Zipf 2.0 Documentation

Benjamin Elbirt

Table of Contents

Software License Agreement

PLEASE READ THIS SOFTWARE LICENSE AGREEMENT CAREFULLY BEFORE INSTALLING OR USING THE SOFTWARE.

BY OPENING THE PACKAGE, DOWNLOADING THE PRODUCT, OR USING THE EQUIPMENT THAT CONTAINS THIS PRODUCT, YOU ARE CONSENTING TO BE BOUND BY THIS AGREEMENT. IF YOU DO NOT AGREE TO ALL OF THE TERMS OF THIS AGREEMENT, RETURN THE PRODUCT TO THE PLACE OF PURCHASE FOR A FULL REFUND, OR DO NOT DOWNLOAD THE PRODUCT.

Single User License Grant: Benjamin Elbirt (“Owner”) grants to Customer ("Customer") a nonexclusive and nontransferable license to use **Zipf 2.0** © ("Software") in object code form solely on a single central processing unit owned or leased by Customer.

Multiple-Users License Grant: Benjamin Elbirt (“Owner”) grants to Customer ("Customer") a nonexclusive and nontransferable license to use **Zipf 2.0** ("Software") in object code form: (i) installed in a single location on a hard disk or other storage device of up to the number of computers owned or leased by Customer for which Customer has paid a license fee ("Permitted Number of Computers"); or (ii) provided the Software is configured for network use, installed on a single file server for use on a single local area network for either (but not both) of the following purposes: (a) permanent installation onto a hard disk or other storage device of up to the Permitted Number of Computers; or (b) use of the Software over such network, provided the number of computers connected to the server does not exceed the Permitted Number of Computers. Customer may only use the programs contained in the Software (i) for which Customer has paid a license fee (or in the case of an evaluation copy, those programs Customer is authorized to evaluate. Customer grants to Owner or its independent accountants the right to examine its books, records and accounts during Customer's normal business hours to verify compliance with the above provisions. In the event such audit discloses that the Permitted Number of Computers is exceeded, Customer shall promptly pay to Owner the appropriate licensee fee for the additional computers or users. At Owner option, Owner may terminate this license for failure to pay the required license fee.

Customer may make one (1) archival copy of the Software provided Customer affixes to such copy all copyright, confidentiality, and proprietary notices that appear on the original.

EXCEPT AS EXPRESSLY AUTHORIZED ABOVE, CUSTOMER SHALL NOT: COPY, IN WHOLE OR IN PART, SOFTWARE OR DOCUMENTATION; MODIFY THE SOFTWARE; REVERSE COMPILE OR REVERSE ASSEMBLE ALL OR ANY PORTION OF THE SOFTWARE; OR RENT, LEASE, DISTRIBUTE, SELL, OR CREATE DERIVATIVE WORKS OF THE SOFTWARE.

Customer agrees that aspects of the licensed materials, including the specific design and structure of individual programs, constitute trade secrets and/or copyrighted material of Owner. Customer agrees not to disclose, provide, or otherwise make available such trade secrets or copyrighted material in any form to any third party without the prior written consent of Owner. Customer agrees to implement reasonable security measures to protect such trade secrets and copyrighted material. Title to Software and documentation shall remain solely with Owner.

LIMITED WARRANTY. Owner warrants that for a period of ninety (90) days from the date of shipment from Owner: (i) the media on which the Software is furnished will be free of defects in materials and workmanship under normal use; and (ii) the Software substantially conforms to its published specifications. Except for the foregoing, the Software is provided AS IS. This limited warranty extends only to Customer as the original licensee. Customer's exclusive remedy and the entire liability of Owner under this limited warranty will be, at Owner option, repair, replacement, or refund of the Software if reported (or, upon request, returned) to the party supplying the Software to Customer. In no event does Owner warrant that the Software is error free or that Customer will be able to operate the Software without problems or interruptions.

This warranty does not apply if the software (a) has been altered, except by Owner, (b) has not been installed, operated, repaired, or maintained in accordance with instructions supplied by Owner, (c) has been subjected to abnormal physical or electrical stress, misuse, negligence, or accident, or (d) is used in ultra-hazardous activities.

DISCLAIMER. EXCEPT AS SPECIFIED IN THIS WARRANTY, ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS, AND WARRANTIES INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE, ARE HEREBY EXCLUDED TO THE EXTENT ALLOWED BY APPLICABLE LAW.

IN NO EVENT WILL Owner BE LIABLE FOR ANY LOST REVENUE, PROFIT, OR DATA, OR FOR SPECIAL, INDIRECT, CONSEQUENTIAL, INCIDENTAL, OR PUNITIVE DAMAGES HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY ARISING OUT OF THE USE OF OR INABILITY TO USE THE SOFTWARE EVEN IF Owner HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In no event shall Owner liability to Customer, whether in contract, tort (including negligence), or otherwise, exceed the price paid by Customer. The foregoing limitations shall apply even if the above-stated warranty fails of its essential purpose. SOME STATES DO NOT ALLOW LIMITATION OR EXCLUSION OF LIABILITY FOR CONSEQUENTIAL OR INCIDENTAL DAMAGES.

The above warranty DOES NOT apply to any beta software, any software made available for testing or demonstration purposes, any temporary software modules or any software for which Owner does not receive a license fee. All such software products are provided AS IS without any warranty whatsoever.

This License is effective until terminated. Customer may terminate this License at any time by destroying all copies of Software including any documentation. This License will terminate immediately without notice from Owner if Customer fails to comply with any provision of this License. Upon termination, Customer must destroy all copies of Software.

Software, including technical data, is subject to U.S. export control laws, including the U.S. Export Administration Act and its associated regulations, and may be subject to export or import regulations in other countries. Customer agrees to comply strictly with all such regulations and acknowledges that it has the responsibility to obtain licenses to export, re-export, or import Software.

This License shall be governed by and construed in accordance with the laws of the State of New York, United States of America, as if performed wholly within the state and without giving effect to the principles of conflict of law. If any portion hereof is found to be void or unenforceable, the remaining provisions of this License shall remain in full force and effect. This License constitutes the entire License between the parties with respect to the use of the Software.

I Zipf 2.0 Product Specification

I Product Overview

Zipf 2.0© is a java application for performing analysis of multi-lingual text using a co-occurrence semantic text/content analysis (Harris, 1957; Hildum, 1963), traditional frequency distributions (Zipf, 1935) and learning/forgetting concepts (Woelfel & Fink, 1980). Controls are provided for exclusion / replacement of characters / tokens, learning and forgetting rates, window size and slide rate, directionality, and distance within window.

MDSJ™ (Brandes & Pich, 2007) - a Java Library for Riemann-Space (non-Euclidean) multidimensional scaling – is used to provide coordinate generation with controls for the number of dimensions to use.

II Software Installation

This software can/should run from any location and does not depend on additional files.

III Memory Expansion

ZIPF 2.0© uses as much memory as made available through the Java environment and the command line execution. To increase the memory, use:

```
java–Xms1M  –Xmx2M –jar zipf.jar
```

where ***1M*** is the amount of memory to use for the initial heap size and ***2M*** is the amount of memory to use for the maximum heap size. The **M** is used to represent megabytes.

```
java  –Xms1000M –Xmx2000M –jar zipf.jar
```

will launch the application with an initial heap of 1 GB and a maximum heap of 2 GB.

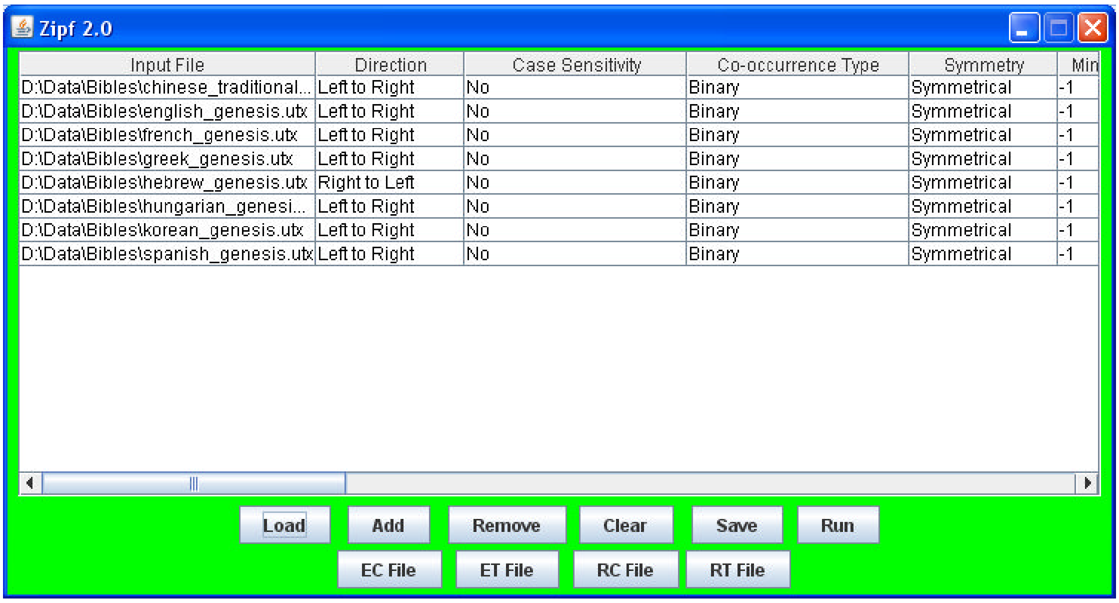
IV Bug Collection & Reporting

Use the "java –jar zipf.jar" command from a prompt (DOS) to obtain error messages from the executable. Please report any bugs / problems to the author, Benjamin Elbirt, at elbirt@elbirttechnologies.com.

VI Future Enhancements

1. Additional Output Metrics
2. Neural-Network learning vs. Co-occurrence
3. Out of Memory Error Trapping

II Data Specification Interface



Input File	Direction	Case Sensitivity	Co-occurrence Type	Symmetry	Min
D:\Data\Bibles\chinese_traditional...	Left to Right	No	Binary	Symmetrical	-1
D:\Data\Bibles\english_genesis.utx	Left to Right	No	Binary	Symmetrical	-1
D:\Data\Bibles\french_genesis.utx	Left to Right	No	Binary	Symmetrical	-1
D:\Data\Bibles\greek_genesis.utx	Left to Right	No	Binary	Symmetrical	-1
D:\Data\Bibles\hebrew_genesis.utx	Right to Left	No	Binary	Symmetrical	-1
D:\Data\Bibles\hungarian_genesi...	Left to Right	No	Binary	Symmetrical	-1
D:\Data\Bibles\korean_genesis.utx	Left to Right	No	Binary	Symmetrical	-1
D:\Data\Bibles\spanish_genesis.utx	Left to Right	No	Binary	Symmetrical	-1

Data Specification Interface

I Control Buttons

These buttons provide the controls for defining Input files and processing configurations.

- **Load** – Use this button to load a pre-existing definitions file created with the *Save* button.

- **Add –** Use this button to add input files (multiple) to the definitions list with the default settings. Duplication of input files for multiple control options on execution is possible. See the Outputs section for information on output file naming.

- **Remove** – This button will remove all selected rows from the definition list.

- **Clear –** This button will remove all rows from the definition list.

- **Save –** This button will save the current definitions to the specified file for use with the *Load* button.

- **Run** – This button will start the analysis process for all defined input files.

- **EC File, ET File, RC File, RT File** – These buttons will load the respective file type and apply it to *ALL* selected files in the definition.

II Control Fields

- **Input File** – This is the name of the file to process for the given input instance. Multiple copies of the same file can be specified with various processing control options. This value cannot be modified directly.

- **Direction –** This is the direction for text processing (Left to Right vs. Right to Left). The default is Left to Right.

- **Case Sensitivity** – This is the Yes/No option for considering the character case (upper vs. mixed). The default is No and all lower case letters will be capitalized.

- **Co-occurrence Type, Symmetry** – These options control the co-occurrence method and can be Binary or Distance; Symmetrical or Directed.

  Binary relations are all incremented by one (1) for all tokens in the window. Distance relations are incremented as a function of the distance between the tokens with the default distance of 1.

  Symmetrical relations consider all the tokens in the window being related to each other irrelevant of position and direction of text flow. Directed relations consider the relations as flowing from Left to Right only (or Right to Left with the Direction option) and thus does not consider the words backward related.

  Note: A token cannot be related to itself and thus all token relations to the self are considered zero (0).

- **Min Token Size, Max Token Size** – These specify the minimum and maximum token size to consider. All tokens outside of the size restrictions will be ignored. Use -1 to disable the restriction.

- **Run MDS, MDS Dimensions** – This option (Yes/No) will determine if MDS coordinate generation should be applied to the co-occurrence matrix. The total dimensions to generate should be at least 1 or -1 to generate all possible dimensions (number of tokens).

  Note: The more dimensions generated and tokens the more memory needed for this function to properly work. Higher dimensionality means less error in the results.

- **Output Type** – This is the type of output file to generate and can be Microsoft Excel or UTF8/Unicode Text.

- **Window Size, Slide Rate** – These are the controls for how the co-occurrence will process. The window size defines the number of tokens to consider for co-occurrence and the slide rate is how far to move the window after each window is processed.

  Window size must be at least 2; Slide rate must be at least 1.

- **Learning Rate, Forgetting Rate** – These are additional controls for use with the co-occurrence process that attempt to simulate the learning/forgetting feature of human beings.

  The learning rate is multiplied by the relation distance to dampen/improve the relation strength for the given relation. This option is useless when the Binary Co-occurrence Type is used as the distance is always 1.

  The forgetting rate is applied by multiplying the value against all relations for all tokens after each window is processed. This has the effect of dampening the relation distance of all relations over time. The more a relation exists, the stronger it will be after forgetting as compared to less existent relations.

  Note : All relations below 0.0000001 are considered zero and removed during forgetting rate processing.

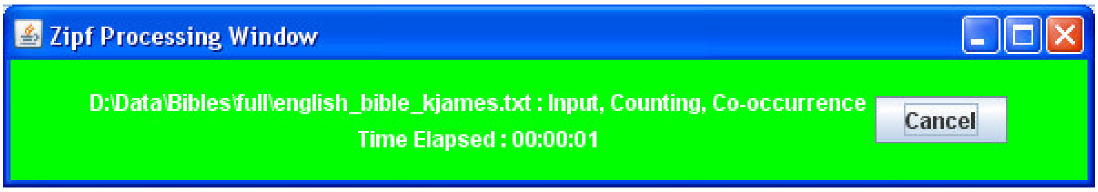
- **EC File** – This is the Exclude Characters file and should contain a list of characters, one per line, in Unicode/UTF8 format to be ignored during input file processing.

- **ET File** – This is the Exclude Tokens file and should contain a list of tokens, one per line, in Unicode/UTF8 format to be ignored during input file processing. Tokens can consist of any non-end of line characters including spaces and special characters.

- **RC File** – This is the Replace Characters file and should contain a list of characters, one per line, in Unicode/UTF8 format to be converted such that the first character is converted to the second in the sequence.

- **RT File** – This is the Replace Tokens file and should contain a list of Tokens, one per line, in Unicode/UTF8 format to be converted such that the first token is converted to the second in the sequence. Tokens can consist of any non-end of line characters including spaces and special characters.

III Processing Window



D:\Data\Bibles\full\english_bible_kjames.txt : Input, Counting, Co-occurrence

Time Elapsed : 00:00:01

Processing Window

I Processing Information

The processing window provides information on the current action in the process. The name of the file being worked on and the action being performed are specified. The Time Elapsed represents total time since the processing began.

II Cancellation of Process

The cancel button will stop the process in the current state. All files processed completely will have the output available; non-completed files will not have any output.

III Errors and Handling

All errors are trapped by the application and provided in an error message window upon occurrence. The process will cancel as though the cancel button were pressed upon error with outputs as previously defined.

WARNING: Out-Of-Memory errors do not trap at this time and will only be displayed if the DEBUGGING method is used for error reporting. Further, the processing window will not “cancel” upon this error and may require a hard termination to reset the application.

IV Outputs

I File Naming

Output files are generated in the same directory the input file is obtained from and given the same file name as the file being analyzed. Additional naming information is appended to the file name preventing the over-writing of the original file. An integer index is used should the file to be generated as output exist. This index will be incremented and tested until a new file can be generated.

Thus, if the input is C:\Temp\Hebrew_Genesis.utx and an Excel Output file is to be generated the output file will be C:\Temp\Hebrew_Genesis.utx.xls. The file name C:\Temp\Hebrew_Genesis.utx.1.xls will be attempted should the non-indexed file exist. The index will be incremented to 2, 3, etc. until a new file can be created.

Note: Multiple copies of the same input file will result in incremental index usage for the outputs. The order of output index will be the order of inputs listed in the definitions adjusting for any pre-existing output files by the same name.

II Microsoft Excel File

The resulting Microsoft Excel file will contain three spreadsheets:

1. Zipf Frequency Data – This will contain each Token, the document frequency, document percentage and token size.
2. Coordinates – This will be all dimensions created with the Token name.
3. Co-occurrence Matrix – This will be an NxN matrix, with Token name column/row header and the relational distance for each token to the others.

III UTF8/Unicode Text File

Three files will be created containing the data as described in the Microsoft Excel file output using UTF8/Unicode compatible character sets. The files will be named using the conventions:

- Hebrew_Genesis.utx.**1.zipf.tokens.utx** – The Zipf Frequency Data.
- Hebrew_Genesis.utx.**1.coords.tokens.utx** – The Coordinate Data.
- Hebrew_Genesis.utx.**1.cmx.tokens.utx** – The Co-occurrence Matrix.

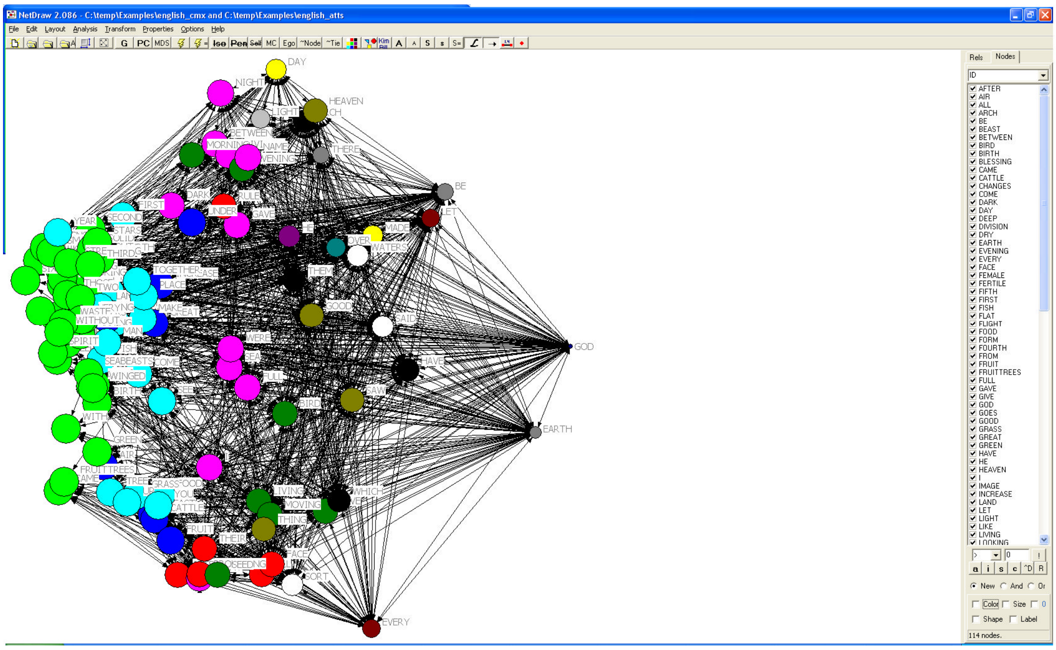
V Example - The Bible (Genesis)

The examples folder provided with this application contains Unicode files of Genesis, the first book of the Old Testament, in Traditional Chinese, Spanish, Hebrew, English, Korean, Hungarian, Greek, and French.

Replace/exclude files for both characters/tokens are provided based on the contents of the data files for use.

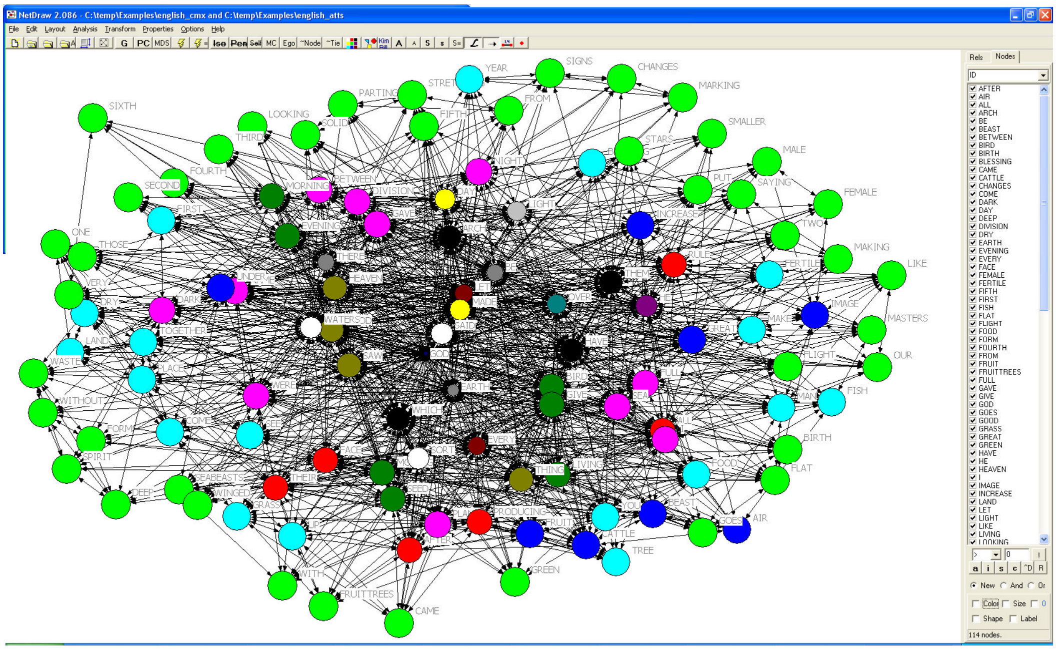
A definition file for loading has been provided with the assumption that the data files are located on a Windows PC in the C:\TEMP\Examples\ directory. Please note that the definition file uses the exclude and replace files and utilizes right to left processing for Hebrew.

The following are screenshots of results graphed using UCINet and NetDraw (Borgatti, Everett, & Freeman, 2002). Regrettably this software does not support UTF8/Unicode text labels and the Hebrew example uses ??? for the symbol labels.



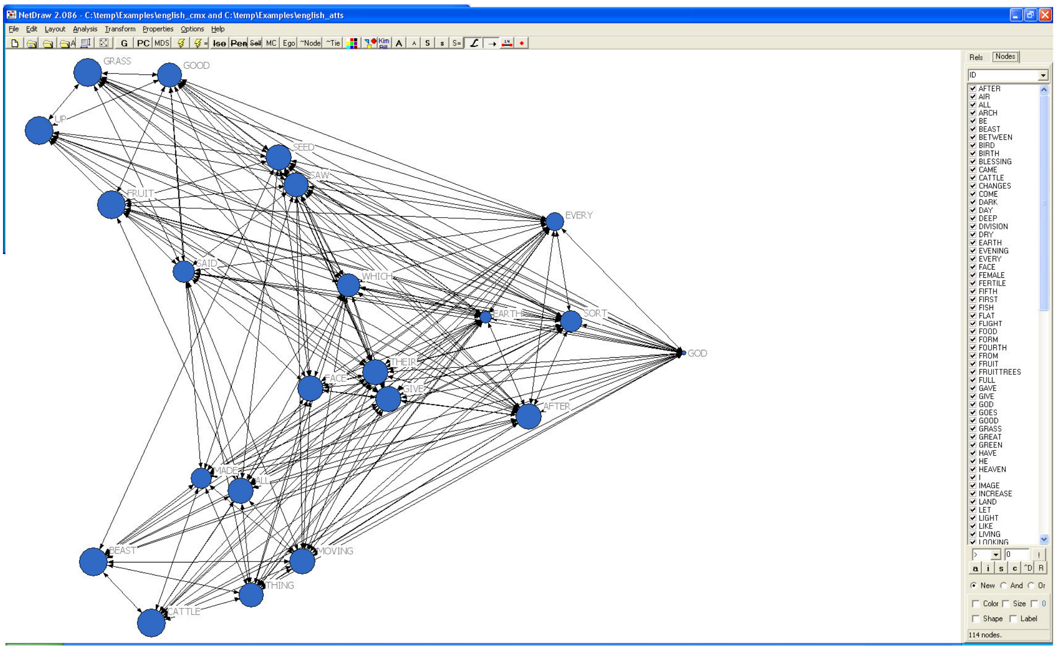
English version of Genesis – Layout 1

This layout shows the various words colored by frequency (similar frequencies have similar colors).



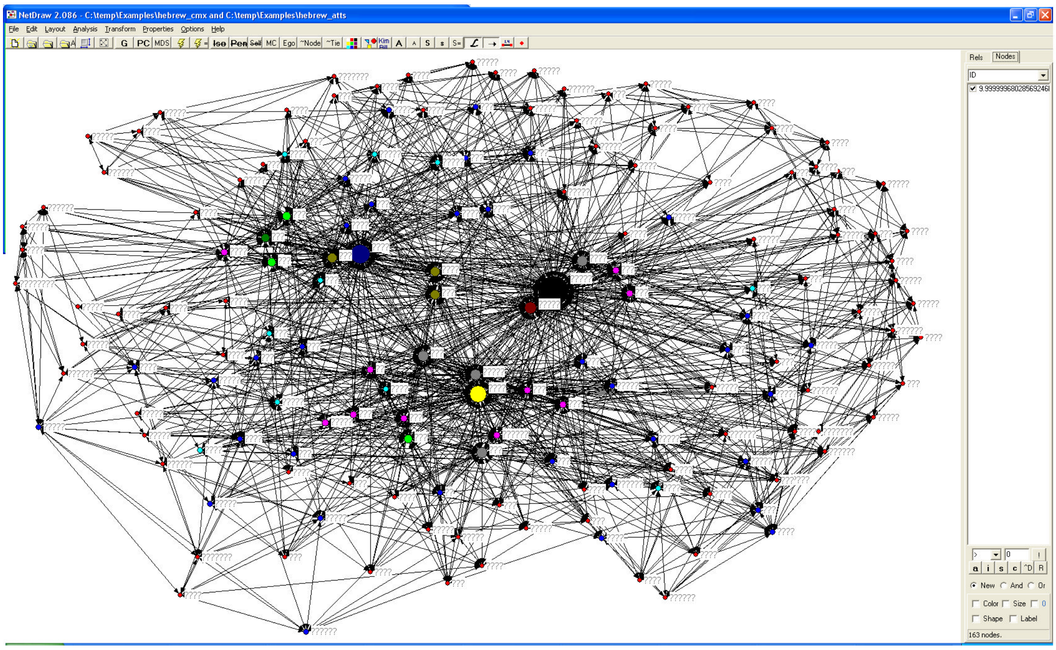
English Version of Genesis – Layout 2

This layout organizes differently but represents the same data as the first layout.



English Version of Genesis – Layout 3 – Ego Network based on the Token ‘GOD’

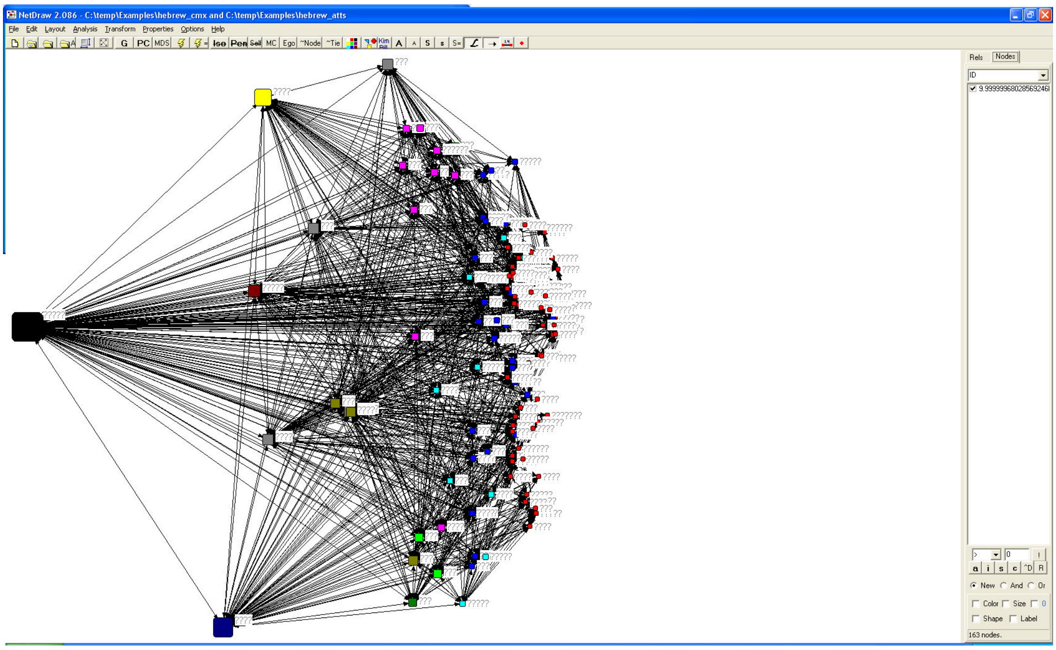
This layout looks at the ego network based on the token ‘God’ as the base node.



NetDraw 2.086 - C:\temp\Examples\hebrew_cmx and C:\temp\Examples\hebrew_atts
File Edit Layout Analysis Transform Properties Options Help
G PC MDS Iso Pen Soil MC Ego ~Node ~Tie A A S S S=
Rels Nodes
ID
☑ 9.99999968028569246
a i s c ~D R
○ New ○ And ○ Or
□ Color □ Size □ 0
□ Shape □ Label
163 nodes.

Hebrew Version of Genesis – Layout 1

This layout represents the Hebrew version. The tokens are improperly labeled due to a lack of UTF8/Unicode compatibility. This layout shows the various words colored and sized by frequency (similar frequencies have similar colors / sizes).


NetDraw 2.086 - C:\temp\Examples\hebrew_cmx and C:\temp\Examples\hebrew_atts
File Edit Layout Analysis Transform Properties Options Help
G PC MDS Iso Pen Sel MC Ego ~Node ~Tie A A S s S=
Rels Nodes
ID
9.99999680285692461
a i s c ~D R
New And Or
Color Size
Shape Label
163 nodes.

Hebrew Version of Genesis – Layout 2 – Ego Network based on the Token ‘םיהלא’ (GOD)

This layout looks at the ego network based on the token ‘םיהלא’ (GOD) as the base node. The layout shows the various words colored and sized by frequency (similar frequencies have similar colors / sizes).

VI Bibliography

Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Brandes, U. and Pich, C., Eigensolver Methods for Progressive Multidimensional Scaling of Large Data. Proc. 14th Intl. Symp. * Graph Drawing (GD '06). LNCS 4372, pp. 42-53. © Springer-Verlag, 2007.

Harris, Z. S. (1957). "Co-Occurrence and Transformation in Linguistic Structure." Language **33**(3): 283-340.

Hildum, D. C. (1963). "Semantic Analysis of Texts by Computer." Linguistic Society of America **39**(4): 649-653.

Woelfel, J. and E. L. Fink (1980). The Measurement of Communication Processes: Galileo Theory and Method. New York, Academic Press.

Zipf, G. K. (1935). The Psycho-Biology of Language, An Introduction to Dynamic Philology. Cambridge, MA, The Riverside Press.