered as part of the text itself but do not normally occur in punctuation. Such characters (e.g. ‘#’, ‘~’, ‘¿’) sometimes occur systematically in source files but should be ignored when looking for words in HAMLET II. Several different characters may even optionally be used in preparatory editing of the text, for example to delimit major text components, but must be chosen to serve this purpose alone, since they must not be allowed to become confused with ordinary text and punctuation. Characters used in this way can be declared to be ignored in the appropriate box in the options window (see below, p. 11). They will then be skipped in looking for matching words.
2.4 Using *HAMLET* - joint frequencies

In the following window:

- at **Text file name**: enter the name of the text file to be read, or click on the **Files** menu and select **Open text file** to search for it. The **Files** menu also allows you to **Reopen** directly files used recently to save having to look for them. Alternatively, use the corresponding speed button, or the usual Windows shortcut key combinations, e.g. **Alt** together with the letter keys indicated in menus and other controls, and **F1** to display context-sensitive help. If the text to be read is in a language other than English, use the pull-down menu on the right of the control panel to select the appropriate lexicographic conventions to apply.

- click on the **Vocabulary** menu, and select **Create new vocabulary list** or **Open existing vocabulary file** to specify the vocabulary items for which to search.

- When creating a **new list**, setting optional maximum sizes for words (<=30 letters) and the number of words (<=100) in the search list helps to reduce unnecessary searching when the program is run:
Clicking **Continue** opens the vocabulary list editor to create the new list:

The items entered may each contain ‘wildcard’ characters (‘*’, '@' – see above p. 6) as required. The first column to the right of the file name list is for the main items or category headings; clicking on these in turn enables you to enter any synonyms or related items for these main entries in the next column to the right.
To assist in this process, clicking on Browse will open a list of words in a selected text file, displayed by frequency and alphabetically in a window to the right. You can select and copy words from this window into the main panel using the mouse. With the left mouse button: clicking on 'Words >' reverses the order of the items listed, and clicking on 'Freq.' causes the columns to swap positions. Clicking anywhere in the vocabulary list with the right mouse button switches the listing between alphabetical and frequency order.

- When opening an existing vocabulary file, the same editor may be used, as required, to make any necessary changes:

![Vocabulary list](image)

Note that 'words' supplied to these lists can be any relevant string of characters, and may contain the 'wild card' characters '@' and '*', with the following effect: When comparing words in the text with the vocabulary list, individual letters corresponding to the position of the character '@', and all characters after, and including, the position of the character '*', will be ignored. This, for example, provides a way of treating words as equivalent which differ only in their suffixes. Care is needed to ensure that the use of 'wild card' characters does not create a logical equivalence between two, or more, vocabulary list entries, which may cause confusion in the searching process. Words may also consist of significant pairs of words, such as 'United Nations' or 'Prime Minister', although care is required to avoid also specifying the same words individually in the search list, which can lead to confusion.

Upper- and lower-case letters may be regarded separately, or treated as equivalent. The latter option, of course, will normally regard words beginning sentences as different from the same words occurring later. Such words will have to be explicitly and separately specified in the vocabulary list if they are not to be missed. Hence the importance of knowing the basic vocabulary of the text before considering the use of HAMLET joint frequencies. Check the box in the options window if searching for words is to be case-sensitive. (Take care here: an inadvertent use of case-sensitive searching when the search list has been specified without
regard to case can lead to unexpected results!)

The present version of HAMLET II allows you to enter up to 100 main vocabulary items, and a total of 500 associated entries. This allows advanced users to create quite elaborate structured search lists or 'dictionaries', where the main entries may be analytical categories and the associated entries for each category are the strings to be assigned to them when searching a number of different texts.

- When finished, save the vocabulary file if it has been altered and close the editor to return to the main HAMLET II window.

- Back in the main HAMLET II window, click on the Options menu to specify the options for the current analysis:

![Options for HAMLET analysis](image)

Select the kind of context unit to be used: for fixed contexts enter the number of words, enter a character to delimit variable contexts, or the span of words (≤ 120) to use for collocations. This last option counts the joint occurrences of pairs of words within a given number of words of each other. If the sentence option is selected, it is possible to amend the punctuation characters if the text has been differently punctuated.

It is also possible to change the coefficient of similarity to be applied to the raw frequency counts. The Sokal coefficient takes account of joint non-occurrences, where this is thought to be
appropriate. The Jaccard coefficient, applied by default, is, however, suitable for most purposes (see above pp. 5-6).

If the text file contains **characters which should be ignored** when making comparisons in **HAMLET**, enter these in the edit box indicated. Characters used in this way must be chosen to serve this purpose alone, since they must not be confused with normal text and punctuation.

Check the box shown if searching for words is to be case-sensitive. If this option is chosen, words in the vocabulary list must also be entered with regard to upper- and lower-case letters if they are not to be missed.

- When finished, click **Continue** to confirm the options indicated, or **Cancel** to ignore them, and return to the main **HAMLET** window.

- Click on **Count joint occurrences** to start the search process. A progress indicator appears while searching takes place. Click on the **Cancel** button by this indicator to stop searching at any time. Patience may be needed when searching large files, or using a large search list, depending on the computing resources you have available. If progress appears very slow, check for continuing disk activity before concluding that there is an error.

- When the process is finished, full results are temporarily displayed in an edit window, which enables you to annotate them as necessary before saving them for further reference.

- Click on the **Files** menu and select **Save results file**, or use the corresponding speed button, to save the output in a named file.

- Click on the **Files** menu and select **Print results**, or use the corresponding speed button, to print the current output file.

**2.5 Cluster Analysis**

**2.5.1 Hierarchical Cluster Analysis**

**HAMLET joint frequencies** automatically offers an optional hierarchical cluster analysis routine, which can help to determine whether further analysis of the similarities matrix, for example by the multidimensional scaling technique offered, will be useful. When the matrix has been saved separately, the same procedure can be applied directly from the main **HAMLET II** window.
• When the panel above appears, toggle the button **Diameter/Connectedness** to select the clustering method to be applied to the matrix of similarities. The alternative methods of hierarchical clustering offered here use different criteria in assigning the individual vocabulary entries to clusters: the "connectedness" or "single link" method looks for the greatest similarity between an unassigned item and those contained in existing clusters; the "diameter" or "complete linkage" method defines the similarity between groups as the similarity between their least similar pair of individual items.

• Click **Display ...** in the window above to see the results of the clustering method selected on the screen as a dendrogram, which may be saved as a bitmap file for later use by clicking on the **Save** button at the bottom of the display:-

![Dendrogram](image)

or

• enter a minimum similarity value (0.0 ≤ s ≤ 1.0) in the box shown on p.16 and click **Show clusters ...** to list the word clusters with this minimum similarity :-
Clicking on **Save clusters** allows the clusters shown to be saved in a separate text file.

Clicking on **Print clusters** allows you to print the results.

The results of cluster analysis are frequently used as an interpretative aid in examining configurations of points resulting from **multi-dimensional scaling**. Where well-defined compact “spherical” clusters exist, these methods tend to produce similar results. The “single link” method, on the other hand, is useful for displaying absence of any distinct structure.

For the purposes of text analysis, although hierarchical clustering and dimensional analysis are not equivalent procedures, the absence of any substantial clustering normally indicates that there would be little point in continuing with a full dimensional scaling procedure. In particular, further analysis should be avoided where any items are joined to the dendrogram only at the highest level, indicating that they have no connection with any other item in the analysis.

**2.5.2 Non-hierarchical Clustering**

When a matrix of similarities has been saved separately, an alternative non-hierarchical clustering procedure is also available from the menu item **Cluster analysis | Non-hierarchical** in the main **HAMLET II** window.

This implements an efficient new algorithm (**BBDIAM**, Brusco, 2003) to partition the matrix into successive numbers of clusters, each containing at least one item. Clusters are mutually exclusive, and exhaustive, in that all items are assigned to a cluster.
It should be noted that, given the number of ties that can occur in minimum diameter partitioning, it is likely that there are many alternative optima in large matrices, but if the results of this procedure are always considered together with those from hierarchical clustering as well as from multi-dimensional scaling, the risk of misinterpretation is considerably reduced.

The matrix of pairwise similarity measures is first converted into dissimilarities. The object is then to develop a partition of the matrix into a given number of clusters, each containing at least one item, where the clusters are mutually exclusive, and exhaustive, in that all items are assigned to a cluster. A commonly-used criterion in clustering is to minimize the within-cluster sum of pairwise dissimilarities, but this has a tendency to produce clusters of approximately the same size, irrespective of the data. For this reason, the enhanced branch and bound algorithm employed by BBDIAM seeks instead to minimize the partition diameter, which is related to Johnson's diameter method for hierarchical clustering. The diameter of a cluster is the maximum pairwise dissimilarity index among objects in that cluster. The partition diameter is defined as the maximum of the cluster diameters. To minimize the diameter of the partition is to minimize the maximum dissimilarity index across all subsets. An advantage of using the partition diameter is that it is not predisposed to produce clusters of particular sizes. It is also computationally simpler than minimizing the within-cluster sum of dissimilarities.

The efficiency of the branch-and-bound algorithm used in minimizing the partition diameter depends on the quality of the upper bound. A good upper bound can frequently be established heuristically using a complete-link clustering algorithm, as in the connectedness method of the hierarchical clustering option outlined above. An algorithm for partitioning then applies two kinds of local-search operations - trial movement of each object from its current subset to each of the other subsets, and pairwise interchange of objects with respect to their subset memberships. These local-search operations are conducted until no further improvement is possible in the upper bound.

Because the resulting solutions may be sensitive to the initial partition, BBDIAM applies a probabilistic process that uses 100 replications and selects linkages using biased sampling. In particular, when considering the next linkage, the algorithm has a 50% chance of selecting the best linkage and a 50% chance of selecting the second best linkage. This biased-sampling version often produces better bounds than the deterministic version.

The non-hierarchical clustering option produces a listing of the partitions generated. For matrices representing up to 15 main entries, the program automatically lists all sets of clusters from 2 to the total number of categories minus 1. For larger matrices, enumeration is restricted to half the matrix size up to a maximum of 20 clusters, and may take some time due to the number of alternative partitions possible.

2.6 Correspondence Analysis of the context unit profiles

As an additional diagnostic tool, or for a representation of the content of a single text, it is also possible to perform Correspondence Analysis of the profiles of the context units containing the frequencies of the main vocabulary entries identified in them. This conveniently represents the row and column categories of the matrix of context unit profiles as points in the same dimensionality.
The configurations generated by the analysis are automatically displayed in graphic form, but because the canonical ("optimal") scores reported are row and column conditional, it is advisable to avoid inter-set point distance interpretation, however tempting this may be!

In correspondence analysis, the matrix of cross-products is first normalized by dividing each row entry by the square root of the product of the corresponding row and column totals, their geometric mean. This removes differences in the marginal totals and expresses each cell as a proportion. If any of these products turns out to be zero, the data are not suitable for correspondence analysis and an error will be reported.

The first eigenvalue of the transformed matrix of cross-products is always 1.0. It corresponds to the independence model of chi-squared expected values, and is ignored in the subsequent analysis. **It is important, however, to check that the eigenvalues remaining after are in fact large enough to justify continuing with the analysis.** Reference should be made to the chi-squared contributions of each dimension of "inertia" reported, and to the overall chi-squared value for the analysis. These values can be inspected by clicking the menu item Data | View/Edit data on the graphic display of the correspondence analysis results.

The configuration of points representing the main vocabulary entries will be broadly similar to that produced by multidimensional scaling (see below under MINISSA). The relative locations of points representing the various context units will indicate how the various categories of the search list are distributed throughout the text. (See below pp.39-42 for a more detailed description of this utility which has been introduced in HAMLET II).

**2.7 Multidimensional scaling (MINISSA)** is offered whenever HAMLET joint frequencies has generated a matrix of joint frequencies and is intended as the main analytical tool, since its results can be used to compare a number of different texts to which the same search list can be applied (see below pp.22-34 for a more detailed description of the multidimensional scaling utilities included in HAMLET II).

The resulting three-dimensional configuration is automatically displayed in graphic form. The first two reference axes appear as a horizontal plane on which labelled points are projected corresponding to words in the text file originally subject to analysis.

The configuration can be rotated and zoomed for closer examination, annotated, and saved for inclusion in other documents.

It is possible to examine and edit the file containing the scaling results on which the display is based by selecting the menu item Data | View/Edit data in the graphical display window.

This is illustrated in the following example.

**2.8 An example to illustrate the simple use of HAMLET joint frequencies**

The following output listing was obtained by searching a file describing the HAMLET package for the following vocabulary list :-

The text is read from the file: C:\Program Files\HAMLET II\HAMLET.txt

Counting Joint Frequencies -

WORD-SEARCHING IS INSENSITIVE TO CASE.

WORD COUNTS

<table>
<thead>
<tr>
<th>WORD</th>
<th>FREQUENCY</th>
<th>% VOCAB.</th>
<th>% TEXT</th>
<th>f[i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>context*</td>
<td>39</td>
<td>11.02</td>
<td>0.97</td>
<td>14</td>
</tr>
</tbody>
</table>

---
4011 words were read from the text file. 354 of these were in the search list, and 34 context-units were counted.

JOINT FREQUENCIES ......................................
for a FIXED CONTEXT LENGTH of 120 words:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>context*</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimension*</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>frequenc*</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hamlet</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>joint</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>12</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MINISSA</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scaling</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text*</td>
<td>8</td>
<td>13</td>
<td>8</td>
<td>11</td>
<td>19</td>
<td>13</td>
<td>9</td>
<td>9</td>
<td>23</td>
</tr>
<tr>
<td>vocabulary</td>
<td>9</td>
<td>8</td>
<td>4</td>
<td>8</td>
<td>11</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>word*</td>
<td>10</td>
<td>13</td>
<td>9</td>
<td>12</td>
<td>17</td>
<td>14</td>
<td>9</td>
<td>9</td>
<td>23</td>
</tr>
</tbody>
</table>

STANDARDISED JOINT INDEX VALUES ........................
Jaccard coefficient - ignores joint non-occurrence

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>context*</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimension*</td>
<td>2</td>
<td>0.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>frequenc*</td>
<td>3</td>
<td>0.35</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hamlet</td>
<td>4</td>
<td>0.27</td>
<td>0.36</td>
<td>0.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>joint</td>
<td>5</td>
<td>0.45</td>
<td>0.39</td>
<td>0.75</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MINISSA</td>
<td>6</td>
<td>0.33</td>
<td>0.67</td>
<td>0.35</td>
<td>0.36</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scaling</td>
<td>7</td>
<td>0.32</td>
<td>0.75</td>
<td>0.33</td>
<td>0.40</td>
<td>0.37</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>text*</td>
<td>8</td>
<td>0.45</td>
<td>0.27</td>
<td>0.37</td>
<td>0.58</td>
<td>0.43</td>
<td>0.31</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>vocabulary</td>
<td>9</td>
<td>0.33</td>
<td>0.17</td>
<td>0.35</td>
<td>0.35</td>
<td>0.38</td>
<td>0.17</td>
<td>0.16</td>
<td>0.53</td>
</tr>
<tr>
<td>word*</td>
<td>10</td>
<td>0.46</td>
<td>0.32</td>
<td>0.43</td>
<td>0.50</td>
<td>0.50</td>
<td>0.32</td>
<td>0.31</td>
<td>0.72</td>
</tr>
</tbody>
</table>

The resultant standardised matrix of similarities was then submitted to multidimensional scaling.

The following smallest space solution reproduces the main topic areas of the documentation file

Kruskal-Guttman-Lingoes-Roskam smallest space coordinates in three dimensions (weak monotonicity):

<table>
<thead>
<tr>
<th>Dimension</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>context*</td>
<td>-0.085</td>
<td>0.372</td>
<td>0.708</td>
</tr>
<tr>
<td>dimension*</td>
<td>0.846</td>
<td>0.016</td>
<td>0.184</td>
</tr>
<tr>
<td>frequenc*</td>
<td>-0.114</td>
<td>0.500</td>
<td>-0.548</td>
</tr>
<tr>
<td>hamlet</td>
<td>-0.103</td>
<td>-0.662</td>
<td>-0.317</td>
</tr>
<tr>
<td>joint</td>
<td>-0.096</td>
<td>0.477</td>
<td>-0.296</td>
</tr>
<tr>
<td>MINISSA</td>
<td>0.876</td>
<td>-0.086</td>
<td>-0.036</td>
</tr>
<tr>
<td>scaling</td>
<td>0.874</td>
<td>-0.143</td>
<td>-0.007</td>
</tr>
<tr>
<td>text*</td>
<td>-0.532</td>
<td>-0.361</td>
<td>0.132</td>
</tr>
</tbody>
</table>
A pseudo-3-dimensional graphic display of these results appears automatically, as follows. Note that, for easier viewing, these configurations are always rescaled according to their largest absolute co-ordinate value.

The plotted graphic may be rotated, dragged within the display window, and zoomed to be able to concentrate on particular areas. Clicking on an individual point will highlight its attached label, which is one of the main vocabulary list items. This is of help in distinguishing points where their labels overlap.

The display can be annotated using the graphic editing buttons shown. It can be separately saved as a graphic image for inclusion in other documents.

Interpretation of MDS solutions may be assisted by reference to hierarchical clustering, as
illustrated above. Smallest space analysis, on the other hand, is generally claimed to produce more easily interpreted geometric solutions in fewer dimensions than metric procedures like factor analysis, as well as being more versatile in detecting ordered structures in the data. For further discussion of the procedures used here, see the references at the end of these notes.

Before settling for any interpretation, however, it is always advisable to check that the MDS solution in question is reliable, in the sense that it represents a good fit to the information contained in the original matrix and is not a so-called ‘degenerate solution’, dependent upon a single extreme value. It is also advisable to check that the solution remains reasonably stable under appropriate variations of the context unit parameters employed to calculate the matrix of joint frequencies on which it is based, and if not, to give careful consideration to the possible reasons for this. In HAMLET II, MINISSA offers the conventional MDS diagnostics, accessed from the Data menu in the above graphic display, to try to ensure that solutions accepted for further analysis, for example, for comparison using INDSCAL or PINDIS, meet the conditions recommended in the authoritative literature of multidimensional scaling.

These powerful multi-dimensional scaling procedures are described in greater detail in the following sections.

3. Multidimensional Scaling

3.1 MINISSA (Michigan–Nijmegen Smallest Space Analysis) takes as its input a file containing a matrix of standardised joint frequencies, as generated by HAMLET. It is possible to repeat scaling for different combinations of items included in the original word list used by HAMLET to generate the matrix.

The resulting three-dimensional configuration will automatically be displayed in graphic form. the first two reference axes appear as a horizontal plane on which labelled points are projected corresponding to words or categories in the text file originally subject to analysis. The configuration can be rotated and zoomed for closer examination, printed or saved for inclusion in other documents.

It is possible to examine or edit the file containing the configuration data on which the display is based using the Data in the graphic display as shown above. Clicking on View/Edit data opens it in a local file editor for this purpose.

Also on the Data menu, there are two additional diagnostic aids in determining the effectiveness of the particular scaling result. The first shows the contributions of the individual points plotted to the final ‘stress’ value reported. This by convention is the measure of the ‘badness of fit’ between the distances between the points in the scaled plot and their original distances represented in the similarity values of the initial matrix. The bar chart of these contributions will reveal any points which it has been particularly difficult to fit using the MINISSA algorithm into the plot of reduced dimensionality displayed. The second diagnostic aid is the Shepard diagram, relating the observed and predicted inter-point distances. A successful scaling is here represented by an even non-linear or linear relationship with little ‘scatter’ evident.

3.1.1 Using MINISSA
At **File name:** enter the name of the matrix file to be read, or click on the **Files** menu and select **Open file** to search for it. The **Files** menu also allows you to **Reopen** files used recently to save having to look for them. Alternatively, use the corresponding speed button or Alt/key combination. The matrix file should have been saved from an earlier run of **HAMLET joint frequencies**. When calling **MINISSA** from within **HAMLET joint frequencies**, a current temporary file containing the scaling results is used automatically.

- When a valid file name is found, or if the scaling option is selected within **HAMLET joint frequencies**, the vocabulary items available are listed as follows:

- It is possible to repeat scaling excluding any of the items included in the current vocabulary list, without repeating the **HAMLET joint frequencies** analysis each time. From the list displayed, simply delete any vocabulary items to be excluded from the scaling process.
You may choose to scale these items in one to three dimensions. Click on **Scale these items** to start the scaling process, or **Cancel** to return to the main MINISSA window.

Once the process has started, as shown by the progress indicator, click on the **Abort** button to abandon scaling at any time.

The resulting configuration is automatically displayed in graphic form. Note that, for ease of viewing, configurations are always rescaled according to their largest co-ordinate value:

![MINISSA display](MINISSA.png)

Click on the **Data** menu item to view or edit the MINISSA output data currently being displayed in detail. For matrices of between 10 and 60 words, this also includes Spence’s approximation for Kruskal’s Stress based on random data with the same configuration (Spence, 1979). The same menu offers, as diagnostic aids, a bar-chart showing the contributions of the individual points to the final stress value, the Shepard diagram corresponding to the results shown, and displays any items which may have been treated as latent categories in the analysis.

Use the control buttons shown to rotate and zoom the configuration displayed. Use the keys indicated to rotate and zoom in and out on the display. Click on the **axis end points** to see the effect of incremental clockwise rotations of the configuration, where appropriate,
with respect to the selected axis. (The numerical keys 1, 2, and 3 have the same result.) Use Configuration to keep track of this process, and save rotated configurations if required.

- Clicking on an individual point will highlight its label,

- Click on menu item Labels to adjust the maximum number of characters of the words displayed as point labels. Click on a point in the display to highlight its label.

- Click on Grid to toggle the internal lines on the horizontal plane on and off as preferred.

- With a mouse button depressed, drag the pointer to a new location and release the mouse button again to move the display around inside the window, to be able to concentrate on particular parts of it.

- Clicking Draw allows you to draw on the display with the mouse, to highlight features of interest. Clicking Line draws straight lines, from a point where a mouse button is depressed to a point where it is lifted. Clicking Ellipse draws an ellipse within a notional rectangle outlined using the mouse in a similar manner. Clicking Text causes a box to appear to enter text. On closing this box, move the mouse to the position required and press a mouse button to add the text to the image displayed. Click on Refresh to clear any lines and annotations added and return to the original image selected. The image as amended must be saved immediately on completion, as the additions will be lost when the display is reoriented.

- Click on Save Display to save the current display as a graphic image file, for inclusion in other documents. The Editor offers limited facilities to draw upon and annotate image files saved by HAMLET procedures.

- From the main MINISSA window, click on the Files menu and select Save scaling results, or use the corresponding speed button, to save the scaling output in a file. Click on the Files menu and select Exit, or use the Alt+F4 key combination, to exit the program. A prompt will appear if any file contents have not been saved.

3.1.1 A note on the treatment of missing values

If MINISSA detects any items in the search list for which HAMLET joint frequencies has found no collocations in the original text, these have to be disregarded in producing the scaling, or what is termed a ‘degenerate solution’ results. However, on completion of the scaling process, an opportunity is nevertheless offered in saving the results, to record them as extreme outliers in an otherwise valid final configuration, for possible later comparison with other results of applying the same vocabulary list to other texts.

It is, for example, not improbable, if a complex, structured list of categories has been developed theoretically for application to a number of text sources, that not all items included in the list will be found to occur explicitly in every text considered. If, however, the user is satisfied that it is justifiable to retain ‘absent’ items of this kind, for example, if they can plausibly be regarded as implicit or ‘latent’, rather than simply missing and disregarded, it is not necessary to restrict an
investigation to cases where all items are explicitly present.

3.2 Individual Differences Scaling (INDSCAL)

When a series of texts have been analysed using comparable search lists, it is possible to compare the resulting co-occurrence matrices using another multidimensional scaling procedure called INDSCAL.

This provides an internal metric analysis of a set of similarity matrices in terms of a weighted distance model, such that each "subject" (in this case, each text source) is regarded as having a set of dimensional weights which systematically "distort" the group space defined by the overall relationships of the search list items. All the text sources are thus to be regarded as sharing a discourse of the same basic dimensions, but applying different weights to them, to produce their observed individual matrices of co-occurrences generated by HAMLET Joint Frequencies.

INDSCAL is an expressly dimensional model and produces a unique orientation of the axes of the group space, in the sense that any rotation will destroy the optimality of the solution and will change the values of the subject weights. Moreover, the distances in the Group Space are weighted Euclidean, whereas those in the private spaces are simple Euclidean. Because of this, it is not legitimate to rotate the axes of a Group Space to a more 'meaningful' orientation, as is commonly done in the basic multidimensional scaling model. As will be seen in the following section, Procrustean Individual Differences Scaling applies a series of transformations of decreasing stringency to the MINISSA configurations generated from the same matrices, in an attempt to bring them into maximal conformity with each other, or with a hypothetical reference configuration. This allows comparison in greater detail than is provided by INDSCAL, which may nevertheless be found useful in exploring the overall relationships between a series of texts.

To use INDSCAL:
first select a series of matrices saved from HAMLET Joint Frequencies, by clicking on their names in the file list window. This will add the names of the files selected to the right-hand window, after checking that the matrices match those already selected in the vocabulary items they contain. (Note that all windows are resizeable.)

Then click on the Run INDSCAL button to execute the procedure. The results will be displayed in graphic form, in three dimensions, as follows:

![INDSCAL graphic](image)

Click on the Next and Previous buttons to switch the display between the Subject Space, showing the text sources, and the Group Space, containing the grouped scaling of the vocabulary items.

Click on Quit or use the Alt+F4 key combination, to close the display. If an input file has been created, or an output listing exists, these can be saved before closing the program.

3.3 Selecting configurations for detailed comparison using PINDIS

SELECT is used to assemble a series of MINISSA configurations for comparison using PINDIS (Procrustean Individual Differences Scaling):

27
Files containing configurations produced by MINISSA can be selected from the list displayed. Their content is shown in the adjacent window. Click on Select configuration to save this in a new or existing input file for comparison using PINDIS. If more than one configuration has been added to an input file, this can optionally be submitted to PINDIS immediately without closing SELECT. The same window can also be used to view the contents of saved PINDIS (.inp) input files.

Alternatively click on Full view to see a full graphic display of the configuration selected:
• Use the keys indicated to rotate and zoom the display.

• Click on the Data menu to see the MINISSA output data being displayed in detail and check the point contributions to stress and Shepard diagram; Save display to save the graphic.

• Clicking on an individual point will highlight its attached label,

• Click on Select configuration to add the configuration displayed to a file to be saved for comparison using PINDIS.

• Close the display and return to the SELECT window to continue.
3.4 Procrustean Individual Differences Scaling (PINDIS)

Where equivalent MDS configurations have been produced for a number of different texts, it is possible to continue the analysis by comparing these results, to produce a picture of the relationships between the various text sources and to examine in detail their similarities and differences.

As already described, SELECT, is used to choose a number of configurations generated by MINISSA for analysis by PINDIS, and automatically offers to call PINDIS to compare the configurations selected. PINDIS can also be called independently and provided with the names of suitable input files already created in this way.

The results of a PINDIS analysis are displayed as a series of graphical windows.

By default, the first of these is the centroid configuration, produced by subjecting the original 'subject' configurations - for the individual texts selected - to a series of transformations which preserve the orderings of the distances between the points corresponding to the words of the original configurations. This represents a kind of median of the individual configurations considered, and acts as a reference point for the examination of the similarities and differences between them. Alternatively, the hypothetical configuration supplied by the user, or, as described below, the configuration adopted by default as the reference for PINDIS comparisons in the event of employing ‘latent categories’ in the input configurations, in the sense outlined earlier, will appear.

The next plot shown is of the ‘subject space’, which shows the relationship of the original configurations (the “subjects”) which are being compared. The plotted coordinates are the optimal normalized dimension weights required to bring the individual subjects into conformity with the reference configuration (i.e. multiplied by the corresponding column sums of squares of the centroid/hypothesis). Unlike the other displays, which, by convention in MDS, are rescaled for ease of viewing according to their largest co-ordinate value, the original values are plotted here to show the relationship between the texts considered more clearly.

It then remains to examine in turn the individual configurations, as transformed and re-scaled by PINDIS, to determine the extent and source(s) of specific departures from the centroid/initial configuration. The individual configurations have been submitted to a succession of decreasingly stringent distortions, the results of which serve to highlight the precise nature of these departures. Although these results may initially appear confusing, they are clearly labelled and will become clearer with experience. The simple example given below, comparing only two configurations containing a limited number of variables, should help to clarify how to read this necessarily extensive information.

After estimating the fit achieved by attaching differing weights to the dimensions, and by combining this with differing dimensional orientations (correlations) – transformations which may already be familiar from comparing results derived by factor analysis - the Lingoes/Borg model goes on to examine the effects of allowing differing weights and orientations (expressed as directional cosines) for the vectors representing the individual words of the vocabulary list in this case. Substantial differences at this level will immediately highlight differences in the contextual relationships of individual words in the texts compared.
(The final set of transformations applied allows both dimensional and vector weights to differ, although it is unlikely that this will lend itself to convincing interpretation in application to data derived from textual analysis.)

3.4.1 Advanced use of PINDIS

- **Using a hypothesis configuration as a starting point**
  As a final stage of analysis, PINDIS also lends itself well to testing goodness of fit of observed configurations to a hypothetical or theoretically determined representation, which may be input to take the place of the centroid which is otherwise determined initially. Care is needed that any hypothetical configuration employed has the same format as the various 'subject' configurations which are to be compared to it, i.e. co-ordinates must always be entered for the same number of vocabulary items ('stimuli') as was used in the original MINISSA configurations, and always for all THREE dimensions requested, even if the values for one or even two of these are all to be zero. In this way, the methods described here can be employed strictly in the development of theoretical models of the structures of associations of ideas as they appear in texts of various kinds.

- **Statistical tests**
  Langeheine (1980) describes tests of significance to permit the evaluation both of single transformations in PINDIS and of improvements in fit between the various transformations. His tables offer criterion values to test the hypothesis that the fit obtained could be generated by purely random configurations.

3.4.2 Using PINDIS

At **File name** : enter the name of a PINDIS input file to be read, or click on the **Files** menu and select **Open file** or **Reopen** to look for it. Alternatively, use the corresponding speed button or **Alt**/key combination.
When a valid file name is entered, you will be prompted to say if you want to use **hypothesis configuration** as the reference for comparison instead of the **centroid**, which is the default option here.

If you opt to use a **hypothesis configuration** you must enter the co-ordinates to be employed in the following window (by default, the values for the first subject configuration are displayed when the window is opened):

The results will automatically appear as a series of graphic displays, as described above. Use the buttons shown to rotate and zoom the configuration displayed. Click on **Next** and **Prev(ious)** to
cycle through the various configurations of the current **PINDIS** output, including the individual ("subject") configurations as transformed in the analysis. With the exception of the subject space, configurations are rescaled, for ease of viewing, according to their largest absolute co-ordinate value.

- Clicking **Draw** allows you to draw on the display with the mouse, to highlight features of interest. Clicking **Line** draws straight lines, from a point where a **mouse button** is depressed to a point where it is lifted. Clicking **Ellipse** draws an ellipse within a notional rectangle outlined using the mouse in a similar manner. Clicking **Text** causes a box to appear to enter text. On closing this box, move the **mouse** to the position required and press a **mouse button** to add the text to the image displayed. Click on **Refresh** to clear any lines and annotations added and return to the original image selected. The image as amended must be **saved** immediately on completion, as the additions will be lost when the display is
reoriented.

- Clicking on an individual point will highlight its label.

- Click on menu item **Labels** to adjust the maximum number of characters of the words displayed as point labels. Click on a point in the display to highlight its label.

- Click on **Grid** to toggle the internal lines on the horizontal plane on and off as preferred.

- With a **mouse button** depressed, drag the pointer to a new location and release the **mouse button** again to move the display around inside the window, to be able to concentrate on particular parts of it.

- Click on **View/Edit Data** to see the full PINDIS results file.

- Click on **Save Display** to save the current display as a graphic image file, for inclusion in other documents. The **Editor** also offers limited facilities to draw on and annotate image files saved by HAMLET procedures.

- Close the display to return to the main PINDIS window. Click on the **Files** menu and select **Exit**, or use the **Alt+F4** key combination, to exit the program. A prompt will appear if any file contents have not been saved.

### 3.5 An example to illustrate the use of PINDIS

The following simple illustration compares solutions from two different multidimensional scaling programs for the well-known Ekman colour-perception data, derived from two different scaling programs. These data, which have frequently been used to illustrate non-metric scaling methods, are based on paired comparisons of stimuli in the form of light displays across the visible spectrum, and are essentially two-dimensional. The stimuli are labelled according to the wavelength of light they represent. However, the configurations produced by the different programs do not immediately look identical in appearance. Observant readers may notice, the solutions are, at least, reflected in the first two dimensions. The question is, do they really differ significantly from the perspective of non-metric scaling?

<table>
<thead>
<tr>
<th>Dimension</th>
<th>W434</th>
<th>W445</th>
<th>W465</th>
<th>W472</th>
<th>W490</th>
<th>W504</th>
<th>W537</th>
<th>W555</th>
<th>W584</th>
<th>W600</th>
<th>W610</th>
<th>W628</th>
<th>W651</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Program One</strong></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W434</td>
<td>-0.283</td>
<td>-0.899</td>
<td>-0.049</td>
<td>0.290</td>
<td>0.730</td>
<td>-0.500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W445</td>
<td>-0.348</td>
<td>-0.827</td>
<td>0.099</td>
<td>0.400</td>
<td>0.720</td>
<td>-0.430</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W465</td>
<td>-0.886</td>
<td>-0.571</td>
<td>-0.085</td>
<td>0.870</td>
<td>0.610</td>
<td>0.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W472</td>
<td>-0.935</td>
<td>-0.491</td>
<td>0.134</td>
<td>0.930</td>
<td>0.530</td>
<td>0.140</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W490</td>
<td>-0.988</td>
<td>0.105</td>
<td>-0.083</td>
<td>0.970</td>
<td>-0.070</td>
<td>0.190</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W504</td>
<td>-0.811</td>
<td>0.661</td>
<td>-0.042</td>
<td>0.760</td>
<td>-0.600</td>
<td>0.270</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W537</td>
<td>-0.492</td>
<td>0.867</td>
<td>-0.229</td>
<td>0.520</td>
<td>-0.860</td>
<td>-0.190</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W555</td>
<td>-0.357</td>
<td>0.927</td>
<td>0.211</td>
<td>0.370</td>
<td>-0.970</td>
<td>0.100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W584</td>
<td>0.525</td>
<td>0.702</td>
<td>0.294</td>
<td>-0.550</td>
<td>-0.660</td>
<td>0.430</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W600</td>
<td>0.787</td>
<td>0.443</td>
<td>-0.140</td>
<td>-0.810</td>
<td>-0.470</td>
<td>-0.060</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W610</td>
<td>1.012</td>
<td>0.042</td>
<td>-0.148</td>
<td>-0.960</td>
<td>-0.070</td>
<td>-0.130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W628</td>
<td>1.008</td>
<td>-0.174</td>
<td>-0.070</td>
<td>-0.990</td>
<td>0.210</td>
<td>-0.060</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W651</td>
<td>0.886</td>
<td>-0.361</td>
<td>-0.155</td>
<td>-0.930</td>
<td>0.400</td>
<td>0.020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For these data, the centroid looks like this:

Clicking on **Next** displays the **subject space** as follows, showing that the two solutions are practically identical, with minor differences only in the third dimension, which can consequently be eliminated:
Clicking on **View/edit data** shows the following detailed information for the PINDIS analysis. This confirms the conclusion reached from visual inspection, that the two scaling solutions compared are indeed practically identical. The differences were probably to be explained by differing numerical precision or convergence criteria in the different programs used. This example also illustrates that MDS solutions which at first sight may appear different can prove to be very similar, under certain transformations, which are progressively applied and fully documented by PINDIS. By examining these results in detail, particularly the ‘vector’ model, which allows individual points to be shifted in distance and direction from the origin, as indicated by their vector weights (lengths) and cosines, it is sometimes possible to highlight the positions of particular stimuli as the reasons for overall differences between the configurations compared. In the present example, comparing the two scaling results for the same data, the effects of allowing differing dimensional weights and correlations are seen to be practically identical, and looking at the vector weights and cosines suggests that the two programs differ, if at all, only very slightly in their placing of the points at opposing ends of the spectrum, program 1 (‘Subject 1’ in the following output), the longest wavelengths, and program 2 (‘Subject 2’), the shortest. The joint weighting model shows them to be, after all, virtually identical.
An example comparing two scalings of the EKMAN colour similarity data

Fit of Centroid Configuration
0.011047
0.011047

Centroid configuration
1   W434   -0.1103   -0.2149   -0.0606
2   W445   -0.1336   -0.1980   -0.0337
3   W465   -0.2596   -0.1139   -0.0091
4   W472   -0.2732   -0.0877   -0.0290
5   W490   -0.2553    0.0726   -0.0979
6   W504   -0.1794    0.2087   -0.0026
7   W537   -0.0893    0.2443   -0.0886
8   W555   -0.0542    0.2685   -0.0909
9   W584   0.1684    0.1666    0.0913
10  W600    0.2348    0.0818   -0.0210
11  W610    0.2660   -0.0339   -0.0172
12  W628    0.2564   -0.0975    0.0098
13  W651    0.2237   -0.1408    0.0117
14  W674    0.2057   -0.1557    0.0919

****  Perspective Origins  ****

Subject
1  -0.1606   -0.2959   -1.3429
2   0.0935    0.3574   -1.1140

****  Analytic Solutions for Individual Configurations  ****

****  Configuration for Subject   1  ****

**** Program one

1)  Similarity Transformations

Normed scalar for unconditional weights:  1.000000

1   W434   -0.1153   -0.2246   -0.0203
2   W445   -0.1322   -0.2022    0.0181
3   W465   -0.2582   -0.1100   -0.0423
4   W472   -0.2720   -0.0860    0.0154
5   W490   -0.2537    0.0735   -0.0430
6   W504   -0.1821    0.2123   -0.0276
7   W537   -0.0845    0.2515   -0.0706
8   W555   -0.0554    0.2622    0.0501
9   W584    0.1648    0.1619    0.0907
10  W600    0.2310    0.0802   -0.0202
11  W610    0.2718   -0.0362   -0.0180
12  W628    0.2590   -0.0930    0.0025
13  W651    0.2200   -0.1369   -0.0231
14  W674    0.2069   -0.1526    0.0883

Fit of subject to centroid/hypothesis ---  S(Z,X) =  0.988953

2)  Dimensional-weighting Transformations

Dimensional Weights:
1.0003    1.0025    0.7870

Dimensional Correlations:
3) Perspective Models - Vector Weighting

Translation of Individual Configuration:

\[-0.0386 \quad 0.0313 \quad 0.1126\]

Stimuli:        Vector Weights:        Vector Cosines:
1 W434          1.0907              0.9916
2 W445          1.0845              0.9863
3 W465          0.9743              0.9946
4 W472          0.9928              0.9992
5 W490          0.9807              0.9951
6 W504          1.0011              0.9973
7 W537          1.0313              0.9983
8 W555          1.0336              0.9931
9 W584          0.9738              0.9999
10 W600         0.9644              0.9999
11 W610         0.9928              0.9997
12 W628         0.9648              0.9992
13 W651         0.8868              0.9915
14 W674         0.9694              0.9999

Fit --- S(ZW,X) = 0.990200

4) Joint Weighting Solution

Dimensional Weights:
1.0055      1.0075      0.8054

Vector Weights:
1 W434        1.0138
2 W445        0.9836
3 W465        0.9874
4 W472        0.9866
5 W490        0.9938
6 W504        1.0105
7 W537        1.0107
8 W555        0.9748
9 W584        0.9922
10 W600       0.9790
11 W610       1.0176
12 W628       0.9968
13 W651       0.9698
14 W674       1.0057

Fit --- S(VZW,X) = 0.990303

**** Configuration for Subject 2 ****
**** Program two

1) Similarity Transformations

Normed scalar for unconditional weights: 1.004231

1 W434 -0.1041 -0.2028 -0.1004
2 W445 -0.1335 -0.1916 -0.0851
3 W465 -0.2582 -0.1165 0.0241
4 W472 -0.2714 -0.0884 0.0422
5 W490 -0.2540 0.0709 0.0237
6 W504 -0.1747 0.2028 0.0224
7 W537 -0.0931 0.2344 -0.1057
8 W555 -0.0524 0.2719 -0.0321
9 W584 0.1701 0.1694 0.0908
10       W600    0.2360    0.0824   -0.0217
11       W610    0.2573   -0.0312   -0.0162
12       W628    0.2509   -0.1010    0.0171
13       W651    0.2250   -0.1431    0.0463
14       W674    0.2022   -0.1570    0.0945

Fit of subject to centroid/hypothesis --- $S(Z,X) = 0.988953$

2) Dimensional-weighting Transformations

Dimensional Weights:
0.9884      0.9864      1.2016
Dimensional Correlations:
0.9999      0.9996      0.9109

Fit --- $S(ZW,X) = 0.990200$

3) Perspective Models - Vector Weighting

Translation of Individual Configuration:
-0.0301      0.0329      0.1144

Stimuli:                Vector Weights:      Vector Cosines:
1 W434              0.8787              0.9832
2 W445              0.8866              0.9753
3 W465              1.0153              0.9965
4 W472              0.9965              0.9996
5 W490              1.0188              0.9954
6 W504              1.0021              0.9968
7 W537              0.9764              0.9988
8 W555              0.9708              0.9936
9 W584              1.0221              1.0000
10 W600             1.0283              1.0000
11 W610             0.9879              0.9994
12 W628             1.0108              0.9994
13 W651             1.0833              0.9945
14 W674             1.0051              1.0000

Fit --- $S(VZ,X) = 0.992053$

4) Joint Weighting Solution

Dimensional Weights:
0.9956      0.9902      1.1956

Vector Weights:
1 W434        0.9883
2 W445        1.0172
3 W465        0.9994
4 W472        1.0026
5 W490        0.9923
6 W504        0.9787
7 W537        0.9810
8 W555        1.0145
9 W584        0.9872
10 W600        1.0089
11 W610        0.9697
12 W628        0.9917
13 W651        1.0215
14 W674        0.9771

Fit --- $S(VZW,X) = 0.990256$

Normalized Dimension Weights
<table>
<thead>
<tr>
<th>Subject</th>
<th>Communality</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9902</td>
<td>0.7701</td>
<td>0.6146</td>
<td>0.1397</td>
</tr>
<tr>
<td>2</td>
<td>0.9902</td>
<td>0.7609</td>
<td>0.6047</td>
<td>0.2134</td>
</tr>
<tr>
<td>Mean</td>
<td>Communality</td>
<td>0.9902</td>
<td>0.5860</td>
<td>0.3717</td>
</tr>
</tbody>
</table>

--- Table of Subject Communalities for PINDIS Transformations ---

Transformation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>Mean</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
</tbody>
</table>

* 0.0 means that particular PINDIS transformation was not used.

For further information about these procedures, see the references below, especially Davies and Coxon (1982) and Borg and Groenen (2005).

4. Correspondence Analysis (CA)

represents the row and column categories of the input matrix as points in the same dimensionality. It is closely related to other procedures which seek to represent row and column variables in the same space, using a singular value decomposition. It is important to realise that CA considers only interactive factors by explicitly neglecting the magnitude effect after decomposition. Because the canonical ("optimal") scores reported are row and column conditional, it is advisable to avoid inter-set point distance interpretation, however tempting this may be, when using correspondence analysis.

The input matrix is first normalized by dividing each row entry by the square root of the product of the corresponding row and column totals, their geometric mean. This removes differences in the marginal totals and expresses each cell as a proportion.

The second step finds the basic structure of the resultant matrix A by singular value decomposition, producing summary row and column vectors (U and V) and a diagonal matrix of singular values d corresponding to the columns of A, so that A = Ud(VT). The matrices U and V are in fact the eigenvectors of the matrices of row and column cross-products of A, and the d values are related to their (identical) eigenvalues (d=sqrt(D*(n-1)), where D is the diagonal of eigenvalues and n is the number of rows in A). The first singular value in d is always 1.0. It corresponds to the independence model of chi-squared expected values, and is ignored in subsequent analysis.

It is important to check that the eigenvalues remaining after ignoring the first one are in fact large enough to justify continuing with the analysis. Reference should be made to the chi-
squared contributions of each dimension of "inertia" and to the overall chi-squared value for the analysis.

The method implemented here is equivalent to HOMALS in SPSS, which uses an alternating least squares algorithm which is considered more suitable for large numbers of cases.

For further information on this procedure see Weller and Romney (1990), Greenacre (1993).

4.1 Using Correspondence Analysis

At File name: enter the name of the file of context unit profiles to be read, or click on the Files menu and select Open file, or use the appropriate speed button, to search for it. From the Files menu, it is also possible to Reopen directly files used recently.

When a valid file name is entered, the categories included are displayed. Select the number of dimensions to be reported (2 or 3), and click Include these categories to start the analysis.

Click Cancel to return to the main HAMLET window. Once the process has started, click on Cancel to abandon the analysis at any time.

The resulting configuration is automatically displayed in graphic form.

Close this display to return to the Correspondence Analysis window shown above.
Click on the **Data** menu item to view or edit the correspondence analysis result currently being displayed in detail.

Use the control buttons shown to rotate and zoom the configuration displayed. Use the keys indicated to rotate and zoom in and out on the display.

Click on menu item **Labels** to adjust the maximum number of characters of the words displayed as point labels. Click on a point in the display to highlight its label.

Click on **Grid** to toggle the internal lines on the horizontal plane on and off as preferred.

With a mouse button depressed, drag the pointer to a new location and release the mouse button again to move the display around inside the window, to be able to concentrate on particular parts of it.

Clicking **Draw** allows you to draw on the display with the mouse, to highlight features of interest. Clicking **Line** draws straight lines, from a point where a mouse button is depressed to a point where it is lifted. Clicking **Ellipse** draws an ellipse within a notional rectangle outlined using the mouse in a similar manner. Clicking **Text** causes a box to appear to enter...
text. On closing this box, move the mouse to the position required and press a mouse button to add the text to the image displayed. Click on Refresh to clear any lines and annotations added and return to the original image selected. The image as amended must be saved immediately on completion, as the additions will be lost when the display is reoriented.

- Click on Save Display to save the current display as a graphic image file, for inclusion in other documents. The Editor also offers limited facilities to draw upon and annotate image files saved by HAMLET procedures.

- Close the display to return to the Correspondence Analysis window. Click on the Files menu and select Exit, or use the Alt+F4 key combination, to exit the program. A prompt will appear if any file contents have not been saved.

5. Basic utilities for looking at word usage

HAMLET for Windows offers the following additional utilities, to assist in exploration of word usage in texts intended for further analysis, and assist in the creation of search lists for use with the main procedure, HAMLET joint frequencies.

5.1 WORDLIST displays a sorted list of words found in a file of text. Any character-string containing letters and numerals may be considered as a ‘word’ for this purpose. WORDLIST can also be used to check how HAMLET will read the contents of a file and help to avoid errors in later analysis. Files created with WORDLIST can be compared with other lists in the same format using the companion procedure COMPARE.

5.1.1 Using WORDLIST

- At File name: enter the name of the file to be read, or click on the Files menu and select Open file to search for it. The Files menu also allows you to Reopen directly files used recently to save having to look for them. Alternatively, use the corresponding speed button or Alt/ key combination. If the text is in a language other than English, use the pull-down menu on the right of the control panel to select the appropriate lexicographic conventions to be employed.
To limit counting to words according to length, enter the minimum and maximum number of letters in the boxes indicated.

To limit counting by initial letters, enter the relevant letters in the box provided, otherwise leave the default entry "list all words occurring ..".

Toggle the appropriate button to list words in forward alphabetical order or descending order of frequency.

Click on **Create a Word List** to start counting words.

The resulting list will be displayed in a temporary edit window when finished. The progress indicator shows text reading in process. Click on the **Cancel** button to stop the counting process at any time.

Click on the **Files** menu and select **Save word list**, or use the corresponding speed button, to save the resulting list in a file.

5.2 **COMPARE** is used to compare pairs of the sorted vocabulary lists produced by the companion program **WORDLIST**. It simply displays in parallel columns the frequencies for any words which appear in both lists. The results can then be saved to compare with other lists produced by **WORDLIST** or by earlier applications of **COMPARE** itself.

Percentage frequencies are shown, based on the words counted by **WORDLIST** in the original text, or are the weighted averages of the percentages for the same words in pairs of files whose frequencies have been combined by **COMPARE**.
5.2.1 Using COMPARE

At Wordlist File One: and Wordlist File Two: enter the names of the files to be compared, or click on the Files menu and select Open file to search for them. Alternatively, use the corresponding speed button or Alt/key combination. If the lists to be read are in a language other than English, use the pull-down menu on the right of the control panel to select the appropriate lexicographic conventions.

Click on Create list of words common to both files to start comparison. The results will be displayed in a temporary edit window when complete. Click on Cancel to stop the process at any time. Patience may be needed when comparing large lists, depending on the computing resources you have available. If progress appears very slow, check for continuing disk activity before concluding that there is an error.

Click on the Files menu and select Save word list to save the resulting list in a file.

5.3 KWIC (Key_Word_In_Context) can be used to produce traditional single line KWIC listings, with one or more keywords in the centre, or display the keyword(s) within a larger block of text.

5.3.1 Using KWIC

At File name: enter the name of the file to read, or click on the Files menu and select Open file to search for it. The Files menu also allows you to Reopen directly files used recently to save having to look for them. Alternatively, use the corresponding speed button or Alt/key combination. If the text to be read is in a language other than English, use the pull-down menu
on the right of the control panel to select the appropriate lexicographic conventions to apply.

- At Search for: enter a word or phrase to search for - this may include the ‘wild card’ characters ‘@’ and ‘*’, causing individual letters corresponding to the position of the character ‘@’, and all characters after, and including, the position of the character ‘*’, to be ignored in searching. This, for example, provides a way to treat as equivalent words which differ only in their suffixes, e.g. plural forms. Click on the adjacent button to see a display of all of the words in the file selected, with their frequencies. You can select and copy words from this window into the main panel using the mouse. With the left mouse button: clicking on ‘Words >’ reverses the order of the items listed, and clicking on ‘Freq.’ causes the columns to swap positions. Clicking anywhere in the vocabulary list with the right mouse button switches the listing between alphabetical and frequency order.

- You can select and copy words from this list into the main panel. Checking the appropriate item below will cause occurrences of the words entered to be listed as a phrase, or separately.

- Enter the size of the context to display the keyword, if found, in lines of text: \((1 \leq n \leq 15)\).

- Click on Create Key_Word_In_Context List to start searching the text for the string to be identified. The results will be displayed in a temporary edit window when completed. Click on Abort to stop the process at any time.

- Click on the Files menu and select Save KWIC list, or use the corresponding speed button, to save the resulting list in a file.

5.4 Text Profile

displays a profile of the distributions of word and sentence lengths for a given text file. This shows mean word and sentence lengths, with dispersion statistics and a histogram for each distribution. It therefore supplements the information available from Wordlist, for the entire text file, without
the restrictions which may be applied in that procedure.

- At **File name:** , enter the name of the file to be read, or click on the **Files** menu and select **Open file** to search for it, or use the corresponding speed button with the same effect. The **Files** menu also allows you to **Reopen** directly files used recently to save having to look for them.

- Select the **Internet** item from the **Files** menu, or the corresponding speed button, to access a file from the Internet.

- Use the pull-down menu at the top right of the display to select the correct character set to use if the text contains words of a language other than English.

- If necessary, amend the set of characters marking the end of sentences in this text.

- Click on **Create text profile** to start the process. The progress indicator shows the text being read. Click on **Cancel** at any time to interrupt the process. The text profile is displayed as follows:
The graphic image may be saved by clicking on the **Save** button at the bottom of the display. Click **Close** to return to the previous window.

On clicking **Save results** on the main menu, or on closing the procedure, you will be prompted to save a file containing the equivalent results in text form.
I. References and further reading


A. P. Brier "HAMLET: A Pascal Program to Count Joint Frequencies of Words in a Text", *Siegener Periodicum für internationale empirische Sozialwissenschaft*, 4,1,(1985),pp.177-196


P. M. Davies and A. P. M. Coxon (eds.) **Key Texts in Multi-dimensional Scaling**, (1982).


M. G. Greenacre, **Correspondence Analysis in Practice**, Academic Press (1993).

H. P. Iker "An historical note on the use of word-frequency contiguities in content analysis",


R.Langeheine Approximate Norms and Significance Tests for the LINGOES-BORG Procrustean Individual Differences Scaling (PINDIS), Institut für die Pädagogik der Naturwissenschaften, University of Kiel (1980).


II. Adapting to other language conventions

HAMLET II for Windows© is supplied with default configuration files to cover the lexical conventions of major European languages. Should these prove unsuitable for a language or purpose not so far considered, this section explains how the file is structured. Future releases of HAMLET II for Windows may be accompanied by configuration files for further languages, taking account of the experience and requirements reported by users.

To add new characters to those recognised by the existing language options, select the items Help | Language options from the main HAMLET II menu (see above, p.4). When the language options window is displayed, you can click on the Show characters button to display, and, if necessary, amend the character set associated with the language code currently selected. On entering previously unrecognised characters, you will be prompted to confirm their upper- and lower-case forms, and, optionally, enter a mapping (into one or two existing characters) to determine their lexical ordering to be applied to them, if this is different from the Windows default ordering. The character '@' as shown above might, for example, be required to be sorted in a position equivalent to '91'. In this way, characters from the keyboard can be assigned their correct lexical ordering, according to the conventions of the language selected.

HAMLET II for Windows and WORDLIST automatically offer similar opportunities to amend the language conventions, on encountering unrecognised characters when reading a text.

-----------------------------------------------------------------------------

It is also possible to define new language convention files, if the language required is not currently available in HAMLET II. The following default file INTERNATWIN can serve as a model for creating your own configuration files as necessary to suit the requirements of other languages. Note that blank lines and those beginning with ";" or "*" are ignored by the program and can therefore be used for annotation. All amendments should, of course, be made using the Windows® character set, which is significantly different from that used by MS-DOS. If in doubt, you can use the utility (available from the Tools | OEM to Ansi conversion menu item) to make sure that any
text files to be read are compatible with **Windows**.

; INTERNAT.WIN - Sample Windows collating sequence
;
; * NB In reading the following file, blank lines and comments
; * (running from ";" or "\*" to EOLN) are ignored *
;
; This file may be edited to include any other characters which may
; be required. Only the characters specified here will be used in
; word searching routines in the HAMLET package for Windows.
; Currently, two sections are defined, as detailed below:
; Section headings may be enclosed by any pair of bracket characters.
;
; [UPCASE]
; This section contains all information needed for to convert characters
; in language set to upper-case equivalents. It also creates a set of
; characters which are used in other routines. Entries must consist of
; pairs of characters, the upper-case form followed by the lower-case
; equivalent. Numerals if they are to be recognised should also be
; included here, as repeated pairs, as should any other valid characters
; for which there is no normal upper-case equivalent in the printed
; language.
;
; [MAPPING]
; The mapping section contains all character mappings significant for
; the correct lexical ordering of lists of words in a given language.
; Only special or accented characters need to be specified, since any
; characters not specified will collate in the normal ASCII order.
; The order of entries in this section is irrelevant.

[UPCASE]
00112233445566778899AaÄäÅåÁáÀàÂâBbCcÇçDdEeÉéÈèÊêFfGgHhIiÍíÎîJj
KkLlMmNnÑñOoÖöÓóÒòÔôPpQqRrSsßßTtUuÜüÚúÙùÛûVvWwXxYyYyZz

*******************************************************
* Mappings for correct ordering of European languages *
*******************************************************

* Germanic
Ä=AE
Ö=OE
Ü=UE
ä=ae
e=oe
ü=ue
ß=SS

* Other
È=EE
e=ee
e=ee
Ç=C
c=c
d=aa
ça
à=aa
à=aa
ù=uu
ô=oo
Having created a similar configuration file for a language not currently covered by HAMLET II for Windows, for example, a file containing Polish language conventions, called POLISH.CFG, you should first copy this to the directory in which the main HAMLET routines are installed (normally this will be called C:\Program files\HAMLET II\).

Next, load the main HAMLET II program, select Help | Language options from the main menu, and click the Edit button to enter the following window:
Simply enter the relevant 2-character Windows language code together with the name of the newly-created file containing the corresponding lexicographic conventions, and click on **OK** to add this code to those already available. You can check the effect of a new file by again clicking on the **Show characters** button, as described above.

### III. Technical details

**HAMLET II for Windows** requires at least 10 Megabytes of space to install on a hard disk or similar medium. All the necessary files are automatically installed in one main directory, as directed at the time of installation. All temporary files are stored in the directory to which permission is assigned in the current Windows setup for this purpose.

Language convention files with the extensions **.win** (or **.cfg**), as described above (pp.36-39), must be placed together with the file **HAMLET.INI** in the main installed directory.

Files with the following default extensions may be saved in the course of working with **HAMLET**, to store the following program data and results:

- **.voc** - vocabulary lists for use with HAMLET Joint Frequencies
- **.mat** - matrix of co-occurrences of word pairs, from HAMLET Joint Frequencies used as input to MINISSA, Cluster Analysis, or INDSCAL
- **.xpr** - context profiles created by HAMLET Joint Frequencies for use with Correspondence Analysis
- **.xpc** - results of Correspondence Analysis
- **.ham** - output listing from HAMLET Joint Frequencies
- **.txt** - listing of word clusters identified by Cluster Analysis
- **.min** – MINISSA output listing, accessed by SELECT to generate input to PINDIS
- **.inp** – PINDIS input file
- **.pin** – PINDIS output listing
.bmp, .jpg - Graphic files to store results displayed by MINISSA, PINDIS, Correspondence Analysis, INDSCAL and PROFILE
.lst - Word list file, generated by Wordlist or Compare
.kwc – Key-Word-In-Context listing

Note that, apart from .bmp and .jpg, these are only conventions, and may be varied to suit the convenience of individual users should they conflict with file extensions already assigned to other software in use on the machine on which HAMLET is installed.