Germanic and its Network: Representing the Germanic Languages Using Median-Joining Phylogenetic Networking

Master's Dissertation

Andrew Peter Fidleir Hall

Student No: 480138

University of the Witwatersrand

Faculty of Humanities

School of Literature, Language and Media (SLLM)

Contents:

Key to Conventions Used in Lexical Structure Formulae

Consonants:

C = consonant

alv = alveolar

app = approximant

cont = continuent

dist = distributed

fric = fricative

gli = glide

intdent = interdental

lab = labial

lab-dent = labio-dental

lat-app = lateral approximant

nas = nasal

obst = obstruent

pal = palatal

pal-alv = palato-alveolar

plo = plosive

postalv = postalveolar

son = sonorant

sib = sibilant

stp = stop

trill = trill

vo = voice

Vowels:

V = vowel

back = back vowel

fro = front vowel

hi = high vowel

low = low vowel

mid = mid-vowel

ro = round vowel

These symbols describe segmental features in the lexical items or parts of lexical items examined for cognacency where they were felt appropriate, such as where similar or related, but not identical, sounds occurred in the same positions across languages. Where the same

sounds occurred in the same positions of languages under examination, standard IPA symbols for specific sounds were used. Standard IPA sysmbols were also used in instances where individual words were transcribed. A plus sign, "+ " before a feature indicates its presence. A minus, "−" its absence. The slash " / " symbol is used in two different ways in this study: (1) the lexical structure formulae, where it appears in the subscript sections in which vowel or consonant features are provided, it indicates that either of two given features occur in the data for the segment in that part of the word or root (for example, [+alv/+pal] indicates the particular segment to which this is attached has either the feature [+alveolar] or [+palatal], depending on the language or variety looked at), and (2) in formulae which show segment conditioning or change it is used as it normally is to indicate "in the environment of" (for example, $C_{[+vo]} \longrightarrow C_{[-vo]}/\_\#$, read as "a voiced consonant is devoiced at the end of a syllable boundary"). An italicised segment in a lexical structure formula represents an instance where varieties have segments which form a class of phonetically or perceptually related sounds but which have a significant number of different segmental features (for example, *r* indicates the presence of a rhotic consonant). When placed around a segment, normal parentheses, (), indicate that the parenthesised segment is widespread but not universal in a set of data items.

Germanic and its Network: Representing the Germanic Languages using Median-Joining Phylogenetic Networking.

Chapter 1

Aim of this study

This study has two main aims. The first is to examine the effects of the use of different lexical items on the generation of a network of the Germanic languages. The second is to determine the effects on network generation of different coding strategies. The model used is the median-joining program Network. The study is ultimately intended to provide information on how applicable this model might be to linguistic data, and to indicate how and under what conditions it may be helpful to linguistic analyses.

Background

The use of a number of bioinformatics tools from the fields of molecular biology and genetics as possible tools for language classification has attracted a degree of interest in recent times. It has, however, received mixed reviews from specialists, with some claiming that they may represent valid ways of classifying languages (Atkinson & Gray, 2005), and others claiming that the use of these methods is highly problematic (Atkinson & Gray, 2005; Heggarty, 2006). This dissertation will examine how the use of different coding strategies for lexical data, as well as different choices of lexical data, will influence the structure of networks generated for the Germanic languages. This study will be based on a study conducted by Forster, Röhl and Polzin (2006), and will use largely the same data, although some varieties of Germanic languages which they did not use, such as Afrikaans and Flemish, will be included. Missing data was handled in one of three ways to determine what the effect of each approach is (removing all missing items, coding missing items as deletions and assigning the most common code to the missing items); different words were also selected for use under different conditions to determine what effects this would have on network generation (for instance influencing where language nodes are placed). It is hoped that, by doing this, potential strong and weak points of the use of median-joining phylogenetic networking will be highlighted, and that some guidelines for assessing this method's accuracy may be arrived at.

The classification of languages can be viewed in many ways as analogous to the classification of biological organisms into groups of species (Atkinson & Gray, 2005). Classification methods and schemes used in diachronic and comparative linguistics are, at an abstract level, based on a number of concepts similar to those used in biology.In both cases comparisons are based on similarities between objects (languages or organisms). In both emphasis is placed on features deemed unlikely to have arisen in parallel, such as conserved

nucleic acid and protein structures in biology, and sound correspondences and shared morphology in linguistics. This has led to a certain amount of interest in using programmes originally designed to compare biological information as a tool for comparing linguistic information (Atkinson & Gray, 2005). One form of model used, which can construct relationships using both inherited and borrowed pieces of information, is termed phylogenetic networking. Phylogenetic networking models are not designed to generate trees or lines of borrowed information, but to look for points of data and connect them in ways which indicate whether or not those individual pieces of data are related to each other; that is, they construct a network of points showing which sequence inputs are related to each other, and what similarities they share (Forster, Toth & Bandelt, 1998). This dissertation will use the program Network (available at http://www.fluxus-engineering.com) to generate a phylogenetic network for Germanic vocabulary items. The network generated by the program will be compared to traditional classifications of the Germanic languages. By doing this, the utility of this method when applied to lexical data can be investigated, as it will allow one to determine how different ways of coding and handling the data may affect the results of the program. This may be useful in determining the reliability of the method. A result which would potentially indicate that the method is reliable would be if the program generated a network in which more divergent languages were found at points further outside the network, and which grouped similar varieties together. If the program were to generate a network which, for instance, placed Gothic alongside Modern English, such a result would be taken to indicate either an unreliable method or a problem with the coding of the data items. This would then be open to investigation. The issue of using these methods to date language divergence will not be investigated due to constraints on the size and scope of this paper. It is hoped that some indication of how reliable this method is for linguistic classification may be reached, and that indicators of future directions in research may be provided.

Rationale

The rationale behind this study is that, theoretically, linguistic and biological data may behave in similar ways. Following this it is thought that, at an abstract level, models designed to handle biological information should be able to handle linguistic information in such a way that they provide useful information regarding how language varieties relate to each other. Forster, Polzin and Röhl (2006) used words from the Swadesh 100-word list, many of which are thought to be less likely to be borrowed at high rates from other languages (this view has been challenged, however (Campbell, 2004)). However, this on its own does not necessarily mean that the words chosen from the individual languages will be historical cognates. One issue with Forster et al.'s (2006) dataset is that some words which are cognate with those in a number of the other languages have not been included as data items. For instance, in Modern English the word *beam* is not included in the dataset at all, despite the fact that it is a modern continuation of Old English *bēam* "tree", and as such is a cognate of the various words for "tree" in the other West Germanic languages. This is despite the fact that its meaning has changed. By not including cognates which have undergone semantic changes, the placement of certain languages within the network may have been oversimplified. Additionally, in instances where data items are missing different ways of handling this situation may have a

sizeable impact on the results. If one is to test the utility of this method such issues need to be investigated.

Chapter 2

Literature Review

Language change can be viewed as more or less analogous to evolutionary change in biology, with a number of linguistic components (such as phonemes, morphemes and lexemes) being viewed in a similar way to biological components, such as nucleic acid base-pairs, genes, and proteins and amino acids. In both instances, change can be largely summed up as descent with modifications (Atkinson & Gray, 2005; Forster, Toth & Bandelt, 1998). This "descent-with-modifications" view encompasses two general forms of change which both languages and genomes undergo: vertical change and horizontal change. Vertical change is change which is language/genome internal, such as a regular unconditioned sound change in a language or a mutation in a gene. Horizontal change is change caused by the introduction of some piece of information from an outside source, exemplified by borrowing between languages in the case of linguistics, and the transference of genetic material from one organism to another (such as a bacterium and a bacteriophage) in the case of biology. The conceptual similarities between language change and biological evolution extend beyond the general way in which change can take place, to encompass a range of much more specific concepts within each field. Atkinson and Gray (2005) have highlighted these similarities in a history of what could be termed "phylogenetic thinking". They (2005) argue as follows: in both instances there exist discrete characters which are open to change. In languages these include phonemes, lexical items, morphemes and syntactic structures, while in organisms these are nucleotides, genes and amino acids. Linguistic cognates, or items which are believed to be related by ancestral descent from a common linguistic ancestor, have their equivalents in the gene homologies of biology, which are those elements of organisms' genomes which are regarded as having descended at some point in the remote past from an ancestral form. Linguistic innovation, which is change in a language variety usually resulting in differentiation from other related varieties over time, is paralleled in biology by mutations, which act to make related genomes and organisms less similar to each other over time. There exist dialect continua and regional language isoglosses, where changes from one variety to another gradually stack up as one moves through the dialect continuum and where uses of language vary geographically. Analogous to these are the geographical clines along which genes in large populations frequently run.

In relation to innovation in linguistics there is borrowing between languages, which has the biological parallel of horizontal gene transfer. Hybrid organisms have their linguistic relative in creoles, which can in many ways be viewed as language hybrids (Atkinson & Gray, 2005). The process of biological cladogenesis, which is the formation of new subgroups of organisms within a family, is analogous to the splitting of groups within language families to form subgroups; both appear to occur along conceptually similar lines (i.e.:

innovations/mutations) (Atkinson & Gray, 2005). It is even possible that the biological idea of environmentally conditioned trait selection has something of a parallel in the uses of different registers, words, pronunciations and varieties: forms which are considered in an unfavourable light in the social environment, as well as those which are under pressure from other varieties for various reasons, can be viewed as being selected against. This is similar to Dawkins' (1976) idea of the meme, which is a piece of cultural information that can be transmitted from one individual to another through social interaction and communication, and which has selection pressures working to either further or stunt its propagation (Millikan, 2004). While biology has the fossils of extinct organisms, as well as those which are ancestors of modern-day creatures, linguistics has records of extinct languages and older forms of some modern languages, as well as archaisms surviving in speech, such as proverbs and set expressions (Campbell, 2004). Even in death there are similarities, with organisms and languages both dying out (Atkinson & Gray, 2005; Hagège, 2011).

Taking into consideration all of these concepts which appear to be shared between the two disciplines, it is not hard to imagine that methodologies designed for studying organisms could be used in the study of languages. Consider, for instance, that methods of linguistic classification are not ultimately very different from methods of biological classification: both essentially look at how similar certain traits are to others in the languages or organism in question. Evidence for relatedness (or lack thereof) stems from comparison of elements considered highly likely to reflect whether or not a given set of languages or organisms stemmed, at some point, from a common ancestor. In the case of languages this evidence is generally garnered from lexical cognates, while in the case of organisms high levels of genome similarity are usually used. In both cases some information is considered more likely than other information to reflect relatedness, such as the usually higher importance which is assigned to closed-class morphological items and sound correspondences in linguistics as opposed to cardinal sentence structure and general typology (Campbell, 2004). In biology greater emphasis is generally placed on genetic similarity as opposed to morphological similarity (Atkinson & Gray, 2005).

The use of computational phylogenetic models is not something new to linguistics (Atkinson & Gray, 2005). The introduction of lexicostatistical methods into linguistics in the early 1950s by Morris Swadesh can be viewed as a forerunner of these methods. Swadesh thought that by analysing lists of basic vocabulary and determining the percentages of shared cognates between two or more languages, distance measures between languages could be arrived at and could be used to group related languages together (Swadesh, 1951; Atkinson & Gray, 2005).Swadesh also proposed, in 1951and 1952 (Atkinson & Gray, 2005; Campbell, 2004; Swadesh, 1951, 1952) that, based on studies of languages and language families with long recorded histories, core vocabulary tended to change with a roughly constant rate of 14% per thousand year period. By extension this information could be used to date the rough dates of divergence between languages. There are a number of problems with this approach. Firstly, it is problematic to assume that all languages have a 14% loss of basic vocabulary every thousand years, as most languages spoken on earth have not left records which could be used to test this (Campbell, 2004). In most cases, if there are written records at all, they date

back only several hundred years, so the "per millennium" part does not hold well. Similar attempts at finding ways of dating languages (such as the use of phylogenetic methods normally employed to date species divergence in biology) have run into a similar problem. In all instances some assumption of a base rate of linguistic change has to be used to make these models feasible (Renfrew & Forster, 2006; Heggarty, 2006). Quantitative phylogenetic methods, developed in biology from the 1960s onwards, provided researchers with increasingly accurate ways of grouping organisms together and providing likely time estimates of species divergence. This resulted in a number of phylogenetic models which, in theory, could be used not only to deal with biological information, but also anything else which could be subjected to phylogenetic analysis or which involved some form of phylogenetic classification scheme; this includes languages (Atkinson & Gray, 2005). However, there was a general lack of interest in such methods in historical linguistics, or scepticism about them, until the late 1980s and early 1990s. In these decades computational phylogenetic methods which were able to cope with large amounts of data were coming into wide use in the field of molecular biology.This had implications for fields of research which to varying degrees overlap with studies of population history, such as archaeology and, to some extent, historical linguistics (Renfrew & Forster, 2006).

Models and methods of phylogenetic analysis which have been used on linguistic information broadly fall into two main groups: tree methods and network methods (Renfrew & Forster, 2006). Tree methods were the earlier of the two. These use algorithms to recover evolutionary trees from a data set by looking for points of similarity and difference, and by calculating the likeliest divergence points in the data based on similarities and differences. This information is then used to construct the tree. There are two main types of tree model, which differ in how they handle data and data representation. The first type uses distance methods, which calculate the overall similarity between data sets, based on the number of characters by which two data sets differ. These use this similarity for tree construction (for instance, neighbour-joining) (Renfrew & Forster, 2006). These are in some ways problematic methods as they do not actually indicate what the points of difference between the different datasets are, and only give an indication of how similar data sets are on average. The second type of tree model uses character-based methods. These utilise the actual linguistic items being used to construct the tree, meaning that the tree can be examined to determine which of the linguistic characters changed along which branches (Renfrew & Forster, 2006). This provides more detailed information for individual characters, and can allow a researcher to pinpoint where changes have occurred and what those changes are, but at the cost of highlighting more general trends; it also potentially produces overly complex trees which are hard to read. A problem with both of these tree models is that they work on the assumption that a tree of divergence points is always going to be present in the data, i.e. they assume that change is lineal, with specific points at which branching occurred, and are unable to indicate to the researcher which items may be borrowings and which not. When data does not behave in a treelike way, it cannot be accurately fed into one of these algorithms without careful consideration and removal of characters which are or may be due to horizontal rather than vertical descent (borrowings) (Renfrew & Forster, 2006). For instance, if German, French, Latin and English lexical data were to be analysed by a tree algorithm, English would be

clustered along with Latin and French, and German would be placed on its own separate branch of the tree, as the large number of English borrowings of French and Latin origin would be analysed as inherited forms.

This would not be a genuine historical classification but a grouping based on lexical similarity. The resulting classification is superficial, does not reflect the actual histories of the languages involved, and obscures the true relationships between them. This problem can be avoided in many instances with careful coding and selection of the data; however, this method itself is open to the criticism that researcher subjectivity could interfere with the coding in the cases of words which have an uncertain history (this criticism is valid for other types of model as well (Heggarty, 2006)). The careful selection of data may be difficult where languages are poorly attested, since removal of certain items may result in data sets which are too small to make an analysis worthwhile. In the case of distance measures, this may result in problems with the degree to which the values used to generate the tree are similar to each other (i.e. languages are treated as being more similar, or more different, than they actually are) (McMahon & McMahon, 2006). Another problem which tree methods face is that of possible parallel evolution of linguistic characters in separate languages or groups of languages. This can sometimes result in tree programmes generating very large numbers of possible trees, which is unhelpful to the researcher. Another problem is that characters may be grouped in ways which do not reflect the actual linguistic history of the varieties concerned (Renfrew& Forster, 2006). For instance, if the palatalization of velar plosives before front vowels in the Romance languages is examined, a tree building programme may produce a large number of trees which show all of the possibilities for how this arose by positing that each of the languages in question is a possible ancestor to the others. It may also imply that this feature was present in their common ancestor, Latin, when in fact it was not (Allen, 1989; Fortson, 2004).

To get around this complication, network models were developed in the early 1990s. Network models do not work on the assumption that there is a tree in the data, but can take data points which appear to behave in a treelike way and construct trees with those, while data points which do not behave in this way result in reticulations (Renfrew & Forster, 2006). These reticulations can be used to indicate where information characters have possibly been transmitted horizontally and not vertically, i.e. they are borrowings and not true cognates. Like tree models, networks can be based on either distance or character data. The problem with distance data in networks is largely the same as in the older tree models: one can see that there are reticulations in the data, and the model is able to present an overall pattern of similarity between two or more sets of data, but the actual characters in the data cannot be viewed, making it harder for the researcher to determine what caused reticulations and splits. This is something which can be ameliorated by character-based models, where the items causing the reticulations and/or splits can be viewed, allowing for a more detailed picture of what is happening in the data (although representing all points of reticulation and/or splitting may result in diagrams which are difficult to read when very large data sets are used) (Renfrew & Forster, 2006).

As explained above, phylogenetic methods, in theory, should be reasonably applicable to linguistic data, as linguistic change and biological change are similar in many respects.Thus, analysis by these methods should not have to be too different, as the algorithms employed are doing largely similar things to pieces of information which behave in similar ways. Nevertheless, there are differences which should be considered in a phylogenetic analysis, such as the fact that a given language can have synonyms, meaning that the use of only one vocabulary item for a given meaning could result in an over-simplified analysis. Certainly these models are potentially valuable ways of processing large amounts of linguistic data which can be used to determine the relationships between languages and language varieties. A study by Forster, Toth and Bandelt (1998) uses a character-based network programme to study the relationships between seventeen Alpine Romance languages; this includes groups which have traditionally been grouped as Rhaetian, such as varieties of Romansh and Ladin, as well as a number of varieties spoken in the Italian Alps and nearby areas, such as Friulian, which are often regarded as varieties of Italian. The modern-day areas in which these languages are spoken (or were recently spoken) are sharply defined geographically, with many of them valleys which are separated from each other by mountainous terrain; this suggests a high likelihood that a number of varieties might be very localised. This gives the authors the chance to see whether or not their model was able to reflect the geographic layout of varieties of the languages under consideration. Their method involves the translation of the Swadesh 100-word list into states which reflect the roots of the various words ("state" refers simply to whether words with the same meaning resembled each other closely enough to be regarded as the same form for that meaning). Lexical items which are binary variables (i.e., have only two distinct states among the languages used, such as the forms for HEAD, which were either of the form *testa* or a variant of *cio/cé/cheu/tgau* (/ʧo/, /ʧe/, /ʧeʊ/ and /ʧaʊ/; from Latin *caput*) are processed without any modification. However, in instances where one language has two synonyms for the given word, the word which causes the least reticulation is used. This is problematic, as it may be quite an arbitrary decision as to which word to use; additionally this can be viewed as creating an artificially simplistic data set which results in a network which may not reflect the actual relationship between the languages examined. However, the authors do suggest that for every non-binary character the most likely evolutionary path should be chosen and then more ambiguous characters should be treated separately in relation to the overall network (Forster, Toth & Bandelt, 1998). Classical Latin is used in the network as an outgroup; this allows for the network to be rooted, and may act as an indicator as to whether the network is performing an inaccurate analysis (for instance, if the network were to show Latin as a descendant of one of the other languages, it would be quite clear that it was either doing something it was not supposed to, or that there was a source of confusion in the data which would then be open to investigation). By providing an outgroup, the network could be used to determine which was the oldest node in the network relative to the others and then hypothesise both the likeliest ancestral node and ancestral states (Forster, 2006). The resulting network is strongly treelike and closely matches the geographic layout of the Alpine Romance varieties, with linguistic subgroups reflecting geographical grouping. It further indicates likely borrowings between and within subgroups, and also indicates hypothetical ancestral language nodes for each subgroup; these are near the base of each subgroup, suggesting that the features of those nodes are largely shared by other

nodes in the respective subgroups. Finally, the Classical Latin outgroup links to the centre of the entire Alpine Romance group, indicating that they descended from a variety closely related to Classical Latin; this could be interpreted as due to differences between Classical and Vulgar Latin, with the Alpine languages descending from the latter (Forster, Toth & Bandelt, 1998). The network also revealed high levels of variation between the Alpine Romance languages which, according to Forster, Toth and Bandelt (1998) were comparable to, if not in some instances greater than, the differences between Italian, Spanish and French; this has quite obvious implications for the classification of these varieties, namely that a number of them are separate languages rather than dialects of the same language.

A later study by Forster and Toth (2003) uses the same method to examine the relationship between Gaulish, the Celtic sub-branch of Indo European, and Indo-European itself; this is further used to attempt dating roughly when the different splits took place. The data consists of glossed words from bilingual Latin-Gaulish inscriptions, along with the same words and/or constructions in Classical Greek, Old Irish, Modern Irish, Modern Scots Gaelic, English, Welsh, Breton, French, Occitan, Italian, Spanish and Basque (which was presumably included as an outgroup). The data analysis focuses mainly on characters which are binary; the rationale behind this is that less variable characters which are spread over a large area would be more likely to reflect any genetic relationships between the languages. Binary state characters are the first to be processed so that initial network complexity is reduced for the sake of workability. After this phase, multistate characters (those characters with more than two forms across the languages being examined, for example, DAUGHTER, which has six different word forms representing the concept across all the languages) are introduced into the network. To avoid creating an overly complex network too early on, the authors use a system whereby multistate characters are treated as binary as they are introduced into the network. This is done by taking the split which partitioned the largest group of linked nodes in the network and then subsequently introducing splits which partitioned sequentially smaller groups of nodes. For instance, because the forms grouped under the letter *b* (*filia/fille/filha/figlia*) make up the majority of forms representing the concept DAUGHTER in the data set, these are split off from the rest of the terms as one large group; the sets of terms marked *a* (*duxtir/daughter/θυγατηρ*, [t^hugate:r]) as the next largest set are split off next from the remaining data set; this carries on until those sets which contain only one term are left. Any items which have more than five different character states (or forms) across the languages are omitted from the analysis as they contribute to very high levels of reticulation and make the network unwieldy. Unlike the research done on the Alpine Romance languages, an element of syntax, namely cardinal positions of verb and subject in relation to each other, are included in the final phase of the analysis. Suffixes are also included. If any character causes a disproportionate level of reticulation it is removed and another character is used instead (Forster & Toth, 2003). While this is most probably due to issues of practicality, it does raise the question of what criteria were used to determine when reticulation is too extreme, as these are not explicitly stated. The resultant network groups all of the Romance languages used in the study together as descendants of Latin, while all of the Celtic languages used in the study emanate from one node. The internal branching of the Celtic languages is somewhat in agreement with more traditional classifications of them. The Celtic languages

are usually partinioned into Continental and Insular Celtic, with the former embracing Gaulish, Galatian and Celtiberian, and the latter splitting into a Brythonic branch (comprising Welsh, Breton, Cornish and Cumbric) and a Goidelic branch (comprising Irish Gaelic, Scottish Gaelic and Manx) (Fortson, 2004)). The branching in Forster and Toth (2003) has a Goidelic branch and a Brythonic branch occurring separately to Gaulish, which could be seen as a Continental branch. In addition to this the network shows Latin, Celtic, Greek and English all emanating from a single ancestral node; this appears to indicate that a distinct Italo-Celtic branch did not actually exist, contrary to some classifications (Barber, Beal & Shaw, 2012; Fortson, 2004). However, the authors point out that the word list they use is very limited and mostly does not consist of items of basic vocabulary, which are thought to be less likely to be regularly borrowed than more specialised terms. The network also places the older forms of the Celtic languages closer to the common ancestral node than their modern counterparts; this could be taken as an indication that the method is reliable. The authors then attempt to date the splits in the network. This is done by taking the number of lexical replacements from the data over the recorded time of the languages' existence, and then averaging this to produce lexeme mutation rate (Forster & Toth, 2003). The tree is then rooted (presumably by using the Basque data as an outgroup), and the complete branch lengths are normalised to AD 2000; this yields dates of 8100BC $\pm$ 1900 years for the breakup of the languages involved from the common node, and 3200BC$\pm$1500 years for the splitting up of Celtic. However, these dates should be viewed critically for a number of reasons. These include the small number of data characters used, the small set of Indo-European languages used (statistically, having a smaller number of characters and data sets would increase the influence of data items which could act as outliers, skewing the averages) and the possibility that rates of change might (and probably did) differ at different times within branches.

A similar study by Forster, Polzin and Röhl (2006) on the evolution of English basic vocabulary within the Germanic languages uses a median-joining model (Network 4.106) to analyse lexical data from English, two varieties of Old English (Beowulf and Alfred the Great), Old Saxon (with data taken from the Heliand manuscript), several dialects of Modern Frisian, as well as Low German, Dutch, Standard High German, Bavarian, Swiss German, Danish, Faroese, Old Norse, Icelandic, Norwegian Nynorsk, Norwegian Bokmål, Swedish and Gothic. Median-joining algorithms identify groups of closely related data points and introduce hypothesised ancestors in order to create a complete network or tree (Forster, Polzin & Röhl, 2006). The lexical items used are the items for each language on the Swadesh 100-word list; thus they are items which are generally thought to be less likely to be replaced in large enough numbers in a short enough time to obscure the relationships between the languages (this has been challenged, however (Campbell, 2004)). List items which do not have representatives at all in any of the languages are removed, resulting in sixty items being analysed. The items are coded in the same way as amino acid codes, which allows for multistate coding without the "false binary" method used in the study on Celtic. One problem with this is the fact that some languages have synonyms for items in the Swadesh-100; Forster's recommendation is to use the most frequently used one (Forster, personal correspondence) because two amino acid codes cannot be used at once for the same

character. The resulting network only partially agrees with traditional classifications of the Germanic languages. In these there are three main sub-branches, East (Gothic and a number of scantily attested languages), West (English, Frisian, Dutch, Low German and High German) and North (Icelandic, Faroese, Norwegian, Danish and Swedish) (Fortson, 2004). In this study they are instead grouped into four branches: Gothic, North Germanic, English and a German branch, which also includes Dutch and Frisian (Forster, Polzin & Röhl, 2006). The network also groups Old Norse with modern Icelandic, indicating high levels of sequence similarity between them, although the other old Germanic languages are grouped much more closely to the centre of the network. Interestingly, modern English is shown as being distantly related through a series of complex reticulations to the Old English varieties of Beowulf and Alfred the Great; this seems to agree to some degree with suggestions that modern Received Pronunciation English is descended from an Anglian dialect of Old English not represented in any known surviving manuscripts, with evidence of some dialect mixing with the southern varieties of Old English (that is, West Saxon and Kentish) (Barber, Beal & Shaw, 2012). An attempt was made at dating the common ancestor of the Germanic languages, yielding a time bracket of 3950 years (between 3600 BC and AD 350). However, there were some questionable character inclusions and exclusions in the data (for instance, Old English *docga*/*dogga* was not included alongside *hund*, and the study did not take into account the fact that much language in Beowulf is used metaphorically, rather than literally, and that much vocabulary may be poetic and not a good representative of everyday speech; for instance the word *guma* was already fairly archaic in the Old English period (OED, 2015)). Additionally the tree was unrooted, which should theoretically reduce the strength of date calculations anyway. Despite the above, most of the criticism of Forster, Polzin and Röhl's 2003 and 2006 studies has revolved around their attempts to date the splits within networks and the emergences of various linguistic varieties (Heggarty, 2006; McMahon & McMahon, 2006); these have focused on the fact that to do this some assumed rate of change has to be used, which is problematic, as a given language may change at a different rate to another language. The problem of a universal rate of change across a language group could be ameliorated by having different rates of change for different branches. However, this would still leave the problem of variable rates of change within one branch (or even one language) at different points in time (Heggarty, 2006; McMahon & McMahon, 2006). There are a number of issues which need to be further investigated regarding the use of these models. To do this, the current study will expand on Forster, Polzin and Röhl's (2006) study; the details of this are explained under the "Method" section.

Chapter 3

On the Organisation and Classification of the Germanic Languages

The Germanic languages are a group of related languages which constitute a branch of the larger Indo-European language family, and which are spoken as a first language by roughly 400 million people (Barber, Beal & Shaw, 2012). One of the main characteristics which divides the Germanic languages from other Indo-European languages is the phonological

chain shift of Grimm's Law (also referred to as the First Germanic Sound Shift; Nielsen, 1989). This resulted in the original Indo-European system of voiceless, voiced and aspirated voiced stops changing their qualities so that the original voiceless stops became fricatives, the voiced stops became voiceless, and the voiced aspirated stops lost their aspiration (Barber, Beal & Shaw, 2012; Fortson, 2004). This is summarised below:

	IE stops		Germanic obstruents
I.	**p, *t, *k, *k^{w}*	⟶	**f, *θ, *x~*h, *x^{w}~*h^{w}*
II.	**b, *d, *g, *g^{w}*	⟶	**p, *t, *k, *k^{w}*
III.	**b^{h}, *d^{h}, *g^{h}, g^{wh}*	⟶	**b~*β, *d~*ð, *g, *g^{w}*

For example, the descendants of the Proto-Indo-European root **ph$_2$tēr*, "father", are Latin *pater*, Classical Greek *patḗr,* and Sanskrit *pitár*, while Old English has *fæder*, Old High German *fater*, Old Frisian *feder*, Old Saxon *fadar*, Old Norse *faðir*, Gothic *fadar* (vocative). This set of words would give something like **fader*, [fader/fadɛr] or **faðer*, [faðer/faðɛr] in Proto-Germanic. Similar patters can be observed across the other two sets of stops, such as Proto-Indo-European **po**d**-*, "foot", continued in Latin as *pe**d**-* and Greek *po**d**-*, but in Old English, Old Saxon and Old Frisian as *fō**t***; and Proto-Indo-European **bher*, with [b^{h}], continued in Sanskrit as *bhar-*, [b^{h}ar], but in Gothic as *bairan*, [bɛran], Old English, Old Saxon and Old High German *beran,* [bɛːran], and Old Frisian and Norse *bera*, [bɛːra]. Some exceptions to this sound shift occurred if one of the Proto-Indo-European voiceless stops was preceded by the voiceless alveolar fricative [s], in which case the stop retained its phonetic value, or if two of the Proto-Indo-European voiceless stops occurred together as a cluster, in which case the first would be subject to Grimm's Law and not the second (Fortson, 2004). Related to Grimm's Law is a sound change which is usually referred to as Verner's Law, and which acted on the series of fricatives produced by Grimm's Law acting on the Proto-Indo-European voiceless stops. Here, the fricatives [s], [f], [θ], [h] ~[x], and [h^{w}] ~ [x^{w}] became voiced if they occurred intervocalically or before a syllabic sonorant, and were not preceded by a stressed syllable. For example, pre-Verner Proto-Germanic **uféri*, [uˈfɛri] underwent the change $C_{[-vo,+fric]} \longrightarrow C_{[+vo,+fric]} / \sigma_{[-stress/accent]}\_\sigma$ to give post-Verner **uƀéri*, [uˈβɛri] Another distinguishing feature of the Germanic languages is the dramatic reduction in the number of noun cases: where Indo-European had eight distinct cases (nominative, accusative, genitive, dative, locative, ablative, vocative and instrumental), Germanic had four (nominative, accusative, genitive and dative), although the presence of a vestigial instrumental in Old English (which came to merge with the dative) and a vestigial vocative case in Gothic suggests that, at least in its early stages, Proto-Germanic may have had six cases (Nielsen, 1989); it is possible that these vestigial cases were restricted to much smaller numbers of nouns, and were possibly archaic even in Proto-Germanic. Number was also reduced in Germanic, from the three numbers of Indo-European (singular, plural and dual) to two (singular and plural), with the dual only surviving in the pronouns, except in Gothic, which has preserved the dual conjugations for verbs (Fortson, 2004; Nielsen, 1989). Additionally, the tense system of Indo-European, which distinguished between the present, the imperfect,

the aorist, the perfect and the future, underwent significant simplification, with Germanic verbs only marked for the present tense and the preterite. The verbal mood system was also simplified in Germanic, with the indicative, imperative and subjunctive,[1] where Proto-Indo-European had the indicative, the subjunctive, the optative and the imperative (Nielsen, 1989). The Germanic languages are further distinguished from the other Indo-European languages by the class of weak verbs, which arose as an innovation peculiar to the Germanic languages and which make use of an alveolar stop morpheme to mark the preterite (Fortson, 2004; Nielsen, 1989); these existed and still exist alongside the strong verbs, which make use of alternations in the stem vowel to mark present versus past tense (e.g.: English *jump* (present tense), *jumped* /ʤʌmpt/ (past tense) versus *run* /ɹʌn/ (present tense), *ran* /ɹæn/ (past tense). A further characteristic of the Germanic languages was a weak and strong declension for certain adjectives; strong adjectives have endings which are the same as those of the demonstrative pronouns, while weak adjectives either end in an unstressed vowel [ə] when preceded by an article in the nominative or accusative case, or a suffix derived from Proto-Indo-European *-*on* when preceded by an article or pronominal adjective in the dative or genitive case (such as in German *das brav**e** Kind*, "the good child", weak declension, versus *ein brav**es** Kind*, "a good child", strong declension; *mit einem brav**en** Kind*, "with a good child", weak declension) (Fortson, 2004).

It is generally agreed that there are three main groups of attested Germanic languages, usually grouped by giving them the names of three of the cardinal directions on the compass, namely North Germanic, West Germanic and East Germanic (Crystal, 2005; Fortson, 2004). The North Germanic group comprises Swedish, Danish, Norwegian, Icelandic and Faroese; traditionally an east-west division has been made within the North Germanic language group, with Swedish and Danish falling under the eastern division, and Norwegian, Icelandic and Faroese under a western group (Crystal, 2005; Barber, Beal & Shaw, 2012; Fortson, 2004). The West Germanic languages are High German, Dutch, English, the Frisian varieties and the Low German varieties spoken in northern Germany. East Germanic is only attested to any appreciable degree by Gothic, which was spoken by the Ostrogoths and Visigoths in the lower Danube region and up into parts of the Balkans, the Iberian Peninsula and Italy, where it is generally thought to have become extinct in the eighth or ninth centuries; a variety of Gothic referred to as Crimean Gothic appears to have been spoken in the Crimea up to the sixteenth and possibly early seventeenth centuries, and has been preserved in the form of several wordlists and a single song (Crystal, 2005; Barber, Beal & Shaw, 2012; Fortson, 2004).

The existence of North and East Germanic as groups is not in question. However, the relationships between these two, and West Germanic, as well as the actual status of West Germanic as a valid grouping, have been the subject of some debate in the past. The first of these questions will be treated below. Jacob Grimm, the founding father of the comparative method and one of the creators of historical linguistics as a serious, academic discipline, proposed a close relationship between Gothic and High German; however, he noted connections between the other West Germanic languages (in his publication *Deutsche*

[1] Which in Germanic did not descend from the original Indo-European subjunctive, but from the optative.

Grammatik, from 1840, these were Anglo-Saxon, Dutch, Frisian and Low German) and the North Germanic languages, which included similarities in vocabulary and the forms of certain morphemes (Nielsen, 1989). This, he proposed, was probably due to contact between each group; he thus suggested that a single language area which incorporated High German (and presumably the other West Germanic languages) and Gothic was broken by the influence of the North Germanic languages on those varieties of West Germanic which lie along the North Sea coast (Grimm, 1840; as cited in Nielsen, 1989)[2]. This line of thought, which suggests a closer degree of familiarity between High German (and presumably at some point the rest of the West Germanic group) and Gothic, was not shared by Holtzmann (1839, 1870; as cited by Nielsen, 1989) who regarded Icelandic as sharing the most similarities with Gothic, suggesting that in fact the North Germanic languages are closer relatives to Gothic than High German. However, it must be noted that his treatment seems to have been geographic, with the similarity between given varieties of Germanic increasing as one moves in a north-easterly direction from the southern bounds of the German-speaking areas of Europe to Scandinavia (Nielsen, 1989); it seems Holtzmann conceived of the Germanic languages as being arranged rather as a dialect continuum, as opposed to distinct separate groups. One of the first groupings of the Germanic languages into three distinct groups, which could be viewed as distinct clades in analogy with biological classification, was proposed by August Schleicher in 1860; this was also one of the first serious uses of the family-tree model with regards to the Germanic languages (Nielsen, 1989). This model represents the Germanic languages as descending from a single common ancestor (a proto-language) and then branching out at the same point in the tree into three groups, namely Gothic, Nordic (embracing the North Germanic languages) and German (which embraces the West Germanic languages). While this representation certainly does reflect the major differences which distinguish the Germanic languages as generally falling into one of three groups, it has the problem that it suggests that each group is equidistantly related to the other two. Although this is not an impossible situation, it is an unlikely one, and unfortunately this conception of the languages rather simplifies the relationships between these groups. For instance, there are a number of similarities between the North Germanic and West Germanic languages which are not present in Gothic and which are great enough to point to Gothic (and probably also the other East Germanic languages, which are not well attested) being more distantly related to them than either is to the other, for instance the presence of functional clitic particles in second position in Gothic sentences (which are present in other Indo-European languages, and which follow Wackernagel's Law, just as those in the other Indo-European languages do) (Fortson, 2004); these are not found in any of the other Germanic languages), full dual conjugations for verbs (in the North and West Germanic languages the dual exists only in the first and second person pronouns (Fortson, 2004)), productive use of reduplication in the past tense of class VII verbs, a fully productive passive voice for verbs (in the North and West Germanic languages the passive exists only in the single fossilised verb "to be called" (Old English *hāttian*, Modern High German *heißen*, Modern Dutch *heten*), lack of Verner Variants in Gothic (Old English *wearþ*, "becomes", *wurdon*, "they

[2] Owing to difficulties obtaining certain primary sources, much use of secondary, cited sources must be made.

became", versus Gothic *warþ* "becomes", *waurþum*, "they became") and the lack of umlaut (Gothic *fōt*, "foot", *fōtjus*, "feet", versus Old English *fōt*, "foot", *fēt*, "feet") (Fortson, 2004).

In addition to the above information, the existence of runic inscriptions from southern Scandinavia (predominantly Denmark and southern Sweden) which appear to show features of North Germanic and West Germanic (such as the sound *ā* instead of Proto-Germanic **ē* and the rhoticisation of Proto-Germanic **z* to *r* or *R*[3]) suggests that the two groups of languages may have had a common ancestor (Nielson, 1989, 2000). Makaev (1965; as cited in Nielsen, 1989) has suggested, based on the evidence of other inscriptions which appear to have Gothic features, and which come from different areas, that a primary split in the Germanic languages is one between Gothic (or more probably, East Germanic) and North-West Germanic. The idea that the North and West Germanic Germanic languages should be more closely related to each other than to the East Germanic languages has also been suggested by Voyles (1968) who lists thirteen phonological parallels between the West and North Germanic languages, such as Proto-Germanic **x* or **χ* becoming [f] between **a, *o, *u, *w* and **l, *n, *r*, as well as **g* being lenited to [w] between **a, *o, *u, *w* and **m*, in opposition to the single feature shared between North and East Germanic of the fortification of geminated glides to geminated stop-glide sequences. The fact that so many correspondences are shared between the North and West Germanic languages is taken as evidence that they are part of a single North-West Germanic group. Bahnick (1973; as cited in Nielsen, 1989) has suggested a grouping, based on phonological and morphological criteria, in which Proto-Germanic divides into Gothic on the one hand and Northwest Germanic on the other; this Northwest Germanic node, which is drawn with a division within it (presumably this is to indicate two separate dialects), divides into Common Nordic (from which the North Germanic languages all descend) and three other nodes (these descend from the side opposite the internal division to Common Nordic), namely Old English, Old Saxon and Old High German. Old English and Old Saxon are shown as having been influenced by the dialect of Northwest Germanic from which Common Nordic descends.

Within the West Germanic languages there has been some debate as to whether or not English and the Frisian languages form a distinct subgroup, known as Anglo-Frisian, or are part of a larger, less defined group which has been termed North Sea Germanic by Maurer (1952; as cited in Nielson, 1989) or Ingvaeonic by Wrede (1919; as cited in Nielsen, 1989). North Sea Germanic and Ingvaeonic have both been taken to include not only English and Frisian, but also Low German; thus, West Germanic in this view would not be broken up into High German, Low German, Dutch and Anglo-Frisian, but would have a tripartition into High German, Dutch and North Sea Germanic. The argument used to posit an Anglo-Frisian subgroup is based on the fact that Old English and Old Frisian show a large number of similarities to each other, and have a number of features which are not seen in Dutch, High German and many varieties of Low German. These include the presence of only one plural form across the different persons (first, second and third) for verbs in the present and preterite, the loss of the alveolar nasal [n] before spirants (with nasalisation and

[3] The letter *R* has traditionally been used in Old Norse studies to represent a phone with a value somewhere between [z] and [r].

compensatory lengthening of the preceding vowel), the fronting of West Germanic **ā* to *ǣ* [æː] (sometimes termed Anglo-Frisian brightening), and the palatalization of West Germanic **g* and**k* to [j] and [ʧ] before front vowels (Fulk, 1998). Wrede (1919; as cited in Nielsen, 1989) has argued that Old Saxon shares in a number of these innovations, such as the loss of the alveolar nasal before spirants and the uniform plural endings on verbs; he also argues that a piece of evidence supporting the notion of Old Saxon being part of a group with Old English and Old Frisian is the forms of the accusative and dative pronouns, which unlike High German frequently have the form CV rather than CVC. This last piece of evidence may be viewed as questionable, as pronouns with the form CVC are attested in Old English (especially poetic or archaic texts, for instance the inscription on the Alfred Jewel, which runs *Ælfred mec heht gewyrcan*, "Alfred had me made", where *mec* is the first person accusative pronoun), raising the possibility that the dropping of consonants in the pronouns may have developed in parallel, with unstressed pronouns undergoing this change first, before it became the norm (compare the presence of "weak" pronouns in Dutch, which generally occur when the pronoun is unstressed, for example, *jij*, "you"(stressed form) versus *je*, "you" (weak form); these are coming to be used in contexts in which the pronouns are not stressed (Stern, 1984)).

It has generally been the case that theorists have grouped English with the other West Germanic languages; however, Forster, Polzin and Röhl (2006), used networking software to compare vocabulary from a number of Germanic languages and their varieties had English grouped quite far away from the other West Germanic languages, in fact positioning it as an outlier on the network, jutting out from a fairly reticulated section of the network lying midway between the North Germanic languages and the West Germanic languages. Interestingly, Gothic was placed closer to the West Germanic languages than English was. These results may be due to the data used in the study, and also possibly to how it was coded. It may in fact be that the network has grouped the Germanic varieties based on lexical similarity. However, the fact that English was grouped in an outlying position, and emanating from a reticulation which was intermediate to the North and West Germanic languages, suggests that, provided one is careful with how the relevant data is used, this type of analysis could be informative as to some aspects of a language's history. By including additional lexical data in the form of vocabulary items from Afrikaans, Flemish and Proto-Germanic, the degree to which this may be the case can be tested. This will be investigated.

A Brief Explanation of Median-Joining

Quite naturally if one wants to examine the potential of a median-joining model such as the one used in the above study, it will be necessary to include an explanation of how the median-joining algorithm actually works. Median-joining algorithms are a subset of algorithms known as maximum parsimony algorithms (Bandelt, Forster & Röhl, 1999); maximum parsimony algorithms calculate the most parsimonious path or relationship

between different sets of data (in relation to any constraints placed on the data). Median-joining is built on the use of median vectors, which are points lying at a median distance between three sequences; the sequence which has the most in common with the other two sequences is made the internal node of the tree, with the other two branching off from it; because of this, there need to be at least three data sequences to be of any real use (Bandelt, Forster & Röhl, 1999)(if only two sequences of data are fed into the algorithm, the resultant tree will simply have two points joined by a line).

Connections between sequences are generated using median vectors in a process known as median generation, in which median points along lines connecting three data points are found and used to shorten the distances between these points. The actual connections generated between sequences and their lengths are calculated using what is known as the "Hamming distance"; this is merely the summation of the number of different characters within each sequence, with the number of different characters being proportional to the length of the connection between sequences. Median generation will occur only within a triplet of sequences which have at least two feasible links (Bandelt, Forster & Röhl, 1999; Röhl, personal correspondence); fewer than two feasible links and median generation is unnecessary.

Data must be aligned before it can be fed into the model; thus, points which correspond with each other must be indicated as such otherwise they will not be read by the algorithm as corresponding points of data (Forster, personal correspondence). The algorithm is based on the assumption that for any of at least three sequences of data, there will be just one tree connecting them (this is assuming the three sequences are comparable sequences of data, and are not completely unrelated, for example: three separate languages would be comparable, while two languages and a DNA sequence would not be). This means that, in theory, there must be one single "most parsimonious" tree for the given minimum three data sequences, i.e. a tree made in which sequences are linked by the shortest of the possible paths generated (Bandelt, Forster & Röhl, 1999). To generate such a tree, the majority state from a group of states for a given data character is taken as the "ancestral" state. Thus, if three languages are analysed and one of them has a word Y where the other two have a word X for a given concept, the algorithm would group the language with Y as being a more distant relative of the two X languages *for this data character*; this is based on the assumption that it would be more likely for a character X to change to Y in one language than a character Y to change to X in two languages, as one step is more parsimonious than two (this is a purely statistical assumption, and this example does not include more detailed linguistic knowledge, which would be applied to an analysis which was being used to seriously investigate the relationships between varieties of languages).

Data characters which are ambiguous (i.e. have an uncertain history) may be removed from an analysis (although this may be less interesting in the case of programmes which display data points on the network) or they can be included, in which case the algorithm highlights the possibility of two different origins for that character by creating a reticulation in the network (joining that particular point to the network in such a way that it joins to two separate nodes in the network, where each node represents a language). Character ambiguity

could be due to there being more than one form of a given character (such as Old English *eaxl* and *sculdor*, both of which mean "shoulder", but have different continuations in modern English in *axle* and *shoulder*, respectively), particularly when it is uncertain which form to prefer in an analysis of linguistic relationships (for instance, when both terms are native terms in a language), or ambiguity can refer to instances where data is missing or a particular form is thought to have existed, but is unattested. To a large extent ambiguous data has to be handled at the level of data coding, which means that instances where there is ambiguous data in the set of sequences (in this case languages) will most probably be very sensitive to how the individual researcher handles these ambiguities. Thus great care is needed when coding this data.

Whether or not more than one potential connection between sequences of data is generated and incorporated into the developing network depends on the value of a particular parameter known as ε. ε is a tolerance parameter which indicates how strongly distances between characters are to be distinguished (Bandelt, Forster & Röhl, 1999), and which has the effect of controlling the number of feasible connections created during the median generation process; the inclusion of ε is important because without this control, the algorithm would attempt to generate a network made up of all possible relationship trees. Generally, this number is so large that it cannot be computed (Röhl, personal correspondence). A low setting of ε will have the effect of reducing network complexity by causing the algorithm to assign ambiguous character states to those of any other characters which are the most similar to the ambiguous characters and which are not ambiguous themselves, and will result in a network which does not fully explore all feasible connections, as only the shortest connections as defined by the particular setting of ε will be generated. For example, if five languages are being examined, and each language is treated as a sequence made up of lexical data, setting ε to zero will result in only the shortest possible links between the languages being calculated and generated. Increasing the value of ε will result in potentially more network complexity as the algorithm is less strict in terms of how it assigns ambiguous states, resulting in it generating more links between sequences in addition to the connections which would have been calculated if the parameter ε had a low setting. This is because the higher the value of ε, the larger the number of links that are generated and treated as being equal (even though, in reality, they may technically not be). Thus, higher values of ε will result in a greater likelihood that characters which are ambiguous and may have more than one connection within the network will be identified as such, and that they will be incorporated into the network as reticulations. For example, if one considers a character X which is ambiguous, a low setting of ε will result in fewer feasible connections between the sequence in which X occurs and other sequences being calculated than a higher setting would. However, higher settings of ε, while possibly allowing for greater accuracy in a phylogenetic analysis, require more calculation time (Bandelt, Forster & Röhl, 1999). An example of the effects of ε might be seen if one were to compare English vocabulary with vocabulary from both the North Germanic languages and the other West Germanic languages; due to the large number of Old Norse words which have entered the English language and replaced the Old English words in general speech, a low setting of ε (such as zero) would increase the likelihood that modern English would be represented as being closer to the North Germanic languages than it

actually is, due to fewer scenarios being explored by the model, as only the shortest paths between English and the North Germanic languages would be calculated (possibly English would be represented as one of the North Germanic languages, or as a very close relative, with the other West Germanic languages being shown as more distantly related to English). Higher settings of ε would increase the likelihood that the words of Old Norse origin would be shown as reticulations in the network between English and the North Germanic languages, indicating that they may have been borrowed (but not showing direction of borrowing). The actual value of the parameter ε is thus important to take into account, as values which are too low may result in too small a number of feasible connections within the data being explored, while very high values might result in too many links being calculated, which may be a waste of time and result in unnecessary complexity.

The actual connections generated during median generation are determined by connection cost. Connections will not be generated if their connection costs (which can be viewed as somewhat representative of the probability of a given connection) exceed the value of $\lambda + \varepsilon$, where λ is the minimum connection cost for all feasible triplets, and ε is the value to which ε was set at the beginning of network generation. At the end of median generation (which may be done a number of times) ε is set to zero again and the most *statistically* likely connections (ignoring likelihoods which the linguistic environment throws up) are calculated from the set generated by the earlier median generation and used to construct the network (Bandelt, Forster & Röhl, 1999; Röhl, personal correspondence). It must be stressed that the algorithm does not detect the correct connections between sequences, but, depending on how ε is set, will highlight more than one potential path of connection between sequences, which can then be investigated (for example, it will not be able to detect that English *leg* is a loan from Old Norse, rather than a true cognate of Old Norse *leggr*, but it will be able to show that English *leg* could be either a borrowing or a cognate of the Old Norse word, depending on the value of ε).

The above process will be used to analyse the lexical items which constitute the data in this study. The results of the different simulations (all languages together, only contemporaneous languages, words with the same meaning, cognates regardless of meaning) will be compared to those of Forster, Polzin and Röhl (2006) and to each other. This should confirm or disconfirm the accuracy of the original study, and provide an idea as to the utility of median-joining phylogenetic networking as a method of language classification, as well as providing some indication of the sensitivity of the program to the coding of data.

Method

For this study, the networking program Network 4.6.1.3, created and distributed by Fluxus Engineering, was employed to carry out simulations of the data. Sixteen simulations were performed, under a number of conditions and with slight differences in datasets to determine the effects of this on the representation of the relationships between the different Germanic languages and Germanic sub-groups. The data consisted of lexical data which was compiled

using the Swadesh 100-word list, from a number of different sources, including the Oxford English Dictionary (OED), the Svenska Akadamiens Ordbok (SAOB; the Dictionary of the Swedish Academy)), the Geïntegreerde Taalbank (GTB; an online Dutch and West Frisian etymological dictionary), the German Duden; the Ordbog over det Danske Sprog (Dictionary of the Danish Language; ODS); Vladimir Orel's (2003) *A Handbook of Germanic Etymology*; Forster, Pölzin and Röhl's (2006) dataset, and a number of individuals at various institutions who helped with the etymologies of words.[4] The datasets for the Old Germanic languages were not complete, as certain words for some of the concepts on the Swadesh 100-word list are not attested in the literature and records of these languages. This resulted in the use of three different coding strategies to handle this missing data. The first of these was to code data under a Majority Wins condition, in which missing data items were simply assigned the same code as the most common form in the data. This was helpful in that it allowed the size of the dataset to be maintained, and was based on the assumption that the probability of a common term existing in the language with the missing item as well is reasonably high. This is however a problematic strategy because, despite allowing the size of the dataset to be maintained, it is quite possible that a cognate word was not used in the language in which the necessary data item is missing and so there may be a risk of underestimating the degree of divergence between languages; the likelihood of this is something which has to be judged according to the data available. As the Old Germanic languages share a very large number of cognates, it was judged that the underestimation of their divergence would be likely to be minimal using this strategy.

Due to the possibility of underestimation, however, two other simulation conditions were also employed. One was termed the Infodelete condition. In this condition missing data items were coded as deletions in the data set. This was done to allow the size of the dataset to be maintained while not resorting to assuming that the most common form for the missing data item must have occurred in the given language. This is problematic itself, as the program lists a deletion as a change *ipso facto*, and so this also carried the risk of influencing how the program calculated the network. The second additional condition was called the Infoclean condition. Under this condition, characters (concepts on the Swadesh list) for which some languages had missing items were simply removed. This effectively removes the dangers of the previous two coding strategies, but reduces the size of the dataset, itself something which could lead to changes in how the network is arranged. Despite this, it was thought that this was the safest route to take as it allows for a dataset in which possibly problematic assumptions about the data could be reduced, and so the majority of the simulations were performed under this condition. Most simulations were also performed under a Semantically Strict condition, in which the existence of cognate words in languages which had undergone large semantic shifts, resulting in quite different meanings to the Swadesh concepts, were left out. This was done to reduce the number of simulations required, and the potential complexitiy involved. However, one simulation, involving both old and modern languages, was Semantically Lax, i.e. these items were permitted. This was done to model the effects of

[4] These individuals and institutions are listed under the section titled "Acknowledgments".

including structural cognates with different meanings in the dataset. The simulations where divided into four subgroups.

The varieties of Germanic used in this dissertation are: Old Norse, Old English, Old Saxon, Old Frisian, Old West Low Franconian (Old Dutch), Gothic and Old High German. The modern descendants of these languages (with the exception of Gothic) that have been used in this study include Norwegian Bokmål, Norwegian Nynorsk (sometimes referred to as Neo-Norwegian in the literature), Swedish, Icelandic, Danish and Faroese; Modern English, Modern Low German, Modern Dutch, Flemish and Afrikaans; three varieties of High German (Modern Standard High German, Bavarian and Swiss German), and four varieties of Modern Frisian (West Frisian, Ferring Frisian, Frasch, and Saterland Frisian). Additionally, reconstructed Proto-Germanic forms are included. The selection of the varieties used in this dissertation was motivated by a desire to have a number of different varieties which would allow for testing of the network program under a number of conditions, rather than for establishing a comprehensive classification of the various Germanic languages and dialects. Certain varieties, such as Yiddish and Pennsylvania German, were not included as there were difficulties in obtaining reliable datasets for them timeously, and their inclusion was not felt necessary for testing the networking program. Intermediate stages of the selected varieties (such as Middle English) were not used due to time and length constraints. The varieties preceded by "Old" are the earliest attested forms of these languages, but are not all chronologically contemporaneous (for example, Old Frisian is attested later than Old English). The earliest and most recent attested forms of each variety were thus used in this dissertation.

Subgroup1: The Old Germanic Languages

The first subgroup examined the Old Germanic languages only. These were Gothic, Old English, Old Saxon, Old High German, Old Dutch (also known as Old West Low Franconian), Old Frisian and Old Norse. Two of the 100 concepts were troublesome as they were represented by a large number of words which had semantic differences which were very hard to judge in several of the languages, and would have led to either potentially very arbitrary decisions having to be made about which forms to choose, or a number of additional simulations having to be performed. These, SMALL and TO KILL, were therefore left out of several simulations. They were used, however, in one of the Old Germanic simulations. The first simulation was a semantically strict Majority Wins simulation. After the removal of the two troublesome characters, the dataset had ninety-eight characters. The coding of most of these is discussed in Appendix A; however, in instances where more than one word for the concept was present in a given language, the choice of a particular form will be discussed in this section. For character 17, TO DIE, both Old Saxon and Old English had two forms attested, one, *sweltan* a cognate of the Gothic term (*ga*)*swiltan* and the other, Old English *stēorfan* and Old Saxon *stervan*, cognate with forms found in the other West Germanic languages, such as Old High German *sterban* (where the bilabial plosive in place of a [+labial] fricative is due to the High German consonant Shift (Salmons, 2012; Fortson, 2004),

Old Dutch *stervan* and Old Frisian *sterva*. Of these, the latter form tends to occur much more widely, with the former being found mainly in poetry; because of this, it was decided to use the second form in the dataset. The Old English and Old Saxon forms were thus coded as cognates of the other West Germanic forms, while the Gothic form was assigned its own code, as was the Old Norse *deyja*. For character 33, TO GIVE, Old English had the forms *sellan*, ancestral to Modern English *sell*, and *gifan*, which is not a direct ancestor of Modern English *give*, but is a cognate of the Old Norse *gefa* from which this was borrowed. Both forms were widely used, but the second form has cognates in all of the other old Germanic varieties, as can be seen from the profusion of forms with the core structure $C_{[+vel/pal]}VC_{[+lab],}$[5] as in Gothic *giban*, Old High German *geban*, Old Saxon and Old Dutch *gevan*, Old Frisian *jeva* and Old Norse *gefa*. The form *sellan* is only found in Old English. Based on this, and the fact that *sellan* has undergone semantic narrowing to mean "to give something in exchange for money", while the *give* form still has a general meaning and has been continued in the majority of Germanic languages with such a general meaning, it was thought to be a better candidate for use in the dataset than *sellan*.

For concept 36 on the list, HAIR, all of the old Germanic languages with the exception of Gothic (which had the word *tagl*, cognate with Old English *tægl*, "tail") had a word with the structure /hV*r*/, as in Old English *hǣr,* [hæːr], Old Saxon, Old High German and Old Dutch *hār*, [haːr], Old Frisian *hēr*, [heːr], and Old Norse *hár*, [haːr]. These all descended from a common ancestor, given by Orel (2003) as Proto-Germanic ***hēran***. These were all coded as cognates. Old English and Old High German also had the words *fēax*, [feaks], and *fahs*, [faxs], respectively; both of these also had the meaning HAIR. These are also given by Orel (2003) as being descended from a Proto-Germanic form, ***faxsan***. Due to the fact that there do not appear to be any major semantic differences between the majority *h*-forms and the *f*-forms, as well as the fact that the *h*-forms are found in all the old Germanic languages and have been continued to this day in all of their ancestors with the meaning HAIR, it was decided to use the *h*-forms for all of the semantically strict simulations except for one additional one.

Two different forms presented themselves in the old Germanic data for concept 38, HEAD. The most widespread of these had the core structure /hV($C_{[+lab])}$(V)($C_{[+obst]}$)/, as in Gothic *haubiþ,* Old English *heafod*, Old Saxon *hōvid*, Old Dutch *hōvit*, Old High German *houbit*, Old Frisian *hāved*, and Old Norse *hǫfuð*. These are all descended from Proto-Germanic ***habudan~*houbidan*** (Orel, 2003), and were therefore coded as cognates. Old High German additionally had the word *kopf*; however, this is originally a borrowing from Latin and initially had the meaning "drinking vessel, container", with the meaning subesequnetly shifting to "skull" and then "head". This was therefore not included in this or any of its related simulations (Duden, 2015). Thus for all but one of the semantically strict old Germanic simulations Old High German *houbit* was used.

[5] Note: a number of the features used in this dissertation are not part of the standard set. The reason for this was that some non-standard features were deemed more useful to the determination of cognates than some of the standard features.

Two forms were present in the data for concept 46, LEAF. The most common of these had the structure lVC$_{[+lab]}$/, as in Gothic *laufs*, Old English *lēaf*, Old Saxon *lōf*, Old High German *loub*, and Old Frisian *lāf*, Old Norse *lauf*. These all descend from Proto-Germanic **louban* (Orel, 2003). These were all coded as cognates. In addition to the above words, Old High German and Old Saxon had *blat* and *blad*, respectively; as these words occured alongside others with the same meaning, and appear to have originally had a broader meaning referring to an object with an edge (compare English *blade*), it was decided to use the *l*-forms.

In the old Germanic languages two words which unambiguously meant MAN (concept 51) were attested. Of these the form which occurs more widely, in a variety of prose genres as well as some poetry, had the structure /C$_{[+cont, +lab]}$V$_{[+mid]}$r/, as in Gothic *wair*, Old English, Old Saxon, Old High German, Old Dutch, and Old Frisian *wer*, and Old Norse *verr*. These are all descended from Proto-Germanic **werwaz* (Orel, 2003) and as such were assigned the same codes. The second form had the structure /gVmV/, as in Old English *guma*, Old Saxon *gumo*, Gothic *guma*, Old High German *gomo* and Old Norse *gume*; these are however largely limited to poetry (OED, 2015), and as such the first form was favoured.

Concept 62, the negator NOT, was largely represented by words with the form /nV/, as in Gothic *ni*, Old English *nē*, Old Saxon *nē*/*ni*, Old High German *ni*/*nē*, and Old Frisian *nē*/*ni*. These are all cognates, being descended from Proto-Germanic **ne* (Orel, 2003), and were coded as such. Old Norse however has two words for NOT: *eigi* and *né*. The latter term is used mainly in poetry (OED, 2015) and, considering the former term is ancestral to all of the words meaning NOT in the modern North Germanic languages, may have been an archaism in the Old Norse period already. The first term was thus favoured, and was assigned a different code to the rest of the terms.

For concept 64, PERSON, three forms occurred in the data. In Gothic the term was *andwairþi*; this was assigned its own code. The two other forms are of particular interest because they presented a slight challenge in terms of how they were to be coded. One form had the structure /mV$_{[+low]}$C$_{[+alv/intdent]}$/ , as in Old English *man*, Old High German *man*, Old Norse *maðr* (this displays the fricativisation of an original [nː] conditioned by the masculine singular nominative marker *-r* (Faarlund, 1994); compare the plural *menn*). The second set of forms appear to be derivations based on this form (OED, 2015; SAOB, 2015), as in Old Dutch *mennisko*, Old Frisian *menneska*, and Old Saxon *menisko*. The question regarding these words was whether to assign them the same codes due to the fact that they are still ultimately cognates, or to assign them different codes so as to capture the fact that, despite being cognates, a derivational process has occurred in one set of words and not the other. It was decided to assign them the same codes for the former reason. They were assigned different codes in a later simulation.

Concept 71, TO SAY, had two widespread forms in the data. The one which occurred more had the root structure /C$_{[+alv,+cont,-son]}$VC/, as in Old English *secgan*, Old Saxon *seggian*, Old High German *sagen*, Old Dutch *sagon*, Old Frisian *sedza*, and Old Norse *segja*. These are all descended from a common ancestor, supplied by Orel (2003) as **sagjanan*, while the OED (2015) gives **sagǣjan* and **sagjan* as possible Proto-Germanic terms. These were all

assigned the same code. The second set of forms has the root structure /k^{w}VC$_{[+alv/dent, +obs]}$/, and are also descended from a Proto-Germanic ancestor, **kweþanan* (Orel, 2003). These forms are Gothic *qiþan*, Old English *cweþan*, Old Saxon and Old Dutch *kwethan*/*quethan*, Old High German *kwedan*/*quedan*, and Old Norse *kveða*. As the first set of terms has more members in the data, it was decided to initially use them for the semantically strict simulations; as an *s*-form is not attested in Gothic, it was kept as *qiþan* and assigned its own code.

Two words for the concept SEED (concept 73) occurred in Old Norse, *sáð*/*sćði* and *frjó*. The first of these has a structure which is shared by a number of the old West Germanic varieties, namely /C$_{[+alv,+fric]}$VC$_{[+alv,+obst]}$/, as seen in Old English *sæd*, Old Saxon *sâd*, and Old High German *sât*. According to the OED (2015), these are all descended from Proto-Germanic **sæđi-*/**sæđo-*, a noun derived from the verb root **sæ-*, "to sow". Orel (2003) however reconstructs the word as **sediz*. The only attested free-standing word with the meaning SEED attested in Gothic is *fraiw*, a cognate of the second Old Norse word. As the use of the *s*-form occurs in the West Germanic languages and Old Norse, but not Gothic, it was decided to use the Old Norse *s*-form for the initial set of simulations; this was therefore assigned the same code as the West Germanic forms, while the Gothic word was assigned its own code.

The concept SKIN $_{(human)}$ (no.75) was predominantly represented by words with the structure /hVC$_{[+alv/intdent]}$/, as can be seen in Old English *hyd*, Old Saxon and Old Frisian *hūd*, Old High German and Old Dutch *hūt*. These are all cognates (OED, 2015) and are supposed to have descended from a Proto-Germanic form such as **hūðiz* (OED, 2015) or **hudiz* (Orel, 2003); as such they were assigned the same codes in the Network program. These all had the specific meaning of "human skin". Old Norse also had the word *húð*, although it appears this originally carried the implication of animal skin, in much the same way as Modern English *hide*. For human skin, Old Norse had a word not attested in any of the other Old Germanic languages, *skinn*. This was therefore chosen as the Old Norse word to be used in the initial set of simulations, and was assigned a code separately to the West Germanic forms. Gothic had the word *fill* with the meaning SKIN; this is cognate with a number of other words in the other Germanic languages, such as Old English *fell*, Old Saxon, Old High German, Old Frisian *fel* and Old Norse *fjall*/*fiall*, although these generally had the connotation of animal skin, especially with the fur or wool still attached to it (OED, 2015). The Gothic form was thus assigned its own code.

The concept SMALL had a number of forms attested for it. One of the most common had the structure /smVl/, as in Old English *smæl*, Old Frisian *smel*, Old Saxon *smal*, Old Dutch *smal*, Old High German *smal*, and Gothic *smals*. These were coded as cognates. The old Norse form was *lítill*; the word *smalr* is attested, but only in a small number of fixed collocations and expressions (OED, 2015), suggesting it was no longer productive in Old Norse; based on this, *lítill* was selected instead. This was assigned its own code.

The most widespread form for the concept SMOKE (no.78) had the structure /*r*VC$_{[-voice,+obst, +vel/+pal]}$/, as in Old English *rēc*, Old Saxon and Old Dutch *rōk*, Old High German *rouh*, Old Frisian *rēk* and Old Norse *reykr*. These forms descend from a common

ancestral form, which is reconstructed as Proto-Germanic **roukiz* (Orel, 2003). These were therefore coded as cognates. Modern English *smoke* is descended from Old English *smoca*, although it is attested later then the *r*-form (OED, 2015) and was a backformation from Old English *smēocan*; it was therefore decided to use the *r*-form for the first set of simulations. Additionally, a word for SMOKE was not attested in Gothic; as the initial simulation used the Majority Wins strategy for missing data, this was assigned the same code as the *r*-forms.

There were two attested words for concept 90, TREE. The more widespread of the two had the structure /bV(C)m/, as in Gothic *bagms*, Old English *bēam*, Old Saxon and Old Dutch *bōm*, Old High German *boum*,Old Frisian *bām*, and Old Norse *baðmr*. These are all cognates, having descended from the same Proto-Germanic word which Orel (2003) has reconstructed as **boumaz* in Proto-Germanic; the OED (2015) reconstructs **baumoz* for Proto-West Germanic, and states that the Gothic and Old Norse forms, with their apparent velar stop and interdental fricative before the nasal, respectively, present phonetic complications which make the sounds of a Proto-Germanic form difficult to reconstruct. These were coded as cognates and chosen for use in the first set of simulations based on the fact that they were more widespread in the data. The second set of words, with the structure /t*r*V/, such as Old English *trēow*, Old Frisian *trê*, Old Saxon *trio*, Gothic *triu*, and Old Norse *tré*, all descended from Proto-Germanic **trewan* (Orel, 2003) or **trewo*-(OED, 2015), were used in an additional semantically strict simulation.

Concept 92, TO WALK, was slightly problematic in that some of the old Germanic languages appeared to not distinguish (at least to any significant degree) between walking (i.e.: moving on foot) and simply going (i.e.: no particular manner of locomotion was implied). This was handled by including the available terms with the closest meaning to TO WALK in a set of simulations, and replacing the concept TO WALK with the more general TO GO. For this simulation TO WALK was used. This resulted in two sets of words. One set had the root structure /$C_{[+vel,+obst]}VC_{[+nas]}$/, as in Gothic *gaggan*, Old English *gān*/*gangan*, Old Saxon *gangan*/*gān*, and Old Norse *ganga*. These are descended from Proto-Germanic **gēnan* (Orel, 2003). Three of the old languages did appear to distinguish lexically between TO WALK and TO GO; in these, the words for TO WALK had the root structure /l$VC_{[+lab]}$/, as in Old High German *loufan*, Old Dutch *lōpan*, Old Frisian *hlōpa*/*hlāpa*. These all descend from Proto-Germanic **hlaupen* (OED, 2015), possibly with the meaning "to jump, spring, run"; this meaning subsequently shifted to "walk". These were assigned their own code, which was separate from the code assigned to the *g*-forms.

For concept 99, WOMAN, the West Germanic languages had a form with the structure /$C_{[+lab,-stp]}VC_{[+lab]}$/, as in Old English, Old Dutch, and Old Saxon *wīf*, Old Frisian *wîf*, and Old High German *wīb*/*wīp*. In addition to this form a number of others occurred, such as Old English *ides* and *cwene*, and Old High German *itis*; these were not used as they had meanings which were more specialised than WOMAN (for example Old English *cwene* took on the connotation of an impudent or forthright woman (OED, 2015)), or uses which suggested they were literary or archaic (for example Old English *ides* is a literary form). Gothic and Old Norse had words which had preserved the older, neutral meaning of *cwene*; these were *qino*

and *kona*, respectively. These were assigned their own code separate from the *w*-forms which were the most neutral forms in the old West Germanic varieties.

For simulation two, the same words were used as simulation one, but where data was missing, the missing items were coded as deletions in a condition named Infodelete. This itself may have been a problematic strategy; this will be discussed under the heading "Discussion". Like simulation two, simulation three used the same words as simulation one; however, concepts which were missing data items were this time removed altogether. This was to try to avoid the problems caused by assuming that the most common form in the dataset was likely to be found in a language for which that item was missing, and to avoid the problems potentially caused by coding these missing items as deletions. This strategy resulted in the following concepts being removed from the analysis altogether: ASHES (No.2), BARK (No.3), TO BITE (NO.7), BONE (No.10), CLAW (No.13), EGG (No.24), FEATHER (No.27), TO FLY (No.30), GREEN (No.35), TO KILL (No.43), LEAF (No.46), LIVER (No.48), LOUSE (No.50), NOSE (No.61), ROOT (No. 68), ROUND (No.69), SEED (No.73), SMOKE (No.78), TO SWIM (No.83), TAIL (No.84), WARM (No.93), and YELLOW (No.100). Combined with the removal of TO KILL and SMALL (which were removed because of the large number of forms with uncertain semantic differences presented in a number of the languages) this reduced the size of the dataset to seventy-six items. This condition was known as the Infoclean condition.

Simulation four was a semantically strict Infoclean simulation which used words which were synonyms or very close near synonyms to the words used in simulation three. For the concept BELLY, in place of *wamba*, Old English *būc* and Old Dutch and Saxon *būk* were used. These were assigned their own code separate from the other words. For concept five, BIG, in place of forms such as *mihhil*, *mikil* and *micel*, Old High German *grōz*, Old English *grēat*, and Old Saxon *grāt* were used; these were assigned their own code and treated as cognates. For concept eight, Old English *blæc* was used instead of *swēart* and assigned its own code. For concept fourteen, CLOUD, the Old English form *scio* was used instead of *wolcen*; this was coded as a cognate of Old Norse *ský*. For concept seventeen, TO DIE, the poetic Old English and Old Saxon cognates of Gothic *(ga)swiltan*, both *sweltan*, were used instead of forms such as *steorfan* and were given the same code as the Gothic form. For concept nineteen, DRY, in place of *trokken*, the synonym *durri* was used for Old High German and assigned the same code as Gothic *þaursus* and Old Norse *þurr*. For concept twenty-eight, Old Norse *funi*, which was generally limited to poetic and archaic literary contexts, was used in place of *eldr*, the usual word for FIRE.

Old English *gifan* existed alongside the word *sellan*, which originally meant TO GIVE (OED 2015), and which underwent a semantic shift to give the modern day meaning of "to sell, to exchange something for money". *Sellan* was used instead of *gifan* in this simulation and assigned its own code. In both Old English and Old High German there existed alongside the respective words *hǣr* and *hār* the words *fēax* and *fahs* with the meaning HAIR; these were used instead of the *h*-forms in this simulation. They were assigned the same code, which was different from that of the *h*-forms. For the concept HEAD (number thirty-eight) Old High German *kopf* was used in the place of *houbit*; originally meaning "bowl, skull" it came to

mean “head” and over time ousted the descendant of *houbit*, *Haupt*, as the primary everyday term for HEAD (Duden, 2015). This was assigned its own code. The ancestor of Modern English *know*, Old English *cnāwan*, was used instead of *witan* for this simulation. There does not appear to have been a major semantic difference between this word and *witan*; *cnāwan* was therefore assigned its own code in the data and used under this semantically strict simulation.

For the concept MAN (number fifty-one) an older, poetic form was used; thus the *wer* forms were replaced by Old English and Gothic *guma*, Old Saxon *gumo*, Old High German *gomo*, and Old Norse *gume*; despite being limited to more literary contexts than the *wer* forms, these words all unambiguously meant “man”; they were thus assigned their own code and included in the simulation. For concept fifty-two, MANY, Old English *fela*, cognate with Old High German *filo*/*filu*, Old Dutch *filo* and Old Frisian *fele*/*felo*, was used in place of Old English *manig*; there does not appear to have been any particular semantic difference between the two; *fela* was assigned the same code as the other *f*-forms. Old English *swēora* was used in the place of *hals* for the concept NECK; this was assigned its own code. For concept sixty-two, NOT (verb negator) the Old Norse poetic word *né* was used instead of *eigi* and coded as a cognate of the forms from the other Germanic languages used in the previous simulations (Gothic *ni*, Old English *nē*, Old Saxon *nē*/*ni*, Old High German *ni*/*nē*, and Old Frisian *nē*/*ni*).

For concept sixty-four, PERSON, the *man* and *mennisko* forms were coded as non-cognates, based on that fact that in the second group a derivational strategy was used, whereas in the first group the base form of the words remained the same; the two groups were thus assigned separate codes. For the concept TO SAY the *s*-forms of the previous simulations were replaced by those with the root structure /k^{w}VC$_{[+alv/dent, +obs]}$/, as seen in Gothic *qiþan*, Old English *cweþan*, Old Saxon and Old Dutch *kwethan*/*quethan*, Old High German *kwedan*/*quedan*, and Old Norse *kveða*. For the concept SMALL (number seventy-seven) the *s*-form used for Gothic was replaced by a form cognate with the Old Norse form, namely *leitils*; this was coded as a cognate of this form. A number of forms which also have the root structure /lVC$_{[+stp,+alv]}$/ occur in the West Germanic languages, such as Old English *lytel*, Old Saxon and Old Dutch *luttil*, Old High German *luzzil*; however, these are thought to be parallel developments (OED, 2015), and so were assigned their own code. For the concept STAR, the Old Dutch word *sterro* was replaced by the word *sterno*; these forms appear to be descended from two different forms which already existed in Proto-Germanic, and were thus assigned different codes (OED, 2015; see Appendix A for more details); Old Dutch *sterno* was thus coded as a cognate of Old High German *sterno*, Gothic *stairno* and Old Norse *stjarna*, instead of being coded as a cognate of Old English *stēorra*, Old Saxon *sterra* and Old Frisian *stera*. For concept eighty-six, THIS, Gothic *þatuh* was coded as not being cognate with the other forms, unlike the previous simulations, due to the fact that it consists of *þat* plus the strengthening particle *–uh*, rather than *þat* plus **se~*si* (possibly meaning “see, behold” (OED, 2015)) as seen in Old Norse and the West Germanic languages. For concept ninety, TREE, the words with the structure /t*r*V/, such as Old English *trēow*, Old Frisian *trê*, Old Saxon *trio*, Gothic *triu*, and Old Norse *tré*, were used in place of the *b*-forms used in the previous simulations. Instead of TO WALK, (concept ninety-two), the more general concept

of TO GO was used in this simulation; this resulted in the *l*-forms of Old Dutch, Old High German and Old Frisian being replaced by Old Frisian and Old Dutch *gān*, Old High German *gēn*; these were coded as cognates of the words in Old English, Gothic, Old Saxon, and Old Norse used in the previous simulations.

Subgroup 2: The Modern Germanic Languages.

Three simulations were performed for the modern Germanic languages. The first simulation was a Majority Wins semantically strict simulation with eighteen separate sequences (languages) and one hundred characters (concepts) used. The coding of these characters was generally easier than with those of the old Germanic languages, as in a number of instances the meanings of the words were easier to ascertain. In addition to this, the lexica of the modern Germanic languages do not have the gaps in attestation which pose a serious problem with the old Germanic languages. Both of these points will be discussed in the section titled "Discussion". As in the above simulations, words were coded as cognates based on their core structure; this method was checked, where possible, by examining the established etymologies published in the above-mentioned dictionaries. In instances where more than one word was found in a given language with the same meaning, or where the meanings were difficult to tell apart (i.e. the terms were basically synonymous), one term was included in the first semantically strict simulation, and the second in the other. In instances where there was only one term, the assigning of codes was easier; for more details on these terms please consult Appendix A. The instances where more then one term existed for a given concept, and the selection of these items, will be discussed below.

For concept number three, BARK, Ferring Frisian had two words, *rinj* and *buark*. It was decided to initially use the first of these, as without frequency data it was not possible to tell which was the more common form. *Rinj* was coded as a cognate of Standard High German *Rind*, Bavarian *Rindn* and Swiss German *Rinde*, based on the shared structure $C_{[+postalv/alv]\ [+trill]}V\ C_{[+alv][+nas]}$; the presence of the palatalised nasal [n^j] in Ferring Frisian was in all probability a later sound change. With the available data it was not possible to tell for certain whether or not this was a borrowing, as the word is not reconstructed in Proto-Germanic (Orel, 2003, gives **skurtaz* and **barkuz*); however, the presence of the word of the word in a number of the old west Germanic languages suggests it is an innovation unique to the West Germanic subgroup. Afrikaans also had two words for the concept BARK, *bas* and *skors*. For this simulation *bas* was chosen and coded as a cognate of the Flemish word *bast* based on their shared core structure $C_{[+vo,+lab,+stp]}\ V_{[+low]}\ C_{[-vo,+alv,+fric]}$. The concept BELLY was represented by two words in Frasch Frisian. In Frasch Frisian the words are *bük*/*buuk* and *lif*. The first of these was coded as a cognate of Ferring Frisian *bük*, Saterland Frisian *Buuk*, Low German *Buuk*, Standard Dutch and Flemish *buik*, Afrikaans *buik*, Standard High German and Bavarian *Bauch*, Swiss German *Buuch*, Danish *bug*, Faroese *búkur*, and Swedish *buk*. The second form was used in an additional semantically strict simulation.

For concept thirty-four, GOOD (adjective) Swedish has two words, *gott/god* and *bra*. The first of these words is the native Germanic word (SAOB, 2015), and is descended ultimately from Proto-Germanic **godaz* (Orel, 2003). This form was coded as a cognate of the words in

the other Germanic languages, which all have the same structure, $C_{[+vel,+obst]}VC_{[+alv]}$; this can be seen in Modern English *good*, Ferring Frisian *gud*, West Frisian *goed*, Saterland Frisian *goud*, Frasch Frisian *gödj*, Low German *good*, Standard Dutch, Flemish and Afrikaans *goed*, Standard High German *gut*, Bavarian *guad*, Swiss German *guät*, Danish *god*, Faroese *góður*, Icelandic *góður*, and Norwegian *god*. The second word, *bra*, is ultimately a borrowing from French (SAOB, 2015); this was included in a second semantically strict simulation.

Concept thirty-six was represented overwhelmingly in the data by words with the structure /hV(*r*)/, as in Modern English *hair*, Ferring Frisian *hiar*, West Frisian *hier*, Saterland Frisian *Híer*/*Häier*, Frasch Frisian *häär*, Low German *Hohr*, Standard Dutch *haar*, Flemish *haer*, Afrikaans *haar*, Standard High German and Bavarian *Haar*, Swiss German *Hòòr*, Standard Danish *hår*, Faroese *hár*, Icelandic *hár*, and Norwegian *hår*. Swedish had the word *hårstrå*, a compound which has come to be used to distinguish a single hair from the collective hairs which *hår* can refer to (SAOB, 2015); it was decided to code this as a cognate of the above words based on the presence in the compound of *hår* for this simulation. An additional simulation designed to capture the fact that the whole word is not a cognate of the other forms, and that a derivational strategy has been used in Swedish and not the other Germanic languages was performed. This is discussed further on.

In the case of concept thirty-eight, HEAD, two separate forms occurred in both Standard High German and Afrikaans. The first of these had the structure /$hVC_{[+lab]}C_{[+obst]}$/, as in Standard High German *Haupt* and Afrikaans *hoof*; the second had the structure /kVp(f)/, as in Standard High German *Kopf* and Afrikaans *kop*. The first of these is the older form, with cognates in most of the other Germanic languages (see Appendix A); it is descended ultimately from Proto-Germanic **habudan~*houbidan* (Orel, 2003). The second is a later borrowing from Latin (Duden, 2015; Krause, 2001), and is additionally found in Swiss German *Kopf*, Saterland Frisian *Kop*, Low German *Kopp*, Afrikaans *kop*, and Bavarian *Koopf*; originally this word had the meaning "drinking vessel, bowl, container"; this was borrowed into Old High German and initially appears to have taken on the meaning "skull", before shifting to mean more generally "head" (this happened in the early Old High German period) (Duden, 2015). While the first word has retained its meaning of HEAD, it has undergone a semantic shift which has caused this to become an almost secondary meaning, with the most commonly used meaning being that of "main thing, boss, main form of something" (Duden, 2015), as in Standard High German *Hauptfach*, "major subject", or *Hauptamt*, "main office", and Afrikaans *skoolhoof*, "school principle", and *hoofkantoor*, "main office"; the more common everyday word with the meaning "head" in both languages has the form /kVp(f)/; this was thus chosen for this simulation.

For concept 43, TO KILL, Bavarian had the words *hiimocha* and *umbringa*; it was decided to use the first of these for this simulation and the second in an addtional semantically strict simulation. A number of languages expressed TO KILL by combining the adjective *dead* with the verb (*to*) *make*, as in Ferring Frisian *duadmaage*, Saterland Frisian *doodmoakje*, Frasch Frisian *düüdj mååge*, Low German *dood moken*, and Afrikaans *dood maak*; it was decided to code the Frisian forms as cognates, and the Low German and Afrikaans forms as separate non-cognates, as it is probable that these arose in parallel. For the concept LEAF

(number forty-six) modern Icelandic presented an interesting problem.In the data two forms were attested, one with the structure /lV(C)/, and the other with the structure /blVC$_{[+alv]}$/ (consult Appendix A for further details); while the other languages made use of one or the other of these, Icelandic used a compound made up of both, *laufblað*. For this simulation it was decided to code *laufblað* as a cognate of the /lV(C)/ forms, and in a later simulation as a cognate of the /blVC$_{[+alv]}$/ forms.

Icelandic had two words for the concept MAN, namely *maður* and *karl*; for this simulation it was decided to use *maður*; this is a cognate of all the words with the structure /mV(C$_{[+alv]}$)/, such as Modern English *man*, Ferring Frisian *maan*, West Frisian *man*, Saterland Frisian *Mon*, Frasch Frisian *moon*, Low German *Mann*, Standard Dutch, Flemish and Afrikaans *man*, Modern Standard High German *Mann*, Bavarian *Moo*, Swiss German *Maa*, Danish *mand*, Faroese *maður*, Norwegian *mann*, and Swedish *man*. These were all assigned the same code for this simulation. For the concept MANY (number fifty-two) Ferring Frisian had two words: *manigen* and *fölen*. It was decided to use the first word for this simulation; this was coded as a cognate of Modern English *many*, Frasch Frisian *maning*, Danish *mange*, Faroese *mangir*, Icelandic *margir*, Norwegian *mange* and Swedish *många*. For Afrikaans it was decided to use the native Germanic *veel* rather than *baie* (from Malay *banjak* (Donaldsson, 1994)), as *veel* only has the meaning "many", while *baie* is more commonly used with the meaning "very". For concept fifty-eight, NECK, both Standard Dutch and Afrikaans have two words, namely *hals* and *nek*; as there is no major semantic difference between them, it was decided to use *hals* for this simulation and *nek* in another.

Bavarian and Swiss German both have two words for the concept TO SIT, Bavarian *sitzn* and Swiss German *sitze*, and Bavarian *hockä* and Swiss German *hocke*. The first of these pairs has cognates in other Germanic languages and varieties, which all have the core structure /C$_{[+alv,\ +fric]}$VC$_{[+alv]}$/, such as Modern English *sit*, Ferring Frisian *sat*, West and Saterland Frisian *sitte*, Frasch Frisian *sate*, Low German *siddn*, Standard Dutch and Flemish *zitten*, Afrikaans *sit*, Standard High German *sitzen*, Danish *sidde*, Faroese *sita*, Icelandic *sitja*, Norwegian Nynorsk *sittja*, Norwegian Bokmål *sitte*, and Swedish *sitta*. These were coded as such and used in this simulation.For concept seventy-five, West Frisian, Frasch Frisian, Afrikaans and Swedish all had two separate words for SKIN (human): West Frisian *hûd* and *fel*, Frasch Frisian *hüd* and *schan*, Afrikaans *huid* and *vel*, and Swedish *hud* and *skinn*. It was decided to use the *h*-forms for this simulation; these have cognates in a large number of the other Germanic varieties, which all have the same core structure, /hVC$_{[+alv/intdent]}$/, as in Ferring Frisian *hedj*, Saterland Frisian *Häid*, Low German *Huut*, Standard Dutch and Flemish *huid*, Standard High German and Bavarian *Haut*, Swiss German *Hut*, Danish *hud*, Faroese *húð* (the interdental fricative is no longer pronounced (Barnes & Weyhe, 1994)), Icelandic *húð*, Norwegian Bokmål *hud*. These were all assigned the same code for this simulation.

The concept SMALL (adjective; number seventy-seven) had a number of words in many of the languages which could have possibly been chosen, and which had either very small, or very difficult to judge, semantic differences. It was thus decided to choose forms which were commonly used in the languages, and which were widespread in the data. The problems

which could possibly arise by doing this will be discussed in the "Discussion" section. Thus, two groups of words were used, one with the structure /lV($C_{[+alv,+obst]}$)($C_{[+alv]}$)/, and one with the structure $C_{[+obst]}$lv(n)/. The first can be seen in Modern English *little*, , Ferring Frisian *letj*, West Frisian *lyts*, Saterland Frisian *litje*/*litjet*, Frasch Frisian *latj*, Low German *lütt*, Danish *lille*/*liden*, Faroese *lítil*, Icelandic *lítill*, Norwegian *lille*/*liten*, and Swedish *lilla*/*liten*; as the root vowel is [+high], and in all instances the root of the word begins with the liquid [l] and ends with an alveolar consonant, it was decided to code them as cognates in this simulation. The second structure was found in Standard Dutch, Flemish and Afrikaans *klein*, Standard High German *klein*, Bavarian *gloa*, and Swiss German *chlai*; these were coded as cognates of each other.

The concept STAR (concept eighty) initially appeared to have forms which were cognate, with the structure /$C_{[-vo,+pal-alv,+fric]}$tV(*r*)(n)/, as in Modern English *star*, Ferring Frisian *stäär*, West Frisian *stjer*, Saterland Frisian *Stiern*, Frasch Frisian *stäär*, Low German *Schtirn*, Standard Dutch *ster*, Flemish *ster*, Afrikaans *ster*, Standard High German *Stern*, Bavarian *Schtean*, Swiss German *Schtern*, Danish *stjerne*, Faroese *stjørna*, Icelandic *stjarna*, Norwegian *stjerne* and Swedish *stjärna*. Orel (2003) reconstructs a single form in Proto-Germanic for these, **sternon*. The OED (2015) however suggests that the nasal-less forms are a common West Germanic continuation of Proto-Germanic strong masculine **ster-*, while the forms with a nasal continue the weak feminine **sternon*, developed in parallel in Germanic; based on this it was decided to treat the forms with a nasal and those without as non-cognates, assuming they were different words in Proto-Germanic. Concept eighty-five, THAT, presented a similar problem in a number of languages to Swedish *hårstrå*, that being that a compound of some sort was used which contained one element which was a cognate of the others, but which because of the compounding had followed a different morphological strategy. This could be seen in Ferring Frisian *detdiar* and Frasch Frisian *dåt(deer)*, where the first element has the structure /$C_{[+intdent/alv,+obst]}$V$C_{[+intdent/alv,+obst]}$/ and is cognate with the forms found in most of the other languages, such as Modern English *that*, West Frisian *dat*, Saterland Frisian *dät*, Low German, Standard Dutch and Flemish *dat*, Standard High German *das*, Bavarian *des*, Danish *det*, Icelandic *þetta*, and Norwegian and Swedish *det* (*där*) (older *detta*). It was decided to assigne the Ferring and Frasch Frisian words the same codes as these owing to the presence of the /$C_{[+intdent/alv,+obst]}$V$C_{[+intdent/alv,+obst]}$/ root structure, despite the fact that an additional element occurs in these words. An additional simulation was performed in which they were coded differently. Afrikaans *daardie* was assigned its own code due to the fact that the constituents of this compound are not etymologically descended at all from Proto-Germanic **þat*, unlike the above forms. Swiss German had the word *säb*, which is descended from the pronoun *selb*, with *l*-vocalisation (Rowley, personal correspondence); this was assigned its own code.

Nynorsk had two words for the concept WE (first person plural pronoun; number ninety-five), *vi* and *me*; for this simulation, *vi* was used; it was assigned the same code as all of the other words for this concept in the Germanic languages, with the exception of Afrikaans *ons*, based on the presence of the structure/$C_{[+lab]}$$V_{[-low]}$/. Afrikaans was assigned its own code. For concept ninety-eight, WHO, Low German had two forms, *weer* and *wokeen*. The first is

ultimately descended from Proto-Germanic **hwaz~*hwez* (Orel, 2003) or **χwaz~*χwez* (OED, 2015), and is a cognate of Modern English *who*, West Frisian *wa*, Low German *weer*, Standard Dutch, Flemish and Afrikaans *wie*, Standard High German *wer*, Bavarian *wea*, Swiss German *wèèr*, Danish *hvem*, Faroese *hvør*, Icelandic *hver*, Norwegian Nynorsk *kvem*, Norwegian Bokmål *hvem* and Swedish *vem*. It was therefore assigned the same code as these and used in this simulation. The second word was used in a later simulation. Ferring Frisian had the word *hoker*, which is a borrowing from Old Low Franconian *hok*, itself derived from *hwelik*, "which" (Hoekstra, personal correspondence); this was assigned its own code, as was Saterland Frisian *wäl*, derived from Low German *welk*, "which", (Hoekstra, personal correspondence). Bavarian was recorded as having two terms for WOMAN (concept ninety-nine), *Weiberz* and *Frau*; the first is common to almost all of the West Germanic languages (although not always with the same meaning or connotations), and is descended from Proto-Germanic **wiƀan* (Orel, 2003), whereas the second is attested later and descends from a word used in the continental West Germanic languages originally with the meaning "a noble woman" (GTB, 2015; Duden, 2015); this later had its meaning generalised to WOMAN. *Weiberz* was used for this simulation and was assigned the same code as Ferring Frisian *wüv*, Frasch Frisian *wüset* (from Old Frisian **wüfshood*, with *hood* cognate with English *head* and having the same distributive function as in "head of cattle' (Hoekstra, personal correspondence)), Saterland Frisian *Wieuw*(*moanske*) and Modern English *woman* (from Old English *wīfmann*). Two other groups of words existed for this concept, one with the structure /f*r*V/ and the other with the structure /k($C_{[+lab]}$)Vn/. The first can be seen in West Frisian *frou*, Low German *Fru*, Standard Dutch and Flemish *vrouw*, Afrikaans *vrou*, Standard High German *Frau*, and Swiss German *Frau*; these were coded as cognates. The second can be seen in Danish *kvinde*, Faroese *kvinna*, Icelandic *kona*, Norwegian Nynorsk *kvinnfolk*, Norwegian Bokmål *kvinne* and Swedish *kvinna*; these were assigned the same code (for details on the differences between them, please consult Appendix A).

The second simulation using data from the modern Germanic languages was also a semantically strict simulation, with the other word options for the above mentioned concepts being used in place of those in the previous simulation. Thus, for Ferring Frisian the word *buark*, which is a borrowing from Old Norse and is quite commonly used by speakers, was used in place of *rinj* and assigned the same code as Modern English *bark*, Swedish *bark*, Danish *bark*, Norwegian[6] *bark*, Faroese *børkur*, and Icelandic *börkur*. In the case of Afrikaans, the word *skors* was used in place of *bas* and was assigned the same code as Standard Dutch *schors*. For BELLY, the Frasch Frisian word *bük*/*buuk* was replaced by *lif*; this is a cognate of West Frisian *liif* and was assigned the same code. For GOOD, the Swedish word *bra* was used instead of *gott*; *bra* is ultimately from French *brave*, via Middle Low German, and has become extremely common and is often used as something of a synonym of *gott* (SAOB, 2015); it was assigned its own code.

For character thirty-eight, HEAD, the High German varieties (Standard High German, Swiss German, and Bavarian) had the form *Haupt* substituted for the *Kopf* forms; this was done because the semantic shift that is evident between Old High German *houbit* and the reflex

[6] If a form occurs in both Nynorsk and Bokmål, then simply Norwegian is used (see Appendix A).

found in the modern High German varieties is not a complete semantic shift, meaning that the older meaning ("the body part known as the head") is still associated with this word; it has simply taken on a wider semantic range, as explained above. This meant that *Haupt* could still be included in a semantically strict situation, and was treated as a less common word for HEAD. This was also the case for Afrikaans. Both the High German varieties and Afrikaans were thus assigned the same code as the other *h*-forms for the concept HEAD. In place of *hiimocha*, the Bavarian word *umbringa* was used for the concept TO KILL. Although it is likely that the constructions found in Ferring Frisian *duadmaage*, Saterland Frisian *doodmoakje*, Frasch Frisian *düüdj mååge*, Low German *dood moken*, and Afrikaans *dood maak* arose in parallel, due to the fact that they are made up of cognate elements, it was decided for this simulation that they would all be assigned the same code, unlike the previous simulation in which the Frisian forms were assigned their own code, and then Low German and Afrikaans each had a code to themselves. This was done in order to capture the fact that, even if these constructions arose in parallel, they are nevertheless made up of the same elements.

The Icelandic compound *laufblað*, under the concept LEAF, which was given the same code as the *leaf*-forms in the previous simulation, was coded as a cognate of the *Blatt*-forms in this one. For MAN, Icelandic *karl* was used in place of *maður*, which was use in the previous simulation; *karl* was assigned its own code. For the concept MANY, Ferring Frisian had the word *fölen* used instead of *manigen*; this was coded as a cognate of Saterland Frisian *fúul*, Low German *vel*, Dutch and Flemish *veel*, Afrikaans *veel*, Modern Standard High German *viele*, Swiss German *fili* and Bavarian *fui*, based on the presence of the structure /fV(l)/ (see Appendix A for an explanation of the Bavarian form). For this character Afrikaans *baie* was used based on the fact that, even though it generally means "very", in some circumstances it can also be used with the meaning "many" (for example, *Daar is baie mense in die kerk*, "There are a lot of/many people in the church", and *Baie mense dink*, "Many people think"), although this is to a more limited extent than *veel*. This was assigned its own code. For character fifty-eight, Afrikaans, Standard Dutch and Flemish had the word *nek* substituted for *hals*; *nek* was assigned the same code as Modern English *neck*.

For the concept TO SIT Bavarian *hockä* was used instead of *sitzn*, and Swiss German *hocke* was used instead of *sitze*; these words are cognates, and were assigned their own code. For SKIN (human), the West Frisian word *hûd* was replaced by *fel*, in the same way that Afrikaans *huid* was, for this simulation, substituted by *vel*. These two were assigned the same code. Frasch Frisian *schan*, which is an early loan from Old Norse which is quite widely used in the modern language, was used in place of *hüd*, as was Swedish *skinn* and Danish *skind* in place of *hud*; these were assigned the same code as Modern English *skin* and Norwegian Nynorsk *skinn*. For SMALL, Modern English *little* was replaced by *small*, which semantically is very similar; the North Germanic plural *små* was not used as its etymological connections to the English form are disputed, and the prevailing view is that they are unrelated (OED, 2015). The continental West Germanic cognates of the Modern English form, such as Standard High German *schmal* and Dutch *small* mean 'narrow, thin" rather

than “small”, and so were not included in this simulation. Modern English *small* was therefore assigned its own code.

Despite the fact that the two main groups of words for STAR (those with a nasal and those without) were explained as having developed in parallel (OED, 2015) and it is therefore likely they already existed as separate words in Proto-Germanic, they still ultimately go back to the same Proto-Indo-European root; this in conjunction with the fact that they have the same meaning was the motivation for assigning them the same code and treating them all as cognates for this simulation. In a similar vein to the *dead + make* constructions above, the Ferring and Frasch Frisian words for THAT, *detdiar* and *dåt*(*deer*), were not coded as cognates of the rest of the dataset but were assigned their own code. Afrikaans *daardie* was not assigned the same code as these due to the fact that the portion of the compound meaning “there” is in a different position (first rather than last) and it is almost certain, based on the fact that the second element comes from the definite article, rather than a demonstrative, that these arose completely separately. This retained its own code. For concept ninety-five, WE, Nynorsk *me* was used rather than *vi*; this was given its own code. Instead of Low German *weer*, for WE, the word *wokeen*, etymologically from *welk een*, “which one” was used; this was given its own code. Bavarian *Frau* was used in this simulation in place of *Weiberz* and was coded as a cognate of West Frisian *frou*, Low German *Fru*, Standard Dutch and Flemish *vrouw*, Afrikaans *vrou*, Standard High German *Frau*, and Swiss German *Frau*.

The last simulation for the modern Germanic languages dataset was performed under the semantically strict semi-cognate condition. For this simulation words which were either compounds containing elements non-cognate to the words in the above simulations, or had morphological elements in them which were not present in the other words used, were assigned different codes to the ones they had originally been assigned in order to model this. For this, Swedish *hårstrå* was assigned a different code to the other words for HAIR due to the fact that it is not the root on its own, but a compound containing an additional word. The Ferring and Frasch Frisian words for THAT, *detdiar* and *dåt*(*deer*), were not coded as cognates of the rest of the dataset under this condition, owing to the fact that they are compounds, but were assigned their own code. Afrikaans *daardie* was not assigned the same code as these due to the fact that the portion of the compound meaning “there” is in a different position (first rather than last) and it is almost certain, based on the fact that the second element comes from the definite article, rather than a demonstrative, that these arose completely separately. This retained its own code. Icelandic *við* and Faroese *vit* were coded as cognates of each other, but not of the other words for WE, as in the previous simulations, due to the fact that the interdental fricative [ð] in the case of the former and the alveolar stop [t] in the case of the latter are not found in any of the other forms; these are the remains of what had originally been a weak form of Old Norse *tveirr*, appended as a clitic to the word *vér*, which formed the Old Norse dual pronoun *vit* (Thráinsson, 1994; this has a cognate in Old English *wit*, from *wē + twā*).

Subgroup 3: Old and Modern Languages-Combined Simulations

The first of the simulations using data from both the old and modern Germanic languages combined both simulations one and seven under the Majority Wins semantically strict condition. One concept (TO KILL) was removed entirely from the simulation as it was very difficult to determine whether or not its primary meaning was "to kill" or "to torture, torment" in the old Germanic languages; this could be achieved by a frequency analysis comparing how both meanings are used in old Germanic texts, but this is outside the range of this study. This resulted in a simulation with ninety-nine concepts and twenty-five sequences. This simulation effectively added the first and seventh simulations to each other as the codes for the individual words remained the same as in these previous simulations, i.e.: the Majority Wins technique within the old Germanic languages was treated as separate to that used in the modern Germanic languages. For example, where the concept BONE was not represented by an attested word in Gothic, the code assigned to it was the same as that assigned to it in simulation one (which involved only the old Germanic languages). This was done as a way of preventing a word which either came about much later, or only took on a particular meaning later, than the old stage of a given language. The codes for each word in the modern and old Germanic languages were thus the same as those in simulations one and seven. For simulation eleven the Majority Wins semantically strict condition was also used, although this time words with largely the same meaning but which were different to the ones used in the previous simulation were included in place of the previous sets of words in the languages in which they occurred. This effectively combined simualtions five and eight. The Majority Wins condition was applied across both the old and modern Germanic languages in the same way as the previous simulation. Simulation twelve was performed under the Infoclean semantically strict condition. Due to the fact that certain lexical items were not attested in a number of the old Germanic languages, the following items had to be removed, as the program is unable to have certain characters (concepts in this case) present for some sequences (in this case languages) and not for others within the same simulation: ASHES (concept two), BARK (concept three), TO BITE (concept seven), BONE (concept ten), CLAW (concept thirteen), EGG (concept twenty-four), FEATHER (concept twenty-seven), TO FLY (concept thirty), GREEN (concept thirty-five), TO KILL (forty-three), LEAF (forty-six), LIVER (forty-eight), LOUSE (fifty), NOSE (sixty-one), ROOT (sixty-eight), ROUND (sixty-nine), SEED (seventy-three), SKIN (seventy-five), SMOKE (seventy-eight), TO SWIM (eighty-three), TAIL (eighty-four), WARM (ninety-three) and YELLOW (one hundred). This resulted in a simulation with seventy-eight concepts and twenty-five sequences. The remaining items were coded as in simulations three and ten. Simulation thirteen was performed under the Infoclean semantically strict condition, with the same concepts listed above removed, leading to seventy-eight concepts and twenty-five sequences. The rest of the words were coded as in simulation eleven. Simulation fourteen was performed under the Infoclean semantically lax condition. For this simulation words which have undergone semantic shifts between the old and modern periods in the Germanic languages, and which are thus present in the modern languages with different meanings from the old languages, were used. The same concepts as the previous two simulations were removed, leading to seventy-eight concepts and twenty-five languages being used. The following

substitutions were made: For BELLY Modern English *womb*, which is descended from Old English *wamb*, "belly" (OED, 2015), was used and assigned the same code as Gothic *wamba*, Old English *wamb*, Old Saxon, Old High German and Old Dutch *wamba*, and Old Frisian *wamme*. Modern English *fowel*, which now has a more limited meaning than *bird*, being usually used to refer to wild birds (especially those which are hunted) or to domestic birds kept for consumption, was used for the concept BIRD; as this is a descendant of Old English *fugol*, it was given the same code as all of the words which are descended from Proto-Germanic **fuglaz* and which have the structure $C_{[-vo,+lab\text{-}dent,+fric]}V_{[+back]}C_{[+vel]}(V)C_{[+lat\text{-}app]}$, such as Gothic *fugls*, Old English *fugol*, Old Saxon *fugal*, Old High German *fogal*, Old Dutch *fogal*/*vogal*, Old Frisian *fugel*, Old Norse *fugl*, Ferring Frisian *fögel*, West Frisian *fûgel*, Saterland Frisian *Fúgel*, Frasch Frisian *föögel*, Low German *Vågel*, Dutch and Flemish *vogel*, Afrikaans *voël*, Standard High German *Vogel*, Bavarian *Foogl*, Swiss German *Foogel*, Danish *fugl*, Faroese *fuglur*, Icelandic *fugl*, Norwegian *fugl* and Swedish *fågel*. Modern English *starve*, which comes from Old English *stēorfan*, "to die, perish", now means specifically to die due to lack of food, or to experience the effects of a lack of food. This was assigned the same code as the Old English word.

Modern English *dog* replaced by *hound*, which is generally now used to refer to a hunting dog; this is a direct descendant of Old English *hund*, which seems to have swapped meanings with Old English *docga*, the ancestor of *dog*, which originally had the narrower meaning, referring to a breed of dog particularly associated with hunting (OED, 2015). *Hound* was assigned the same code as Old English *hund*. While the Old Norse word for FAT was *smjǫr*, the meanings of most of the descendants of this word have all shifted slightly to mean "butter", as in Danish *smør*, Norwegian *smør*, Icelandic *smjer* and Swedish *smör*; these were used under the semantically lax condition and assigned the same code as the Old Norse word. For concept thirty-three, TO GIVE, the Old English *sellan* was used; this had the general meaning of "to give" but underwent a semantic shift to give Modern English *to sell*, "to give something in exchange for money"; this was used in place of Modern English *to give* and assigned the same code as Old English *sellan*; these thus stood alone in the data. In addition to *witan* and *cnāwan*, Old English also had the word *cunnan/cannan* for TO KNOW; this gave rise to the Modern English modal verb *can*, with the meaning having shifted from "to know" to "to know how to do something", probably via sentences such as *Cannst þū lǣsan?*, "Do you know how to read?", to 'to be able to do something" (OED, 2015), as in *Can you read* versus *Do you know how to read?*, which have a clear semantic difference in Modern English. These were both used in this simulation, and were assigned the same code. Old Norse *karl* underwent semantic shifts in both Danish and Swedish; in the case of Danish *karl* has gone on to refer to a farmhand (ODS, 2015), while in Swedish it has survived with the meaning "man", but with prominent connotations of manliness (SAOB, 2015), which are absent from the word *man*. These were thus included in the semantically lax simulation, and were assigned the same code as the Old Norse word. The Modern English descendant of Old English *hȳd* is *hide*; this underwent a semantic narrowing so that, instead of referring generally to skin, it now refers specifically to the skin of an animal, especially a mammal (OED, 2015); this was used in place of Modern English *skin* for this simulation and assigned the same code as the other *h*-forms. Likewise, Old English *rēc* has survived in Modern

English as *reek*; this underwent a semantic shift from the meaning "smoke" to 'strong, foul odour". *Reek* was used in place of *smoke* in Modern English for this simulation, and was assigned the same code as Old English *rēc*. The Old English word *bēam* has survived as *beam* in Modern English, with the semantic shift being from "tree" to "a pole-like structural support, a shaft-shaped object, a shaft of light" (OED, 2015). *Beam* was included in this simulation and assigned the same code as Old English *bēam*. Old Dutch *wīf* and Old High German *wīb*/*wīp* have both survived in the modern day languages as Dutch *wijf* and Standard High German *Weib*, but with distinctly pejorative connotations which generally imply a hussy, slattern or loose and disreputable woman (Duden, 2015; GTB, 2015); it was for this reason that they were included under only the semantically lax condition. Both were assigned the same codes as their Old Dutch and Old High German ancestors.

Subgroup 4: The Old and Modern Germanic Languages with Proto-Germanic

For this last quartet of simulations, reconstructed Proto-Germanic forms were included in the datasets. These forms were effectively added to the pre-existing datasets of simulations ten and twelve. In a number of cases there were two reconstructed Proto-Germanic forms for a given concept; as the meanings of these reconstructed forms is usually not one hundred per cent certain, and many are listed as having a number of possible related meanings, there was generally no particular reason for choosing to use a particular form in a given simulation; thus the choice of which of a pair of reconstructed forms with apparently identical or very similar meanings to use in a given simulation was somewhat arbitrary. Two conditions were used for these simulations: a Majority Wins semantically strict condition, and an Infoclean semantically strict condition. The reason for avoiding a semantically lax condition with these was that it was felt that this might carry the risk of having to include a very large number of different terms in the daughter languages which are descended from, or related to, the reconstructed forms, but which have undergone a range of semantic changes, resulting in a dataset which would require a number of different simulations in this subgroup alone, and which would extend the size of this study far beyond the practicality of this exploratory analysis. The Majority Wins pair of simulations thus had ninety-nine concepts, and twenty-six languages sequences. The Infoclean pair of simulations had seventy-eight concepts and twenty-six languages sequences. The potential pitfalls of included unattested, reconstructed items in such an analysis are examined in the section titled "Discussion". Below only those instances where two Proto-Germanic forms were reconstructed for the same concept are discussed; for more information on the other forms, please consult Appendix A.

Under the Majority Wins condition, for the concept BARK (concept three) Orel (2003) reconstructs two Proto-Germanic words, **skurtaz* and **barkuz*; the first of these was initially used, and assigned the same code as Old High German *skorza* in the simulation marked Word Choice One. The second was selected for use in a simulation marked Word Choice Two, and was assigned the same code as Old Norse *börkr*. For the concept CLAW (thirteen) the forms in the descendant languages of Proto-Germanic were descended from two different forms: the words in the West Germanic languages are believed to be descended from either the Proto-Germanic root **klawâ-~*klæwâ-* (OED, 2015) or the verb **klawjanan* (Orel, 2003), while the North Germanic words are thought to be descended from Proto-Germanic **klôh-*, related to

the verb stem **klâ-* (OED, 2015). For Word Choice One, the first root was included in the Proto-Germanic sequence, and was assigned the same code as the West Germanic forms; for Word Choice Two the second root was used and was assigned the same code as the North Germanic forms. Two reconstructed words for FAT exist in Proto-Germanic, **smerwa* (OED, 2015) or **smerwon* (Orel, 2003), and **faito-* "fat" (adj) or **faitido-*, the past participle of the Proto-Germanic verb **faitjan*, "to fatten' (OED, 2015). The first of these was used under Word Choice One and was assigned the same code as the forms found in all of the old Germanic languages; the second was used under Word Choice Two and was assigned the same code as most of the modern Germanic languages (with the exception of Frasch Frisian). Two separate words for FIRE are reconstructable for Proto-Germanic, **fuwer~*fur* (Orel, 2003), **fûir* (OED, 2015) and **fon* (Orel, 2003). The first of these is the proposed ancestor for the words in the data with the structure /fV*r*/ (largely represented by West Germanic languages), while the second is the putative ancestor of Gothic *fon* and the poetic Old Norse *funi*. The first of these reconstructions was used in the simulation marked Word Choice One, and the second in the simulation marked Word Choice Two. For HAIR, Orel (2003) reconstructs both **hēran* and **faxsan*; the first of these was used for the simulation marked Word Choice One and given the same code as the *h*-forms in the various attested Germanic languages. The second was used in the simulation marked Word Choice Two. The concept MAN also has two reconstructable Proto-Germanic words, **wiraz* and **gumon* (Orel, 2003); the first of these was used in the first Majority Wins simulation, and was assigned the same code as the various *wer* forms in the old Germanic languages, while the second was used in the second Majority Wins simulation and was assigned its own code. The concept MANY also had two reconstructed Proto-Germanic words, **managaz* and **felu*; the first of these was used in the first Majority Wins simulation, and the latter in the second; both were coded as cognates of the respective words in the various daughter languages. Two reconstructions for FLESH are possible: Orel (2003) gives **flaiskaz* for the first form, while the OED (2015) supplies **flaiskoz*, while for the second form Orel (2003) reconstructs **huldan*. The first of these was used in the Word Choice One simulation and the second in the Word Choice Two simulation. Both were assigned the same codes as their respective cognate forms in the daughter languages (the same code as Old English and Old High German in the first case, and the same code as Old Norse in the second). There are two reconstructed words for TO SAY in Proto-Germanic: Orel (2003) gives **sagjanan* while the OED (2015) gives **sagǣjan* and **sagjan* as the ancestor of those forms in the modern Germanic languages which have the core structure /$C_{[+alv,+fric]}$V(C)/, while those which have the structure /$k^{w}VC_{[+alv/intdent,\ +obst]}$/ have the reconstructed ancestor **kweþanan* (Orel, 2003). The first of these was used for the simulation marked Word Choice One, the second with the simulation marked Word Choice Two; each was given the same code as the words with the same respective structures. For the concept SEED three possible Proto-Germanic words have been reconstructed. One of these is the form thought to be ancestral to all the words with the structure /$C_{[+alv,+fric]}VC_{[+alv,+obst]}$/ in the data; this is reconstructed as **sediz* (Orel, 2003); the OED (2015) has **sǣđi-*/**sǣđo-*, a noun derived from the verb root **sǣ-*, "to sow". This was the Proto-Germanic form used in the simulation labled Word Choice One. The second reconstructed form is ancestral to the Gothic and North Germanic words with the core structure /fr(j)V/ and is given by Orel (2003) as **fraiwan*. This was used in the simulation marked Word Choice Two. The third form is

**semon*, which is ultimately ancestral to the words in the data with the structure /$C_{[+alv, +fric]}$VmV/, such as Standard High German *Samen*, Bavarian *Saama* and Swiss German *Saame*; this was included in a separate simulation. The concept SKIN (human) likewise had more than one possible reconstructed word in Proto-Germanic; one of these is **hūðiz* (OED, 2015) or **hudiz* (Orel, 2003), which is the reconstructed ancestor of the words with the structure /hV$C_{[+alv/intdent]}$/, such as Old English *hȳd*, Old Saxon and Old Frisian *hūd*, Old High German and Old Dutch *hūt*, and Old Norse *húð*, as well as the forms descended from these. This form was used in the simulation marked Word Choice One. The second reconstructed form supplied by Orel (2003) is **skenþō*; this was used under Word Choice Two and assigned the same code as Old Norse *skinn*. For STAR, the OED (2015) suggests that the forms in the Germanic languages which occur with a post-vocalic nasal are parallel formations to those without; Orel (2003) only reconstructs one form, **sternon*, although the OED (2015) has two, **sterron* and the parallel formation **sternōn*; it was decided to use the forms found in the OED and use the first form without a nasal under Word Choice One. This was assigned the same code as the nasaless forms in the data. The second form was used under Word Choice Two and was assigned the same code as those forms in the data which possess a post vocalic nasal in the first syllable. Proto-Germanic appears to have had two words for TAIL: **stertoz* (OED, 2015) or **stertaz* (Orel, 2003), and **taglan* (Orel, 2003), although the original sense of the second word appears to have been "the hairy tail of an animal"; due to this, only the first form was used; it was assigned the same code as those words with the core structure /stV(r)t/. The demonstrative pronouns under THIS created an interesting problem when Proto-Germanic data was to be included. In the North and West Germanic languages THIS was conveyed by suffixing **se~*si* (possibly meaning "see, behold" (OED, 2015)) to the original Proto-Germanic **þat*, while Gothic attached the strengthening particle *–uh* to **þat* (OED, 2015); due to this, there is no certain reconstruction of the Proto-Germanic word for THIS; it was thus decided to remove the missing item and remove the concept THIS from the simulation. For concept ninety, TREE, two Proto-Germanic words have been reconstructed, **boumaz* (Orel, 2003) and **trewan* (Orel, 2003) or **trewo-* (OED, 2015); the first of these is the ancestral form to all the words in the data with the core structure /bVm/ and was coded as a cognate of these and was used in the simulation marked Word Choice One; the second was coded as a cognate of the forms with the core structure /trV/ and was used in the simulation marked Word Choice Two. For WOMAN two words which can be reconstructed are **wiƀan* and **kwenan*; the first of these was used in the simulation marked Word Choice One and was assigned the same code as the words in the rest of the dataset with the structure /$C_{[+lab,-stp]}$V$C_{[+lab]}$/, such as Old English, Old Dutch, and Old Saxon *wīf*, Old Frisian *wîf*, Old High German *wīb*/*wīp*, Ferring Frisian *wüv*, Frasch Frisian *wüset*, Saterland Frisian *Wieuw*(*moanske*), Bavarian *Weiberz*, and Modern English *woman* (from Old English *wīfmann*). The second was used in the simulation marked Word Choice Two and was assigned the same code as those words with the structure /k($C_{[+lab]}$)Vn/, such as Gothic *qino*, Old Norse *kona*, Danish *kvinde*, Faroese *kvinna*, Icelandic *kona*, Norwegian Nynorsk *kvinnfolk*, Norwegian Bokmål *kvinne* and Swedish *kvinna* (consult Appendix A for more detailed information). Orel (2003) also reconstructs the word **diso*; this was left out as the semantics of it, and its descendants, were unclear.

For the two Infoclean semantically strict simulations the strategy for coding items was the same as in the two simulations described above, with the same word choices being made, the exception being those items which were not attested in the data and which were thus left out. The concepts which were removed as a result were: ASHES (concept two), BARK (concept three), TO BITE (concept seven), BONE (concept ten), CLAW (concept thirteen), EGG (concept twenty-four), FEATHER (concept twenty-seven), TO FLY (concept thirty), GREEN (concept thirty-five), TO KILL (forty-three), LEAF (forty-six), LIVER (forty-eight), LOUSE (fifty), NOSE (sixty-one), ROOT (sixty-eight), ROUND (sixty-nine), SEED (seventy-three), SKIN (seventy-five), SMOKE (seventy-eight), TO SWIM (eighty-three), TAIL (eighty-four), WARM (ninety-three) and YELLOW (one hundred).

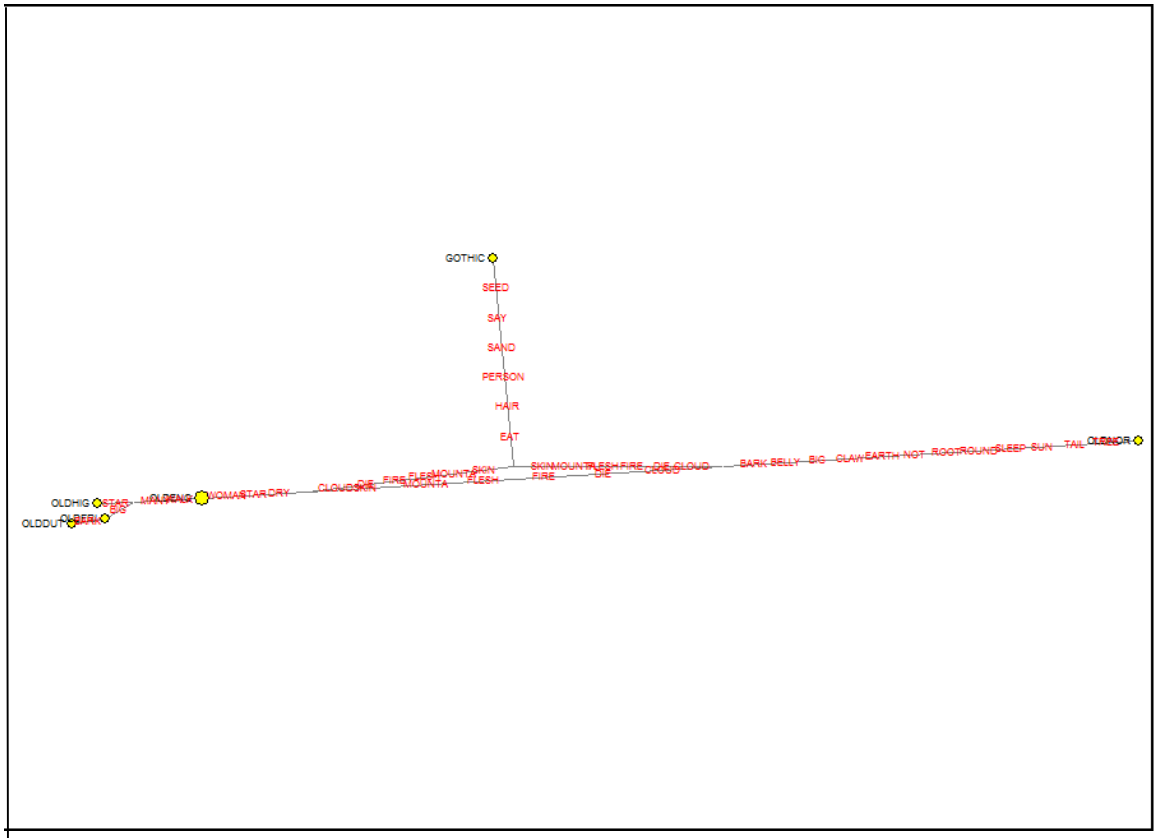
Chapter 4

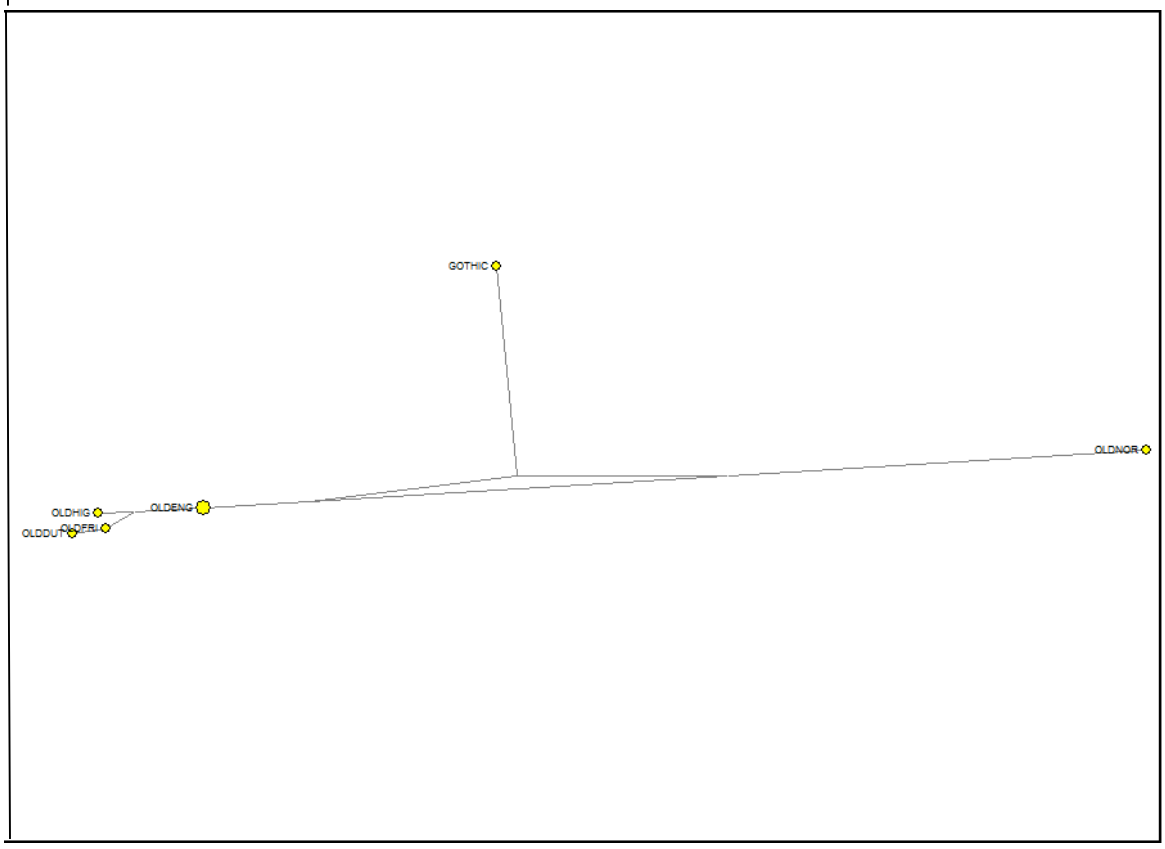
Results[7]

Subgroup 1: The Old Germanic Languages

The first simulation within this group was the Majority Wins semantically strict simulation with seven languages and ninety-eight concepts. This was run at ε=0, following the coding scheme given in the method section. This resulted in the network in the diagram on the following page.

[7] In some instances sections of network diagrams were not clear. Please consult Appendix B for closeups of parts of the following networks: Subgroup 2 – Simulations 1 and 2, Subgroup 3 – Simulation 2, Subgroup 4 – Simulations 1 and 3.

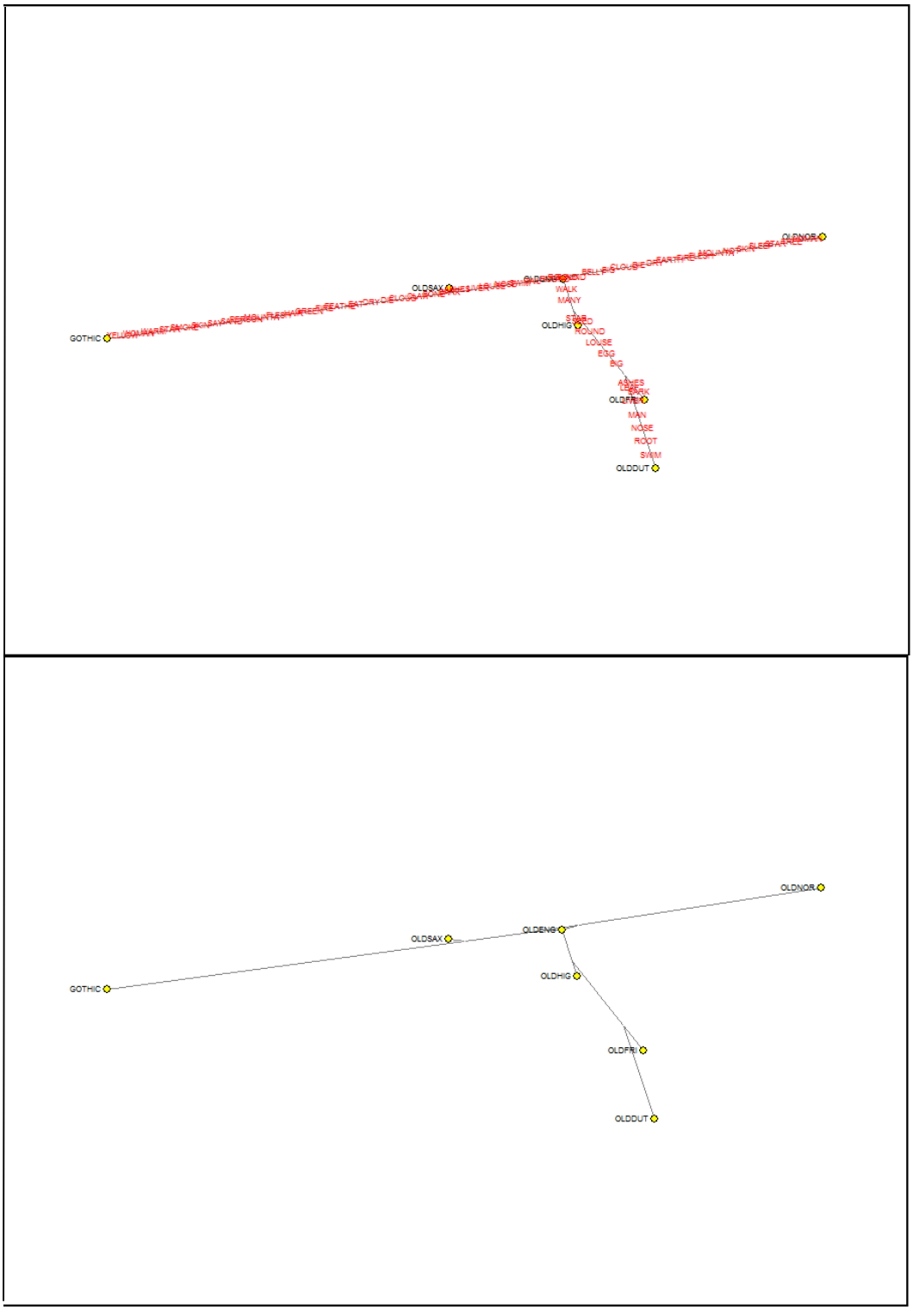




Old Germanic Languages, Majority Wins, Semantically Strict Simulation, ε=0, Concept n=98, Language n=7

This network has three main branches: one for Gothic, one for Old Norse, and one on which the old West Germanic varieties are clustered. Concepts listed on each branch are those by which nodes further down a given branch differed from those further up the branch. The node marked OLDENG is larger than the others as it is made up of both the Old English and Old Saxon sequences; this is because under this condition all of the Old English and Old Saxon words were cognates, causing them to be treated as the same sequence which has occurred twice.The concepts listed on a branch of the network between two or more nodes indicate which concepts were different between those nodes. The Gothic sequence differed from both the old West Germanic varieties and Old Norse by the words for the concepts SAY, SAND, SEED, PERSON, HAIR, EAT, SKIN, FLESH, MOUNTAIN, FIRE, DIE, and CLOUD. Old Norse differed from both Gothic and the West Germanic languages by the words for the concepts TREE, SUN, SLEEP, NOT, EARTH, BIG, CLOUD, DIE, FIRE, FLESH, MOUNTAIN, BARK, BELLY, CLAW, ROOT, ROUND, TAIL and SKIN. The Old West Germanic varieties collectively differed from Old Norse and Gothic by the words for the concepts WOMAN, STAR, DRY, CLOUD, SKIN, MOUNTAIN, FLESH, FIRE and DIE. Within the West Germanic group there was a division between Old English and Old Saxon and Old High German, Old Dutch and Old Frisian. This initial subdivision was caused by differences in the codes for WALK and MANY, with the Old English and Old Saxon words being different from those in the other old West Germanic varieties; Old Dutch, Old Frisian and Old High German had words with the core structure/lV$C_{[+lab]}$/ (Old High German *loufan*, Old Dutch *lōpan*, Old Frisian *hlōpa*/*hlāpa*) for the concept WALK, while for MANY they had words with the structure /fVl/, such as Old High German *filo*/*filu*, Old Dutch *filo* and Old Frisian *fele*/*felo*. Old English and Old Saxon on the other hand had words with the core structure /$C_{[+vel,+obst]}$V($C_{[+nas]}$)/ for WALK (for example, Old English and Old Saxon *gān*/*gangan*), and with the core structure /m$V_{[+back]}$(n)-/ (for example, Old English *mānig* and Old Saxon *manag*). Old High German differed from Old Dutch and Old Frisian by the word for the concept STAR. Old Frisian and Old Dutch differed fom the rest of the West Germanic varieties by the word for the concept BIG; in both cases in this simulation they had words with the structure /gr$V_{[+back]}$t/, such as Old Dutch *grōt* and Old Frisian *grāt*. Old English and Old Saxon differed from each other by the word for the concept NECK. This simulation presents evidence for the subgrouping of the West Germanic varieties together based on the levels of shared vocabulary items from the Swadesh 100-word list, but indicates that the old West Germanic varieties are slightly more closely related to Gothic than they are to Old Norse in terms of basic vocabulary. However, this relationship is still fairly weak, and Gothic is shown as differing significantly from both of them. This particular network reflects the traditional division in the Germanic languages as an essentially tripartite one, but is different in the sense that it does not group Old Norse and the old West Germanic varieties as forming their own subgroup distinct from Gothic (East Germanic). This grouping of the old West Germanic varieties and Old Norse together as a Northwest Germanic subgroup is based primarily on sound correspondences; this difference in grouping suggests that, despite similarities in their sound systems, the ancestors of Old Norse and the old West Germanic languages underwent changes in the use and selection of words which constitute those on the Swadesh 100-word list.

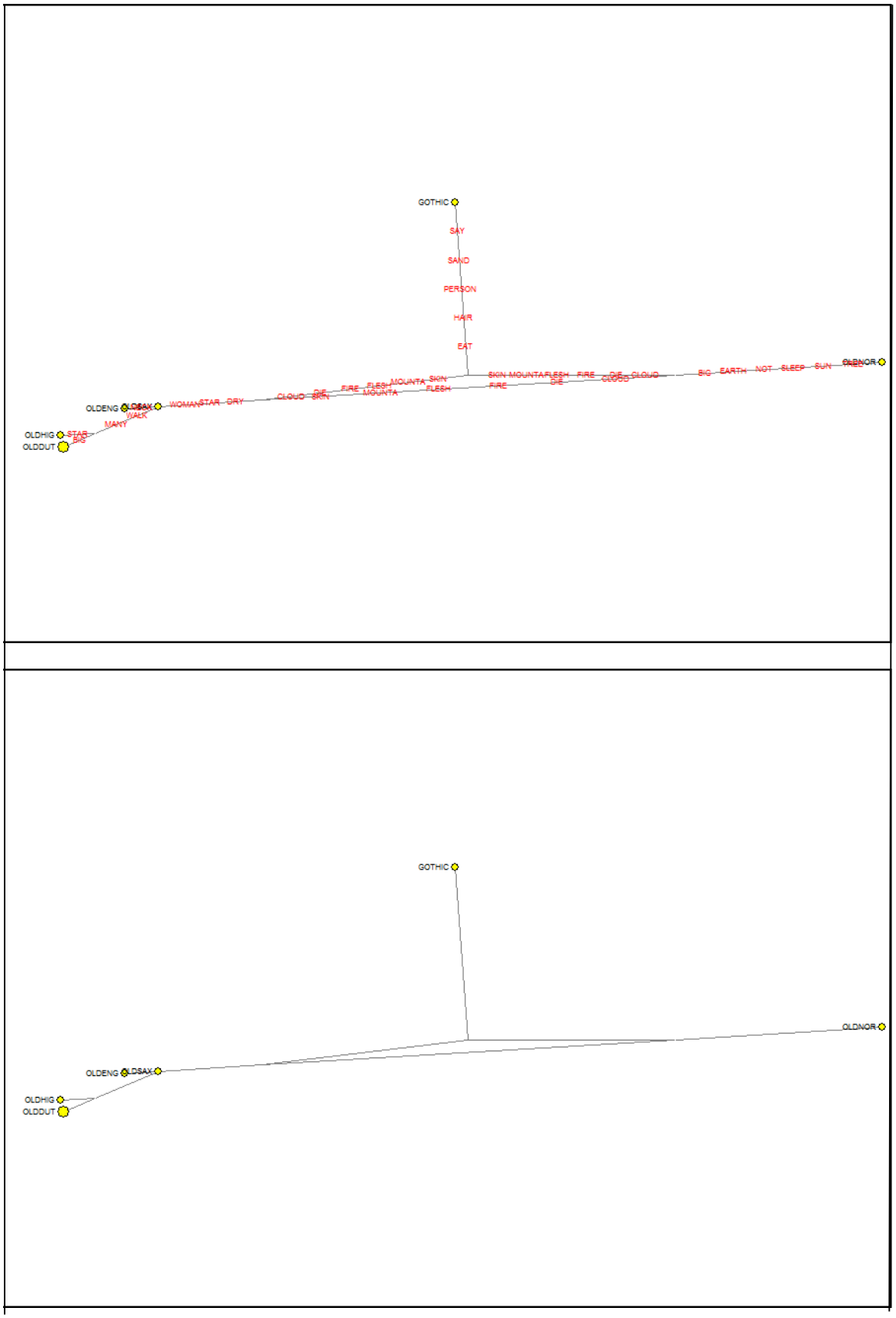
The second simulation using data from the old Germanic varieties was carried out under the semantically strict Infodelete condition. This resulted in the following network.



Old Germanic Languages, Infodelete, Semantically Strict Simulation, ε=0, Concept n=98, Language n=7

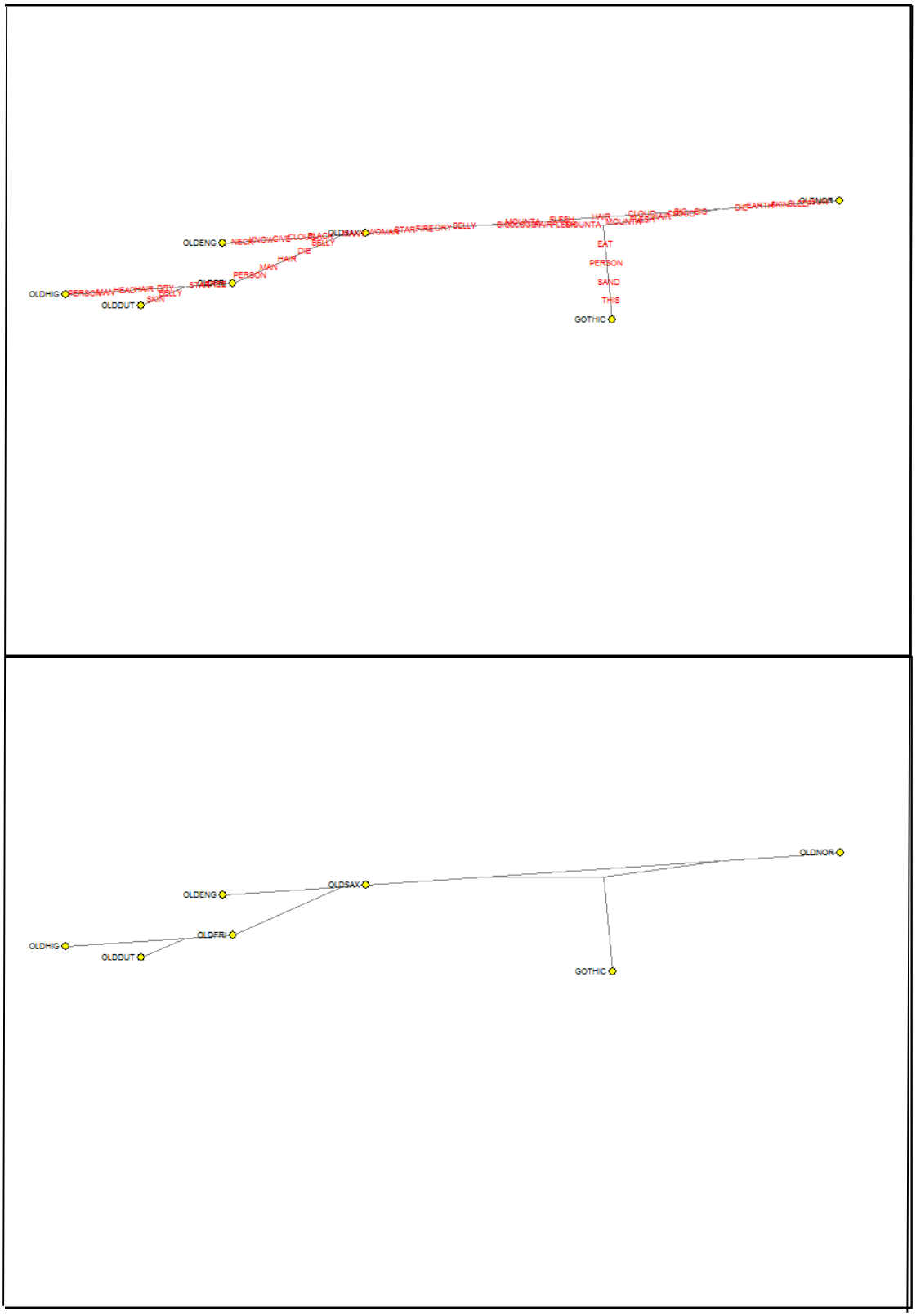
Old Dutch and Old Frisian both differed from Old High German by the words for the concepts BIG, EGG, LOUSE, ROUND, and SEED. These were the differences from Old High German which they shared with each other (i.e. for each of these concepts they both had the same code, which was different to the Old High German code). Of these, EGG, LOUSE, ROUND and SEED were unattested in Old Dutch and Old Frisian, and were thus coded as deletions. The word for the concept BIG differed in this simulation between Old High German (which had *mihhil* in this simulation) and Old Dutch and Old Frisian (which had *grāt*). Old Frisian branched off the common branch separating it and Old Dutch from Old High German due to the fact that words for the concepts ASHES and BARK were not attested and so were coded as deletions. Old Dutch branches off from the branch it shares with Old Frisian because of differences in the codes of LEAF, LIVER, MAN, NOSE, ROOT and SWIM. The words for these concepts were coded as deletions due to the lack of attestation of them. The branch of the network along which the Old High German, Old Dutch and Old Frisian nodes sit diverged from Old English, which was basal to it, due to differences in the coding of MANY and WALK. In both cases Old High German, Old Dutch and Old Frisian shared words for these concepts which were not used in any of the other languages, or were not attested with the same meaning. For MANY they had words with the structure /fVl/, such as Old High German *filo*/*filu*, Old Dutch *filo* and Old Frisian *fele*/*felo*; in the other languages a word with the root structure /mV$_{[+back]}$(n)-/ (for example, Old English *mānig*) was used. For WALK, all three languages made use of a word with the core structure /lVC$_{[+lab]}$/ and which originally meant "to leap", as in Old High German *loufan*, Old Dutch *lōpan* and Old Frisian *hlāpa*/*hlōpa*. SWIM, TAIL, NOSE, LIVER and ROUND were unattested in both Gothic and Old Saxon; the coding of these as deletions in the data resulted in both languages branching away from the section of the network in which the rest of the West Germanic languages branch. Old Saxon additionally had no attested word for ASHES, which caused it to branch off from the branch separating them from the other Germanic languages. Gothic was shown as being significantly divergent; of the ninety-eight concepts used in this simulation, it differed from Old Saxon, which in this case is represented as its closest relative, by twenty-two concepts. Of these, BARK, BONE, CLAW, CLOUD, and FEATHER; GREEN, SMOKE, WARM and YELLOW were unattested and were coded as deletions. Old Norse was shown as differing from Old English, which was positioned closest to where it diverges from the rest of the Germanic languages, by sixteen concepts. All of these were attested words, and there were no concepts which were coded as deletions in this branch. As is clear from the above, coding of missing data as deletions within a dataset can result in significant divergence between sequences in the network owing to the fact that the algorithm treats deletions as differences in and of themselves. This indicates that this particular strategy is not reliable and should best be avoided, especially in instances where a given language may have large numbers of unattested words.

The next simulation was performed under the Infoclean condition and resulted in the network on the following page.



Infoclean, Semantically Strict Simulation, ε=0, Concept n=76, Language n= 7

This network has three main branches: one for Gothic, one for Old Norse, and one on which the old West Germanic varieties are clustered. Concepts listed on each branch are those by which nodes further down a given branch differed from those further up the branch. The point marked “Old Dut” is larger than the other points because it represents both Old Dutch and Old Frisian; this is due to the fact that under this simulation condition all of the words used in the Old Dutch and Old Frisian sequences were cognates.The concepts listed on a branch of the network between two or more nodes indicate which concepts were different between those nodes. The Gothic sequence differed from both the old West Germanic varieties and Old Norse by the words for the concepts SAY, SAND, PERSON, HAIR, EAT, SKIN, FLESH, MOUNTAIN, FIRE, DIE, and CLOUD. Old Norse differed from both Gothic and the West Germanic languages by the words for the concepts TREE, SUN, SLEEP, NOT, EARTH, BIG, CLOUD, DIE, FIRE, FLESH, MOUNTAIN and SKIN. The Old West Germanic varieties collectively differed from Old Norse and Gothic by the words for the concepts WOMAN, STAR, DRY, CLOUD, SKIN, MOUNTAIN, FLESH, FIRE and DIE. Within the West Germanic group there was a division between Old English and Old Saxon (which differed from each other by one concept) and Old High German, Old Dutch and Old Frisian. This initial subdivision was caused by differences in the codes for WALK and MANY, with Old English and Old Saxon sharing cognates for WALK and MANY; Old Dutch, Old Frisian and Old High German shared their own cognates for these concepts. Old High German differed from the rest of the old West Germanic varieties by the word for the concept STAR. Old Frisian and Old Dutch differed fom the rest of the West Germanic varieties by the word for the concept BIG. Old English and Old Saxon differed from each other by the word for the concept NECK. This simulation presents evidence for the subgrouping of the West Germanic varieties together based on the levels of shared vocabulary items from the Swadesh 100-word list, but indicates that the old West Germanic varieties are slightly more closely related to Gothic than they are to Old Norse in terms of basic vocabulary. However, this relationship is still fairly weak, and Gothic is shown as differing significantly from both of them. This particular network reflects the traditional division in the Germanic languages as an essentially tripartite one, but is different in the sense that it does not group Old Norse and the old West Germanic varieties as forming their own subgroup distinct from Gothic (East Germanic). This grouping of the old West Germanic varieties and Old Norse together as a Northwest Germanic subgroup is based primarily on sound correspondences; this difference in grouping suggests that, despite similarities in their sound systems, the ancestors of Old Norse and the old West Germanic languages underwent changes in the use and selection of words which constitute those on the Swadesh 100-word list. The second simulation using data from the old Germanic varieties was carried out under the semantically strict Infoclean condition. This resulted in the following network (see next page).



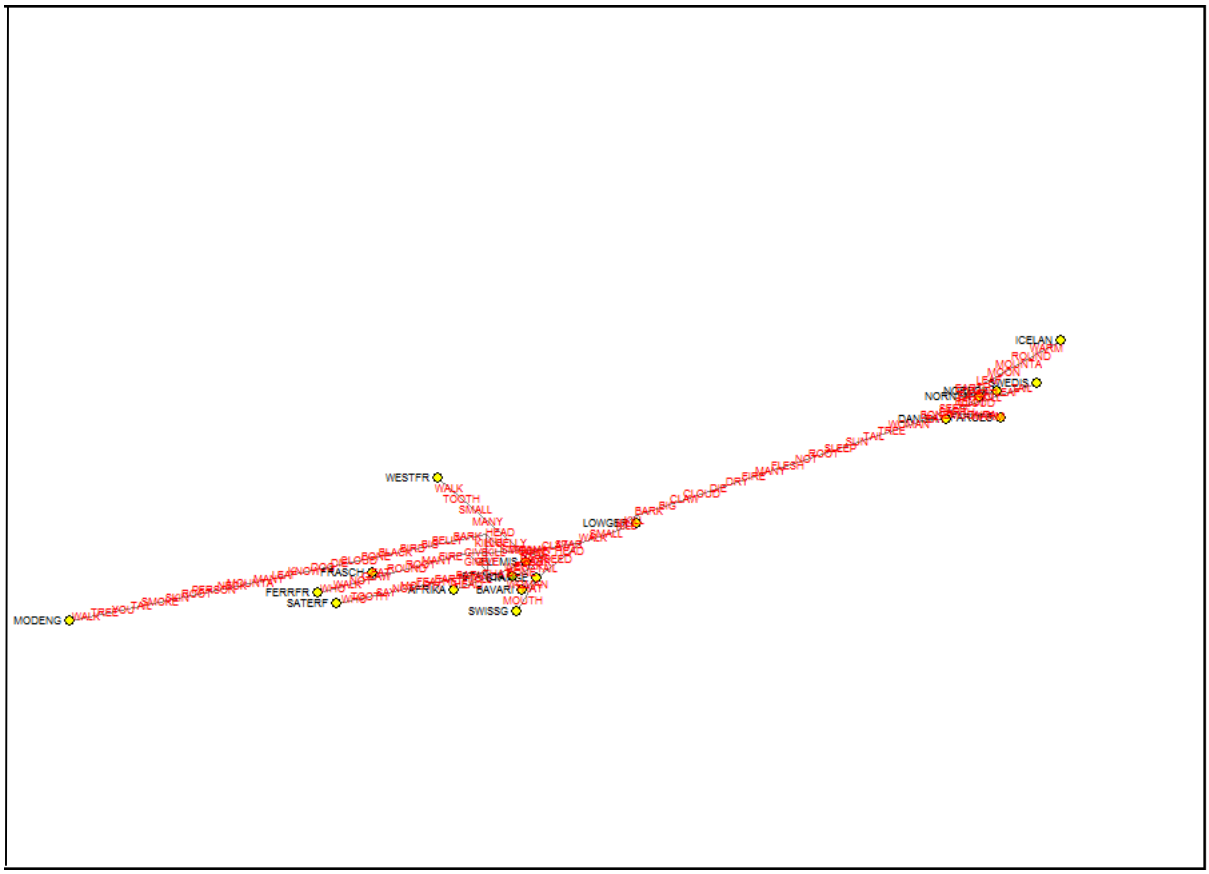
Infoclean, Semantically Strict Simulation: Synonyms and Near-Synonyms, ε=0, Concept n=99, Language n=7

The Infoclean simulation which included both synonyms and near-synonyms produced a network with three distinct groupings of the languages. Like the other simulations, Gothic and Old Norse occupied their own very divergent branches, while the West Germanic varieties formed their own section of the network. The West Germanic languages differed from both Gothic and Old Norse by the words for the concepts WOMAN, STAR, FIRE, DRY, MOUNTAIN, FLESH, BELLY and CLOUD. In this simulation Old Saxon differed from the rest of the old West Germanic varieties by the word for the concept MANY; this was due to Old English *mānig* being replaced by its synonym *fela*, which is a cognate of Old High German *filo*/*filu*, Old Dutch *filo* and Old Frisian *fele*/*felo*. Old English branched off from the rest of the old West Germanic varieties based on the words for the concepts BLACK, CLOUD, GIVE, KNOW and NECK. In the case of BLACK, KNOW, GIVE and NECK, this was due to the use of words which are not attested in the other Germanic languages with these meanings, such as *blæc*, *cnāwan*, *sellan* and *swēora*. In the case of CLOUD, Old English *wolcen* was replaced by *scio*, which is a cognate of Old Norse *ský*. Old Frisian, which had a higher overall sequence similarity to Old English and Old Saxon than either Old Dutch or Old High German did under this condition, branched off based on the words for the concepts BELLY, DIE, HAIR, MAN, and PERSON. Where Old English and Old Saxon had *būc* and *būk*, respectively, for BELLY, Old Frisian only had *wamme*. Old Frisian also did not have an attested word for HAIR which was a cognate of Old English *fēax* and Old Saxon *fax*. Of the concepts shown on the Old Frisian branch of the network, DIE was the only one which was shared across Old Frisian, Old Dutch and Old High German. This was due to the fact that the Old English and Old Saxon words for DIE from the other simulations were replaced by the poetic form *sweltan*, which has no attested cognates in any of the other old West Germanic varieties. In Old English and Old Saxon, the poetic words for MAN, *guma* and *gumo*, respectively, were used; this word is not attested in Old Frisian, which had the word *wer* retained for this simulation. The Old Frisian word for PERSON was a derived form based on the same root as the Old English and Old Saxon forms. Both Old High German and Old Dutch differed from Old Frisian, Old English and Old Saxon by their words for TREE and STAR. Old Dutch branched off on its own because of the codes for its words for BELLY and SKIN; Old Dutch had the word *būk* for BELLY in this simulation, rather than *wamba*, which both Old High German and Old Frisian had; for SKIN it had *hūt* rather than a word with the form /fVl/ (such as Old English *fell)*. Old High German branched off based on DRY, HAIR, HEAD, MAN and PERSON; of these HAIR (*fahs*), MAN (*gomo)* and PERSON (*man*) were cognates of the Old English and Old Saxon words, while DRY (*durri*) was a cognate of Gothic *þaursus* and Old Norse *þurr*. Old Norse, Gothic and the old West Germanic languages all differed from each other by the terms for BIG, CLOUD, FLESH, HAIR and MOUNTAIN. Gothic branched off based on the terms for EAT, PERSON, SAND and THIS; the first three are not shared by any of the other branches (although *andwairþi*, PERSON, seems to be a construction ultimately based on the same root as *wer* (OED, 2015)), while the last is a construction of which one part is a cognate with the other, and another is not. Old Norse differed from the other old Germanic languages by its terms for DIE, EARTH, SKIN, SLEEP and SUN.

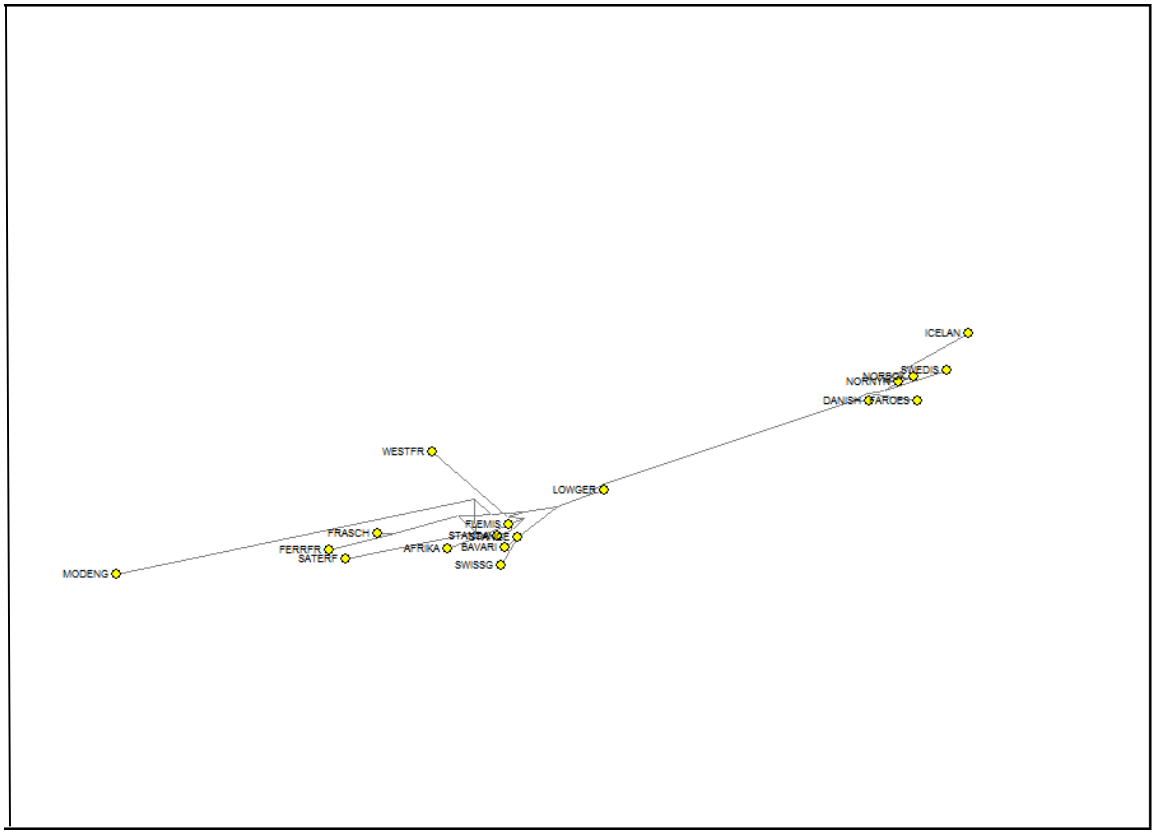
In all of these simulations the different languages have largely been grouped in the same ways, with a tripartite main division dividing the Germanic languages. Gothic and Old Norse have consistently occupied their own branches of the network; in all but the Infodelete simulation Gothic has been marginally closer to the West Germanic languages than to Old Norse. This is in marked contrast to the picture given by sound and morphological correspondences, which has led to classifications which generally posit that Gothic (East Germanic) is more distantly related to both North and West Germanic, which are shown as clustering together.

Subgroup 2: The Modern Germanic Languages

The first simulation in this subgroup was a semantically strict simulation. All of the concepts listed on the Swadesh 100-word list were attested in each modern language, which meant strategies such as those used above for the handling of missing data in the old Germanic languages were unnecessary. The network generated in this simulation is shown below.



Modern Germanic Languages, Semantically Strict Simulation, ε=0 Language n=18, Concept n=100



Modern Germanic Languages, Semantically Strict Simulation, ε=0 Language *n*=18, Concept *n*=100

This network has two distinct clusters of languages: one made up of the various North Germanic languages and the other of the West Germanic languages. Within the North Germanic group, Danish branched off basally to the rest of the group due to its using *spise* for EAT; this was originally borrowed from Middle Low German (ODS, 2015). Of the North Germanic languages, Danish had the most in common with the West Germanic languages in terms of basic vocabulary, largely due to the larger number of terms borrowed from Middle Low German which have become standard parts of the Danish vocabulary. The rest of the North Germanic languages differed from Danish by the word for BONE, which in Danish is the Middle Low German loan *knogle*; the other languages have all preserved a descendent of the original Old Norse *bein*. Faroese branched off on its own based on the words for EARTH, MOUNTAIN and WARM. Swedish, Norwegian and Icelandic all shared cognates for the word for SEED (all words with the core structure /fr(j)$V_{[+ro]}$/) which were not used in either Faroese or Danish, and which caused them to be grouped off of a branch above both Faroese and Danish. Swedish branched off next based on its words for CLOUD, KILL, LEAF and TAIL, which were *moln*, *döda*, *löv* and *svans*. Icelandic, Norwegian Bokmål and Norwegian Nynorsk branched off from the common branch they shared with Swedish based on the concept BELLY, for which they all had cognate words and Swedish did not. The two Norwegian varieties differed from each other in this simulation based on each one's word for

EAT, with Norwegian Bokmål having *spise* (due to the heavy historical influence of Danish in its development) and Norwegian Nynorsk using *eta*; this last is descended from Old Norse *eta* and is cognate with Swedish *äta* and Faroese *eta*. Icelandic branched off from these based on its words for EAT, EARTH, LEAF, MOON, MOUNTAIN, ROUND and WARM. Of these, EAT (*borða*), MOON (*tungl*), ROUND (*kringlóttur*) and WARM (*hlýr*) were unique to Icelandic amongst the modern Germanic languages. EARTH and MOUNTAIN were shared by Icelandic and Faroese (in both cases *mold* and *fjall*, respectively). The Icelandic for LEAF, *laufblað*, was coded as a cognate of Swedish *löv* for this simulation; however, the second element in this compound is a cognate of the Norwegian, Danish and Faroese words for LEAF– this placing is thus due to an arbitrary choice in how to code this item. Items such as this may present formidable problems for a method like this; this will be discussed further on.

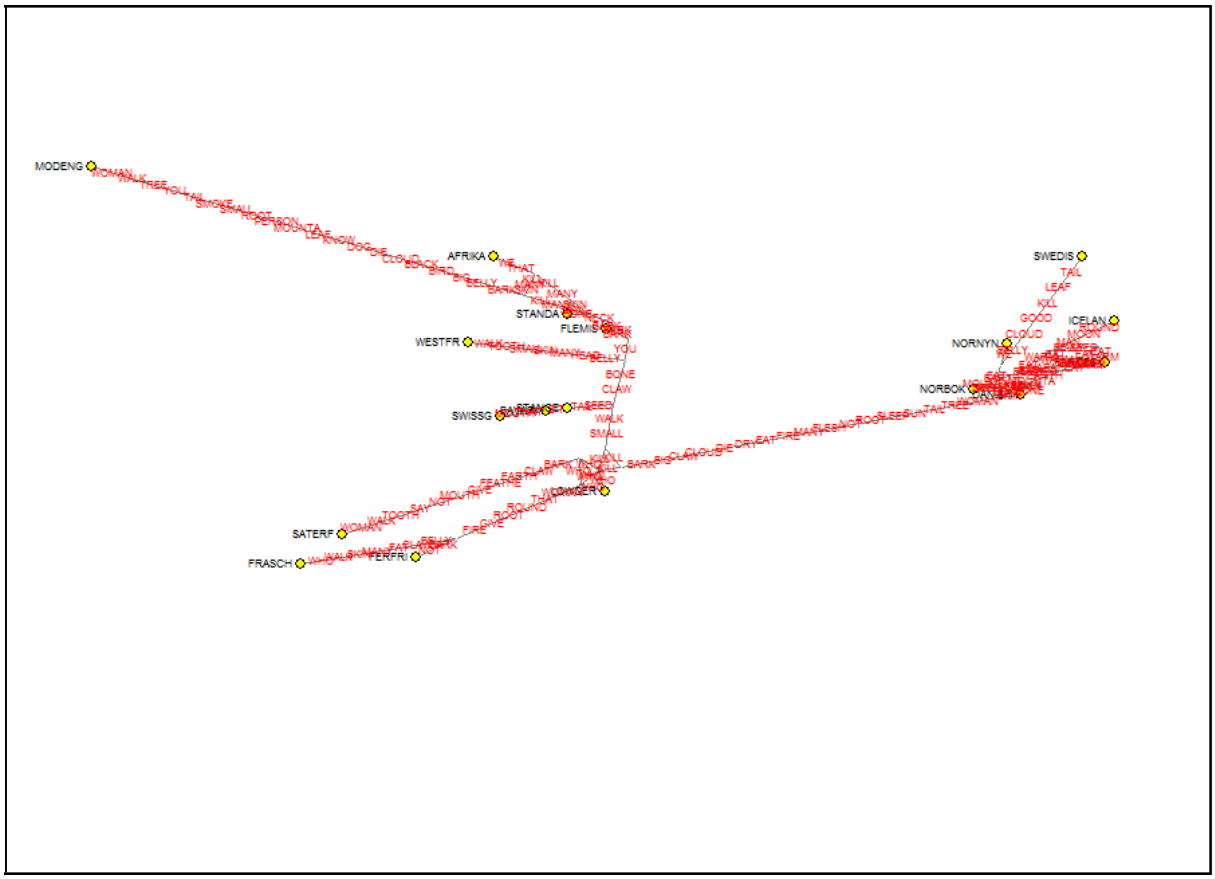
Low German branches off basally to the rest of the West Germanic languages, differing from the branches immediately to the left of it by KILL, SMALL and WALK. The first branch of the rest of the West Germanic languages to diverge from this basal branch is that on which the High German varieties lie. These differ from the preceding branch by their words for the concepts HEAD, SEED and TAIL. In all of the High German varieties used in this simulation, the everyday word for HEAD has the structure /kVpf/, as in Standard High German and Swiss German *Kopf* and Bavarian *Koopf*; this form is only found outside of this branch as the neutral, everyday word for HEAD in Afrikaans (*kop*). For SEED they have a word with the core structure /$C_{[+alv,+fric]}$VmV/ (Standard High German *Samen*, Bavarian *Saama* and Swiss German *Saame*), and for TAIL all had *Schwanz*. Under this condition Bavarian differed from Standard High German by only one concept, WOMAN, where Bavarian had *Weiberz* and Standard High German *Frau*. Swiss German was shown as being more divergent in terms of basic vocabulary, with differences in the words for the concepts MOUTH and THAT; in the case of the former this is due to the fact that the Swiss German for MOUTH is *Muul*, a cognate of Standard High German *Maul*, "snout", while in the case of the latter this is due to the fact that Swiss German uses *säb* (a form of the reflexive pronoun) in the place of the demonstrative *das*. The next division in the network separated the Frisian varieties, English, Standard Dutch, Flemish and Afrikaans from the High German varieties based on the concepts BONE, CLAW and STAR; this is because for BONE all of the varieties above this division use a word descended from Proto-Germanic **bainaz*, while those below it use a word with the core structure/$C_{[+vel/pal,-vo]}$nV $C_{[+vel/pal,-vo]}$/; for CLAW all of these varieties retained words descended from Proto-Germanic **klawâ~*klǣwâ-* (OED, 2015) or **klawjanan* (Orel, 2003), while the High German varieties make use of a different word attested only from the 16th century (*Kralle*); for STAR, all of the West Germanic varieties emanating from this branch make use of a word lacking a nasal, while the High German varieties have the nasal (see Appendix A). This branch divides again, with Modern English, Frasch Frisian, Saterland Frisian and Ferring Frisian being shown as related on the one hand, and West Frisian, Standard Dutch, Flemish and Afrikaans occupying the second branch. West Frisian was shown in this simulation as diverging from the same branch as Standard Dutch and its near relatives; this suggests that overall basic vocabulary similarity between West Frisian and Dutch is significantly higher than between it and the other Frisian varieties. West Frisian branches off based on the words for the concepts BELLY, HEAD, MANY, SMALL,

TOOTH and WALK. Of these, HEAD (*holle*), MANY (*in soad*) and WALK (*rinne*) occurred only in West Frisian, while SMALL (*lyts*) and TOOTH (*tosk*) had cognates in other Frisian varieties (in the case of the former there were cognates in all of the other Frisian varieties, and in the case of the latter Saterland Frisian had the cognate *Tusk*). Flemish, Standard Dutch and Afrikaans all shared the same innovation regarding the concept YOU, by which they diverged from their common point with West Frisian. Standard Dutch diverged from the branch occupied by Flemish and Afrikaans based on the words for the concepts BARK (*schors*) and BONE (*bot*). Both Flemish and Afrikaans had *bas*(*t*) for BARK and *been* for BONE. Afrikaans differed from Flemish by the concepts WE,THAT, KILL and HEAD; Flemish shares the same terms as Standard Dutch for these.

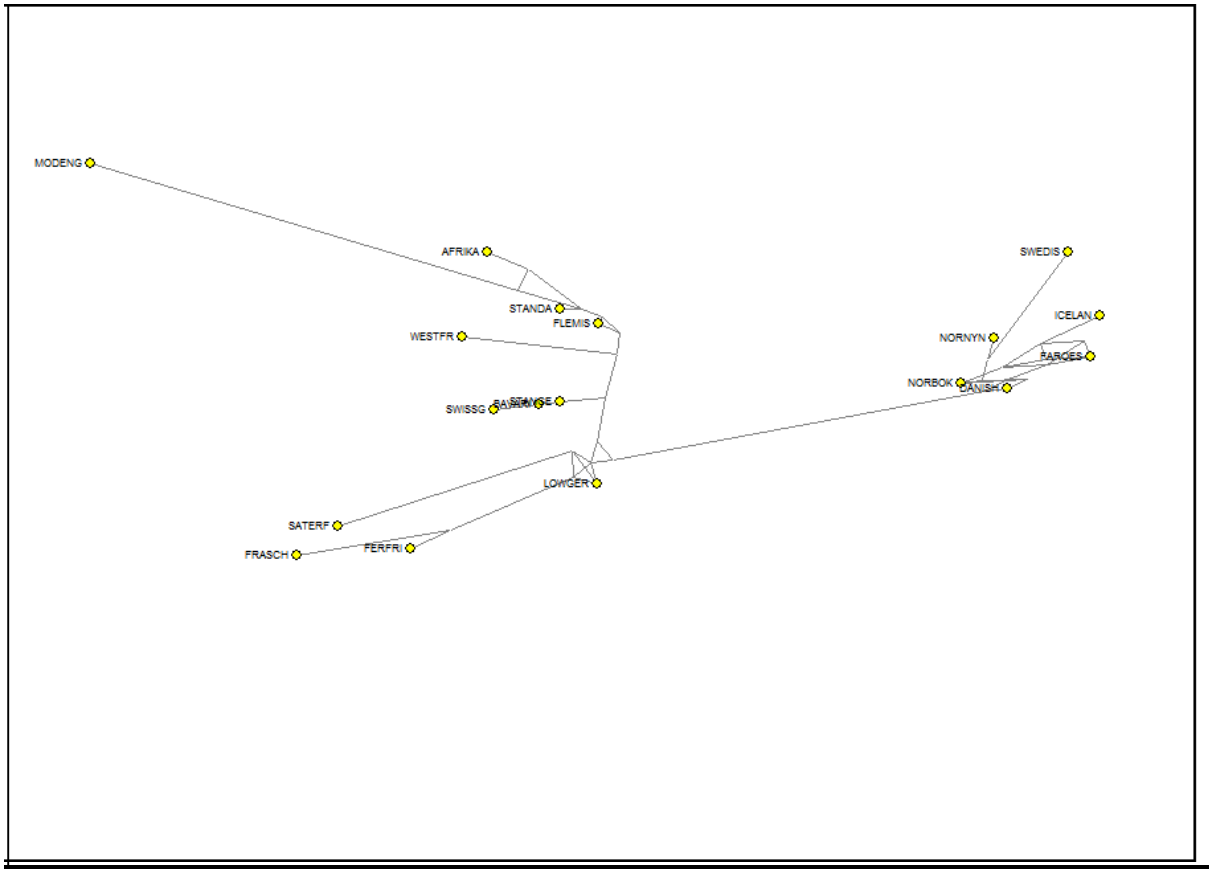
Modern English, Ferring Frisian, Frasch Frisian and Saterland Frisian all branch off from the branch which connects them to the other West Germanic varieties based on the concept WOMAN; all of these languages use a word descended from Proto-Germanic ***wiƀan***, while all of those within the West Germanic group excluded from this division use a variant of *Frau*/*vrouw*, which originally had the meaning "noble woman". A word which also appears to mark these varieties out within the West Germanic varieties, and which is shared with West Frisian is the word for SMALL (of diminutive size), for which Modern English has *little*, Ferring Frisian *letj*, Frasch Frisian *latj* and Saterland Frisian *litje*/*litjet*. Aside from West Frisian, all of the West Germanic varieties not grouped above this division use a word with the structure /$C_{[+obst]}$lV(n)/, such as Standard High German *klein*, Standard Dutch *klein*, Bavarian *gloa* and Swiss German *chlai*. The division which separated modern English from the northern and eastern Frisian varieties occurred with the word for the concept KILL, where these Frisian varieties all had "dead+make" constructions, while English had the word *kill*, the etymology of which is uncertain, and which appears to be unattested in all of the other Germanic languages. Despite being shown as having the most similarities to the northern and eastern Frisian varieties, Modern English was still shown as diverging greatly from the other West Germanic languages; it differs from the languages beneath its immediate branching point by twenty-three terms. Of these thirteen occurred exclusively in Modern English as the normal, everyday words for their concepts; these were KILL (*kill*), BELLY (*belly*), BIG (*big*), BIRD (*bird*), BLACK (*black*), CLOUD (*cloud*), DOG (*dog*), KNOW (*know*), MOUNTAIN (*mountain*), PERSON (*person*), SMOKE (*smoke*), TAIL (*tail*), YOU (*you*) and WALK (*walk*). Two of these are definite borrowings (*mountain* and *person*), and another is of uncertain etymology but has been suggested as a possible borrowing (*kill*, which has been suggested as a possible Old Norse loan, although the word is not actually attested in Old Norse (OED, 2015)). The other ten are all known to be native terms attested from the Old English period (OED, 2015); these therefore seem to be innovations unique to English. The northern and eastern Frisian varieties were grouped as more closely related to each other than to Modern English; in addition to KILL, the division separating them from Modern English included GIVE. Frasch and Ferring Frisian were grouped on the same branch, which separated from that on which Saterland Frisian was placed at the point of the concept GIVE; Frasch and Ferring Frisian both had cognate terms for GIVE, *düünj* and *du*, respectively, which were both derived from the verb "to do"; Saterland Frisian had *reke*, which is derived from the verb "to reach" (Hoekstra, personal correspondence). Frasch and Ferring Frisian

additionally shared terms for FIRE, MANY, ROOT and ROUND. Frasch Frisian branched off based on the word for FAT, for which it had *smeer*, unlike Ferring Frisian which had *feet*. Ferring Frisian was shown as having had changes from the sequence posterior to it for the concepts CLAW (*kral*), NOT (*ei*), WALK (*gung*) and WHO (*hoker*). Of these, NOT has a cognate in Saterland Frisian, CLAW cognates in the High German varieties, and WALK cognates across the North Germanic languages. Saterland Frisian branched off form the sequences posterior to it in the network based on the words for the concepts BARK (*Boarke*), EARTH (*Gruunde*), FEATHER (*Fugge*), MOUTH (*Mule*), NOT (*ai*), SAY (*kwede*), TOOTH (*Tusk*) and WHO (*wäl*). Of these, six had cognates with the same meanings in other languages.

The second simulation in this subgroup used additional words from the modern Germanic varieties which had the same meanings as those in the above simulation. This resulted in the following network.



Modern Germanic Languages, Semantically Strict, Other Words Simulation, ε=0, Language n=18, Concept n=100



Modern Germanic Languages, Semantically Strict, Other Words Simulation, ε=0, Language *n*=18, Concept *n*=100

The use of words which had the same or very similar meanings to those in the first simulation, but which were different resulted in a network which was still divided into two distinctive clusters of languages, the North Germanic and West Germanic. However, within these groups the placing of languages differed markedly. Danish was still placed basally to the rest of the North Germanic languages, although this time it was shown as branching off based on the concept SKIN, which it shared with some of the languages further in the network (Norwegian Nynorsk and Swedish); it differed from the Norwegian varieties by the concept BELLY (*bug* instead of *mave*) . It was also depicted as being more closely related to both varieties of Norwegian than in the previous simulation. Norwegian Nynorsk differed from Bokmål not only by the concepts EAT and SKIN but also by WE, which, along with BELLY, resulted in a division between it and Swedish. Swedish was placed as being a considerable outlier with it branching off based on the concepts BELLY, CLOUD, GOOD, KILL, LEAF and TAIL. Icelandic was grouped on the opposite side of the North Germanic languages to the Norwegian varieties, which in the previous simulation were shown as being its closest relatives within the group. It is now shown as arising from a complex network of reticulations, and has been grouped as a closer relative of Faroese, although it is still something of an outlier. The reason for this may be that substitiuting *karl* for *maður* for the concept MAN and coding *laufblað* as a cognate of the Danish, Norwegian and Faroese forms

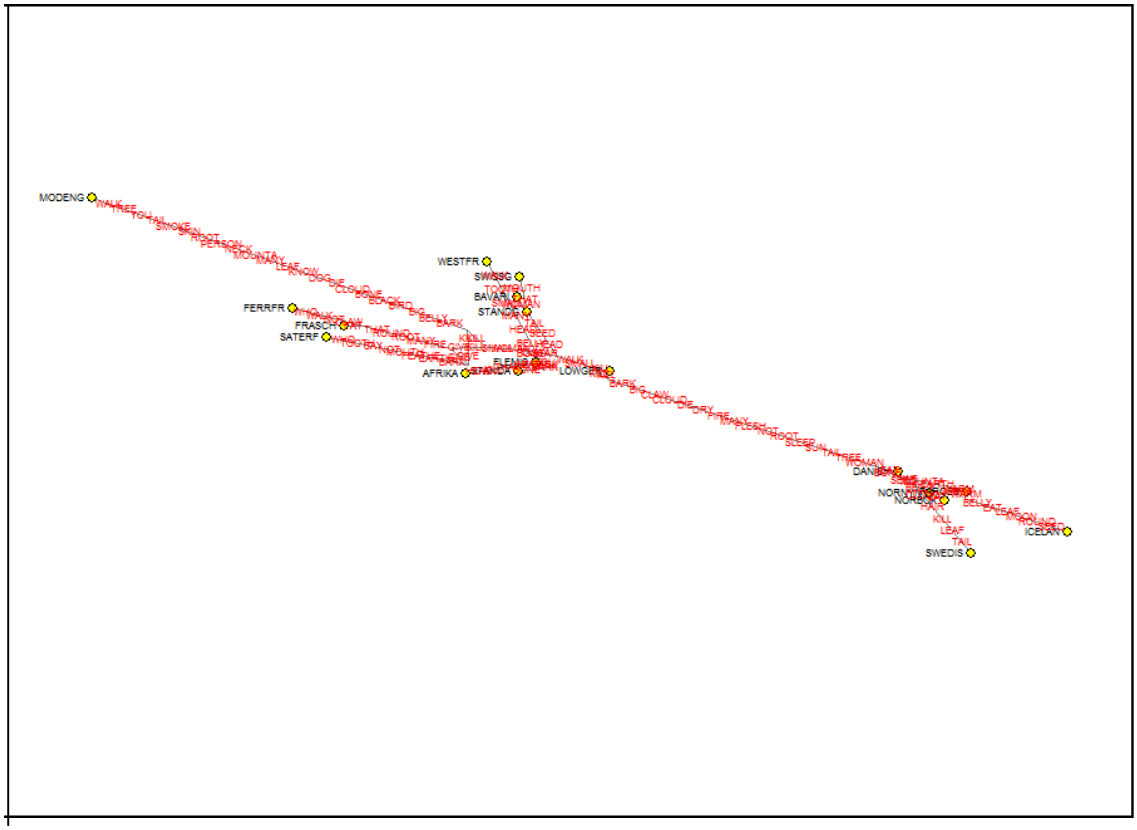
made its sequence similar enough to these sequences, which had a fairly high number of differences in vocabulary between them, to cause the algorithm to attempt to locate it in such a way that it would be situated nearer to the more basal sequences, while at the same time the dissimilarities between its sequence and these caused it to be placed as an outlier.

Within the West Germanic group the greatest change brought about by using different lexical items was that Modern English, Saterland Frisian, Frasch Frisian and Ferring Frisian were no longer grouped as emanating from a common node. Modern English was shown as an outlier emanating from a point from which Standard Dutch and Afrikaans also emerged, while the northern and eastern Frisian varieties were grouped as being more closely related to Low German than they were in the previous simulation. By substituting *skors* for *bas* for the concept BARK in Afrikaans, Flemish differed from both of them by the concept BARK and was thus placed basally to them. Standard Dutch branched off from the branch it shares with Modern English and Afrikaans due to the concept BONE, for which it has *bot* while the others have a word with the structure /bVn/. Afrikaans, Modern English and the branch including both Standard Dutch and Flemish all had different words for the concepts KILL, MANY and SKIN. In the case of MANY, Afrikaans had *baie* substituted for *veel*, while for SKIN *vel* was used instead of *huid*. Afrikaans also differed from the other languages because of the concepts THAT and WE. Modern English was grouped with Afrikaans, Standard Dutch and Flemish on the basis of the concept NECK, for which the word originally used with them, *hals*, was replaced by *nek*, which was coded as a cognate of Modern English *neck*. Modern English differed from the languages immediately posterior to it in the network by the same twenty-three concepts as in the previous simulation. West Frisian was grouped as more closely related to Dutch and its close relatives that it was to the other Frisian varieties, although it was still fairly divergent. The relationship between the High German varieties and the rest of the West Germanic languages, with the exceptions of Low German and the northern and eastern Frisian varieties, has not changed much, and the division between them and West Frisian, the Netherlandic (including Afrikaans) varieties and English was based on the words for SEED and TAIL (for which the High German varieties had words which were not found in the other West Germanic languages).CLAW and BONE were the concepts which, in this simulation, distinguished West Frisian, the Netherlandic varieties and Modern English from the rest of the languages in the West Germanic section of the network. The relationships between the High German varieties has remained largely unchanged; this is due to the fact that, even though some of the codes for concepts were changed, as different words were used, the number of cognates remained the same between them. As the algorithm using the Hamming distance, or number of different characters between sequences, to generate the network the fact that different words were used thus made no difference, as this did not alter the Hamming distance.

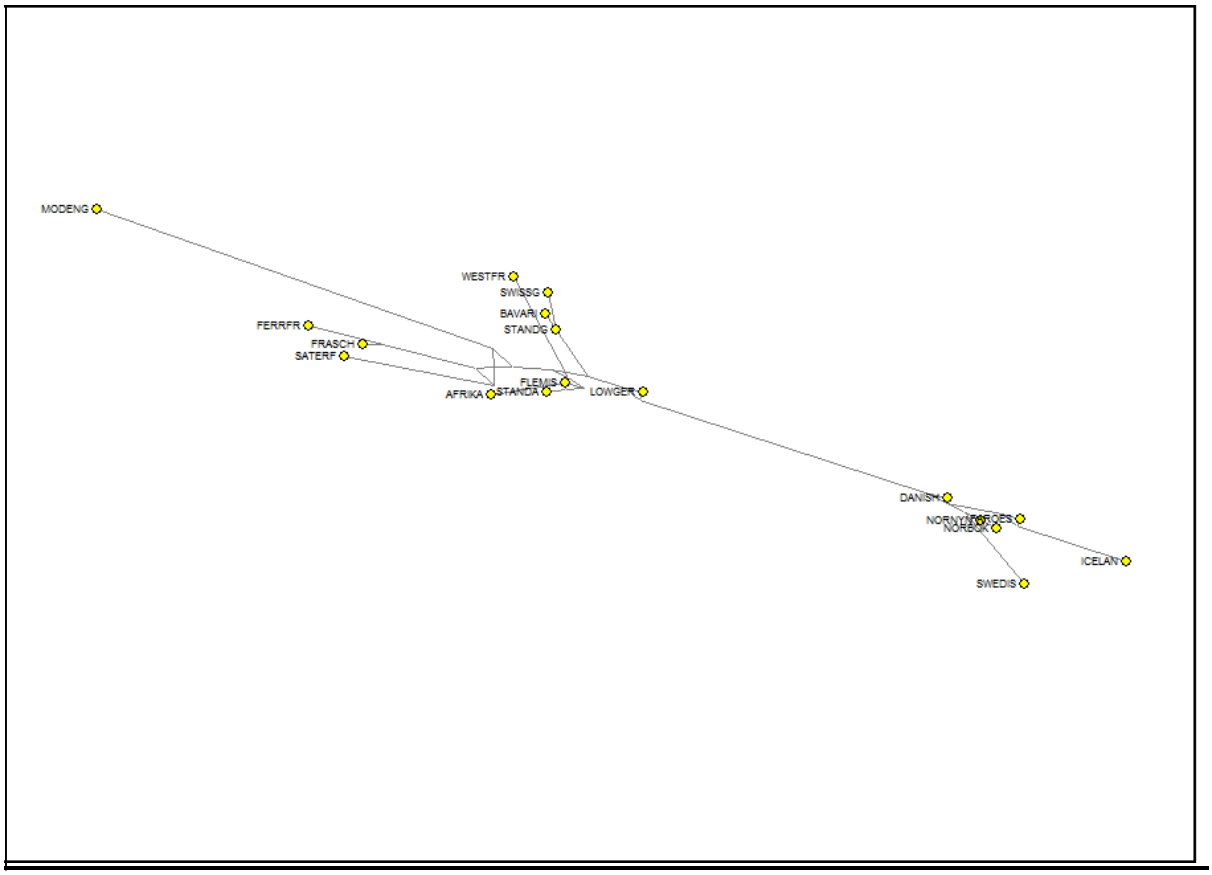
The grouping more closely together of Low German, Saterland Frisian and Ferring Frisian was due to the coding of their words for the concepts KILL and WHO as cognates; in the case of the former, where they all made use of a *dead + make* construction, this was done because of this similarity, while in the latter case all three languages had words which originally meant “which” and which came to mean “who”. Despite not having a cognate

amongst these for WHO, the Hamming distance of Frasch Frisian from Ferring Frisian was still low enough for the algorithm to group them together. As the northern and eastern Frisian varieties were shifted in the network to a region where they branch off from a different set of sequences, the concepts by which some of them were shown as different from those languages below them in the network changed. Whereas in the previous simulation Frasch Frisian was shown as branching from the same branch as Ferring Frisian by only the concept FAT, in this one it branches off by BELLY, CLAW, FAT, MANY, SKIN, WALK and WHO. This is due to the fact that the sequences which are posterior to it and to Ferring Frisian are now different to those in the previous simulation; this has resulted in some words being treated as innovations in this simulation when they were treated as retentions from an ancestral sequence in the last one. Ignoring words which were changed for this simulation, this has effectively caused some of the concept changes to be swapped over as innovations from Ferring Frisian to Frasch Frisian (WALK, WHO and CLAW). This may have serious implications for which data is selected for use in this method. A similar scenario led to the marking of CLAW, GIVE, WALK and WOMAN as innovations on the Saterland Frisian branch in this simulation; these were not marked as such in the previous simulation.

The third simulation performed on the modern Germanic languages was a semantically strict semi-cognate simulation, in which words which were compounds containing both cognate and non-cognate elements were assigned different codes, rather than the same codes as was the case in the previous simulations. This led to the network below.



Modern Germanic Languages, Semantically Strict Semi-Cognate Simulation, ε=0 Language n=18, Concept n=100



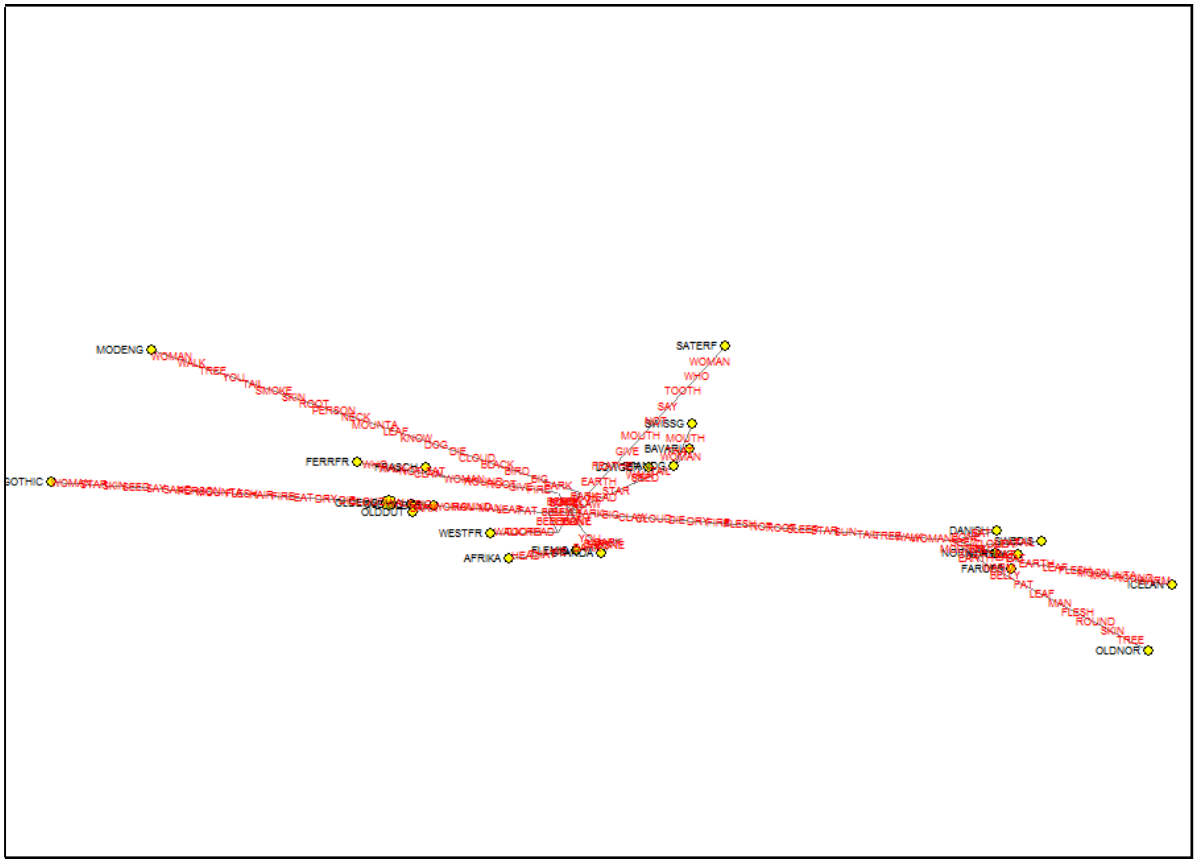
Modern Germanic Languages, Semantically Strict Semi-Cognate Simulation, ε=0 Language *n*=18, Concept *n*=100

This simulation resulted in a network which was largely the same as that in the first semantically strict simulation for the modern Germanic languages. Within the North Germanic section of the network, the coding of Swedish *hårstrå* as being a non-cognate of the other words in the data, based on the presence of *strå*, did not radically alter Swedish's position relative to the other North Germanic languages; it merely lengthened its branch slightly. This indicates that this small change did not alter the Hamming distance enough to have a significant impact on how Swedish was placed in the diagram. On the other hand, the coding of the Faroese and Icelandic words for WE (*vit* and *við*, respectively) as cognates of each other but not of the words in the rest of the data set based on the fact that they were formerly dual pronouns had a significant impact on how Faroese was placed. In the first simulation Faroese was placed far below Icelandic and nearer to Danish; this was due to the fact that, overall, in the first simulation Icelandic, Norwegian and Swedish had greater sequence similarity because of the number of cognates they shared; this meant that its Hamming distance from Danish was lower than it was from many of these. However, it still shared a number of cognates with Icelandic which neither shared with the other North Germanic varieties. By recoding WE so that both Icelandic and Faroese had the same codes for this concept, the Hamming distance between these was lowered enough so that the position of Faroese was shifted so that it was closer to Icelandic and formed a branch separate

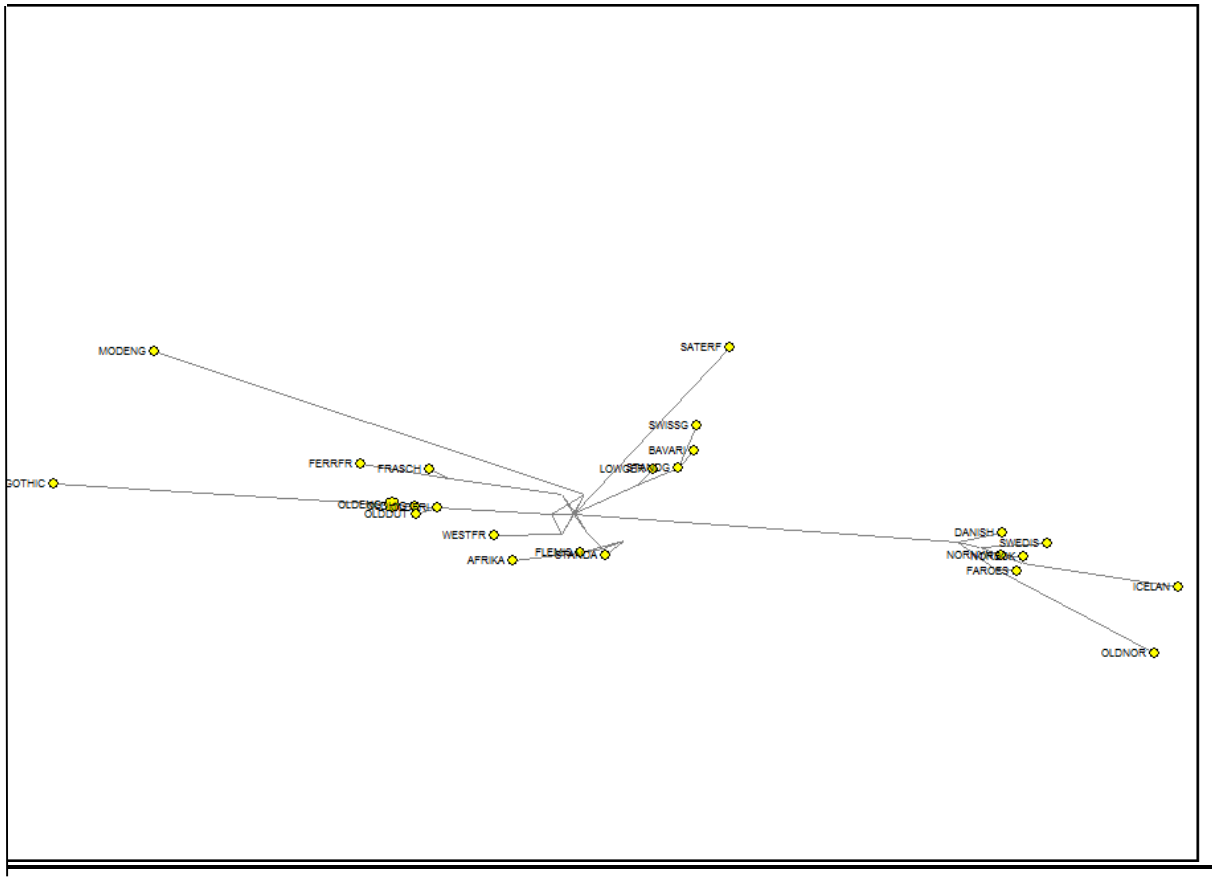
from that of Norwegian and Swedish. Within the west Germanic languages the placing of languages relative to each other has not changed at all; the recoding of the Frasch and Ferring Frisian words for THAT, *dåt*(*deer*) and *detdiar*, as cognates of each other but not of anything else merely resulted in the concept THAT being added to the branch which separates them from Saterland Frisia and Modern English. This indicates that this did not alter their Hamming distances from the other varieties sufficiently to result in their being moved elsewhere in the network.

Subgroup 3: Old and Modern Germanic Languages– Combined Simulations

These simulations included data from both the old and modern Germanic languages. They effectively combined earlier simulations from the previous two groups as outlined in the methods section. The first of these simulations was a Majority Wins semantically strict simulation. This simulation was based on the first and seventh simulations. The network below was generated based on this data.



Old and Modern Germanic Languages, Majority Wins, Semantically Strict Simulation, ε=0
Language n=25, Concept n=99



Old and Modern Germanic Languages, Majority Wins, Semantically Strict Simulation, ε=0
Language *n*=25, Concept *n*=99

With the exception of Old Norse, the old Germanic languages all occupied one main branch of the network. Within this branch, the old West Germanic languages clustered very closely together, while Gothic was an outlier node. Gothic differs from the languages in the node directly beneath it, Old Saxon and Old English (these were represented as one large node due to the fact that under the Majority Wins condition their sequences became identical, and so were treated as the same sequence occurring twice), by fifteen concepts. These were CLOUD, DIE, DRY, EAT, FIRE, HAIR, FLESH, MOUNTAIN, PERSON, SAND, SAY, SEED, SKIN, STAR and WOMAN. Of these, DIE, EAT, FIRE, HAIR, FLESH, MOUNTAIN, PERSON, SAND, SAY, and SKIN did not occur in any other languages used in this simulation. The Gothic for CLOUD, *milhma*, had as a cognate Swedish *moln*, although this did not have any significant influence on the placing of Gothic or Swedish; neither did the fact that Gothic and Old Norse had cognates for DRY (*þaursus* and *þurr*) and WOMAN (*qino* and *kona*). This did not result in great enough changes in their Hamming distances for the algorithm to group them with each other. Old English and Old Saxon differed from Old High German by the concept WALK; these and Gothic used words with the structure /$C_{[+vel,\ +obst]}V(C_{[+nas]})$/ (such as Old English *gān*/*gangan*) while Old High German, Old Dutch and Old Frisian had words with the core structure /l$VC_{[+lab]}$/ (such as Old High German *loufan*). Old High German and the languages occurring above it differed from Old Dutch and Old Frisian by the concept BIG; this was due to the fact that, for this simulation, Old High

German, Old English, Old Saxon and Gothic all used words with the core structure /mV$_{[+high,+front]}$C$_{[+obst, -vo]}$Vl/, such as Old High German *mihhil* and Old English *micel*, while Old Dutch and Old Frisian used *grōt* and *grāt*. Old Dutch branched off based on the concept BARK, for which it had *skorsa*; under the "majority wins" condition Old Frisian was given the same code for this character as Old English *rind*(*a*). This section of the network emerged from a series of reticulations from which also emerge, at different points, the modern West Germanic languages. The section of the diagram which is made up of the old West Germanic languages and Gothic branches out of this reticulate network based on the concepts BELLY, MANY, BONE, FAT, LEAF, MAN, ROUND and WOMAN. The words with the core structure /wV$_{[+low]}$m (b)V/ have not survived into the modern Germanic languages with the meaning BELLY; neither have the words for MAN with the structure /C$_{[-obst,+lab]}$V$_{[+mid]}$r/, except in words such as *werewolf* in which they are fossilised. Particular words unambiguously meaning ROUND were not attested in most of the old West Germanic languages, as well as Gothic; under the Majority Wins condition the code for Old English *sinwēal* and Old High German *sinwel* was applied to these languages. Neither of these forms has survived to modern times. In the case of BONE, with the exception of Modern English, a semantic shift has replaced the original words for BONE in the West Germanic languages with other forms; in most instances the original forms have developed the primary meaning of "leg". Similarly, the original words for FAT (such as Old English *smēoru* and Old High German *smero*) have changed meaning and been replaced by words with the form /fVt/. For WOMAN the old form descended from Proto-Germanic ***wiƀan* has in many instances survived; however in Standard Dutch, Flemish, Low German, Standard High German, West Frisian and Swiss German it has taken on distinctly pejorative connotations, and has been ousted by words with the form /frV/.

West Frisian was shown this time as emerging from a reticulate section at the point where the cluster of the old West Germanic languages and Gothic emanate from a series of reticulations. This may be due to the fact that some of the old West Germanic varieties, particularly Old Dutch and Old Frisian, had a very high number of terms under this condition which were cognates of West Frisian terms; as West Frisian was still quite divergent in the previous simulations with the modern Germanic languages, despite apparently being closest to the Netherlandic varieties, it is probable that with the inclusion of the old Germanic languages a high number of cognates shared between West Frisian resulted in the algorithm having to reposition it so as to be closer to these. West Frisian branched off on its own based on the words for the concepts BELLY (*liif*), TOOTH (*tosk*), HEAD (*holle*) and WALK (*rinne*). Modern English has diverged significantly from all of the Germanic languages; in this simulation it differed from the sequences making up the reticulations immediately posterior to it by twenty-one terms. Of these, eleven occurred only in Modern English in this simulation with the meanings used in compiling the Swadesh 100-word list. These were BELLY (*belly*), BIG (*big*), BIRD (*bird*), BLACK (*black*), CLOUD (*cloud*), DOG (*dog*), KNOW (*know*), MOUNTAIN (*mountain*), PERSON (*person*), SMOKE (*smoke*), TAIL (*tail*) and WALK (*walk*). Four of these had the same meanings in Old English, but were not used in the Old English data for this simulation; these were KNOW, BLACK, SMOKE and TAIL.

Afrikaans differed from Flemish by the concepts WE, THAT and HEAD; both branched off from the sequence common to them and Standard Dutch by the concept BARK, for which they had *bast* (Flemish) and *bas* (Afrikaans), while Standard Dutch differed from the sequences posterior to it by the concept BONE, for which it had *bot*. All three branched off from the network based on the concept YOU, the words for which (Standard Dutch *jij*, Flemish *jij*/*gij* and Afrikaans *jy*) are an innovation by the common ancestor of these three varieties.

Saterland Frisian emerges from the reticulate area of the network from a different point to the branch on which Ferring and Frasch Frisian are located. This is due to differences in the words for MANY and FIRE in their sequences; for the former Saterland Frisian has *fúul* while Ferring and Frasch Frisian have *m*-forms (*manigen* and *maning*), and for the latter Saterland Frisian has *Fjúur* while Ferring and Frasch Frisian have borrowings from Old Norse (*ial* and *iilj*). Saterland Frisian differs from the sequences making up the reticulation by the concepts BARK, EARTH, FEATHER, GIVE, MOUTH, NOT, SAY, TOOTH, WHO and WOMAN. Of these, BARK had cognates in Modern English and the North Germanic varieties, MOUTH in West Frisian and Swiss German, NOT in Ferring Frisian, TOOTH in West Frisian, and WOMAN in the old West Germanic varieties, Modern English, Ferring Frisian, Frasch Frisian and Bavarian. Its overall sequence similarity to Standard Dutch and its relatives was still high enough for it to be grouped nearby them in the network. Ferring and Frasch Frisian branched off from the area of reticulation based on MANY, FIRE, GIVE, ROOT, ROUND, and WOMAN. ROOT had cognates in Modern English, Old Norse, and the modern North Germanic languages, while WOMAN had cognates in the old West Germanic varieties, Modern English and Bavarian. Frasch Frisian branches off the line it shares with Ferring Frisian due to the presence of the word *smeer*, for FAT, which is not cognate with the Ferring Frisian word (*feet*) and is not found in the immediately posterior sequences with the same meaning. Ferring Frisian differs from the sequences immediately posterior to it by the concepts CLAW, NOT, WALK, and WHO; CLAW has cognates in all of the old and modern West Germanic varieties with the exceptions of Frasch Frisian, the High German varieties and Low German. NOT has a cognate in Saterland Frisian. WALK has cognates in this simulation in Gothic, Old Norse, Old English, Old Saxon, Low German and all of the modern North Germanic varieties. Ferring Frisian had a unique form for WHO.

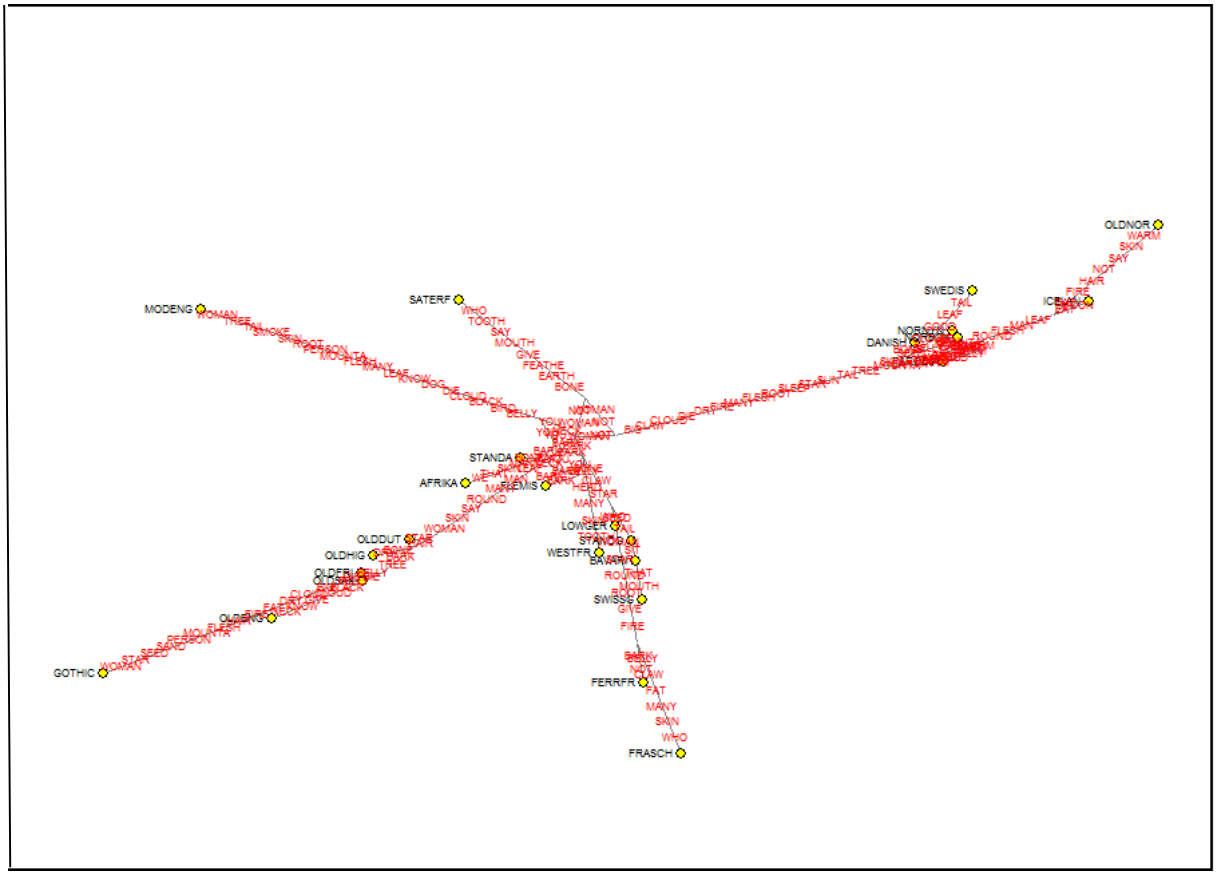
Low German and the High German varieties branched off from the West Germanic reticulation based on the concepts for CLAW, HEAD and STAR. For CLAW, which they shared with Frasch Frisian, they have words with the core structure /k*r*VC$_{[+vo,+alv,+obst]}$/, such as Standard High German *Kralle*. These words are attested only from the 16th century, and appear to be an innovation which began in High German (Duden, 2015). For STAR they have words with a root-nasal, such as Standard High German *Stern*; this is cognate with the words for STAR in Old Norse and all of the modern North Germanic languages with the exception of modern-day Icelandic. For HEAD they had words with the core structure /kVC$_{[-vo,+lab,+obst]}$/, such as Standard High German *Kopf* and Low German *Kopp*. Low German diverges from the sequences immediately posterior to it by the concept WALK, for which it has *gån*, a cognate of the North Germanic, Old English, Old Saxon, Gothic and Ferring

Frisian words; the sequences which are posterior to it, and with which it had greater overall similarity, have words with the core structure /lVC$_{[-vo,+lab]}$/. The modern High German varieties branched off based on the concepts SEED and TAIL. For SEED, the modern High German varieties all have words with the core structure /zVmV/, such as Standard High German *Samen*. For TAIL they had words with the core structure /C$_{[-vo,+pal/alv,+fric]}$vVnC$_{[-vo,+sib]}$/, such as Standard High German *Schwanz*; these had a cognate in Swedish *svans*, which is a borrowing from Middle Low German (SAOB, 2015). Bavarian diverged under the semantically strict condition from Standard High German by WOMAN, while Swiss German diverged by THAT and MOUTH.

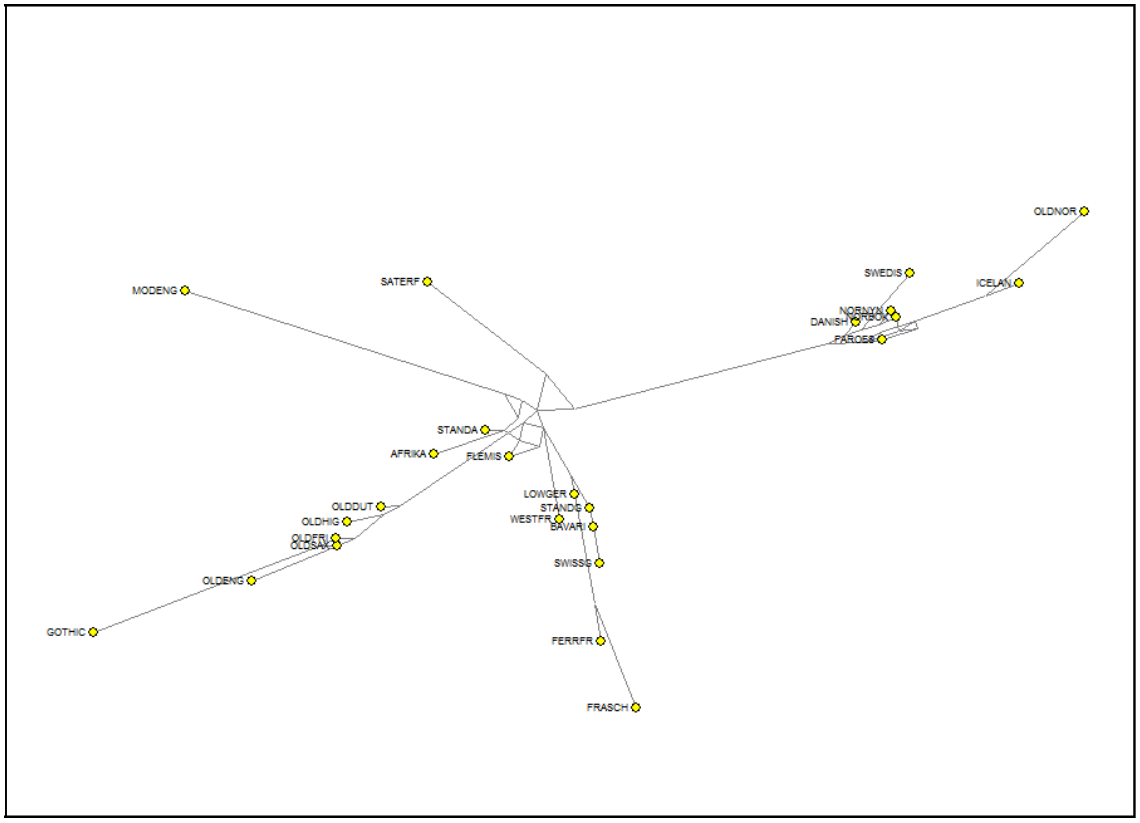
The North Germanic languages again formed their own distinct group. In this simulation the group split into three from the same point. One of these branches was occupied by Danish, which branched off from the preceding branch (along which items which distinguished all of the North Germanic languages occurred) by the concepts BONE and EAT. For BONE it had *knogle*, which is a loan from Middle Low German and which has come to replace *bein* as the neutral, default word for BONE. As in the case of BONE, EAT (*spise*) is a Middle Low German loan; this was shared by Norwegian Bokmål, which developed under heavy Danish influence. On the basis of the word for SEED for which they all had words with the structure /fr(j)V/ (such as Swedish *frö*, Norwegian *frø* and Icelandic *fræ*), both Norwegian varieties, Swedish and Icelandic formed their own branch. Within this branch Swedish was the first to branch off, with differences from the preceding sequences being recorded by the algorithm for CLOUD (*moln*), LEAF *(löv*) and TAIL (*svans*). *Moln* may ultimately be a cognate of Gothic *milhma* and *svans* is a loan from Middle Low German (SAOB, 2015). *Löv* is the retained descendant of Old Norse *lauf* (SAOB, 2015). The presence of two terms with cognates not attested in the North Germanic languages was not enough to alter its Hamming distance from these for it to be grouped outside of the North Germanic cluster. Icelandic and both varieties of Norwegian shared cognates for the concept BELLY which did not occur in the posterior sequence, namely words with the structure /mVgV/ (Icelandic *magi* and Norwegian *mage*). They each differed from each other by the term for EAT: Norwegian Bokmål had *spise*, Norwegian Nynorsk had *eta* and Icelandic had *borða*. Of these, the Nynorsk form is a continuation of the Old Norse word, while Bokmål *spise* is a loan, via Danish, from Middle Low German, and Icelandic *borða* is the product of a semantic shift from "table" to "eat" (Axelson, personal correspondence). In addition to EAT, Icelandic differs from the preceding sequences by the concepts EARTH (*mold*), LEAF (*laufblað*), FLESH (*hold*), MOON (*tungl*), MOUNTAIN (*fjall*), ROUND (*kringlóttur*) and WARM (*hlýr*). Of these EARTH, FLESH, MOUNTAIN and ROUND have cognates in this simulation in Old Norse. LEAF, *laufblað*, was coded as a cognate of Swedish *löv* rather than Norwegian *blad*; this resulted in it being recorded by the algorithm as a change in the sequence. Old Norse and Faroese branched off together from the posterior sequence based on MOUNTAIN and EARTH; in the case of the former both had *fjall*, while for the latter they had *mold*. Faroese branched off from this sequence due to the word it had for WARM, *heitur*, which was not attested in any of the other Germanic languages with this meaning. Old Norse branched off from the preceding sequence based on BELLY (*magi*), FAT (*smjǫr*), LEAF (*lauf*), MAN (*verr*), FLESH (*hold*), ROUND (*kringlóttr*), SKIN (*skinn*) and TREE (*baðmr*).

On addition to MOUNTAIN and EARTH, FLESH, LEAF and ROUND had cognates in Icelandic. Despite this, in this simulation the overall sequence similarity between Old Norse and Faroese was greater than between Old Norse and Icelandic (which had an overall greater similarity to Norwegian), resulting in this grouping. The presence of a cognate for LEAF in Swedish likewise did not cause their Hamming distances to become close enough to group them more closely together.

The next simulation used synonyms or very near synonyms in some of the languages (see the methodology section above). This resulted in the generation of the following network.



Old and Modern Germanic Languages, Majority Wins, Additional Words, Semantically Strict Simulation, ε=0 Language n=25, Concept n=99



Old and Modern Germanic Languages, Majority Wins, Additional Words, Semantically Strict Simulation, ε=0 Language n=25, Concept n=99

The inclusion of words which were synonyms or near synonyms resulted in a network which was very similar to the previous network. Like all of the previous simulations, the network could be divided into two main sections: a North Germanic section, and a section which included all of the modern West Germanic varieties, as well as Gothic and all of the old West Germanic varieties. Within the North Germanic section of the network, however, there were some changes in how the languages were arranged. The use of Norwegian Nynorsk *skinn*, Danish *skind* and Swedish *skinn* for SKIN resulted in a split in the section with Danish, Swedish and both varieties of Norwegian emanating from the same branch. Danish branched off this because of the concept BONE. The use of Swedish *bra*, a French borrowing which has become very common in Swedish (SAOB, 2015) for GOOD increased the the number of items by which Swedish differed from the sequences posterior to it. Using the Norwegian Nynorsk *me* instead of *vi* for WE resulted in it branching off based on this concept. Norwegian Bokmål was shown as differing form the sequence posterior to it by the concept SKIN, for which it still had *hud*; this meant that it did not share a cognate for SKIN with any of the other varieties along that branch of the North Germanic languages under this condition. This was the opposite of the case in the previous simulation, when it had been Norwegian Nynorsk which had not had a cognate for this concept for any of these other varieties. The second major branch within the North Germanic languages was occupied at various positions by Faroese, Icelandic and Old Norse. This branched off based on the concepts MOUNTAIN and EARTH, for which Faroese, Old Norse and Icelandic all shared cognate terms. Faroese

differed from these by the concepts WARM, BELLY, SEED, ROUND, FLESH, LEAF and MAN. Of these Old Norse *sáð*/*sæði* was replaced by *frjó* for this simulation; in the previous simulation the first term, which is cognate with Faroese *sáð*, was used. Additionally, Old Norse *maðr* and Icelandic *maður*, both cognates of Faroese *maður*, were replaced by *karl*. Icelandic branched off of this shared branch based on its words for MOON and EAT, which were the same as in the previous simulation. Old Norse branched off from the posterior sequence based on FAT, FIRE, HAIR, NOT, SAY, SKIN and WARM. Of these, FIRE (*funi*), HAIR (*fax*), NOT (*né*) and SAY (*kveða*) were different to the previous simulation. These substitutions may have decreased the Hamming distance between Icelandic and the Norwegian varieties, and Old Norse and Faroese, enough so that the similarities between them were greater than between any other sequences, resulting in this change in grouping. The same holds for Swedish and both Norwegian varieties in relation to Danish. This suggests a grouping in which the North Germanic languages are divided into Continental (covering Norwegian, Swedish and Danish) and Insular (Faroese and Icelandic), with Icelandic and Old Norse also generally held to be more similar to each other than to any other North Germanic varieties (Henrikson & van der Auwera, 1994).

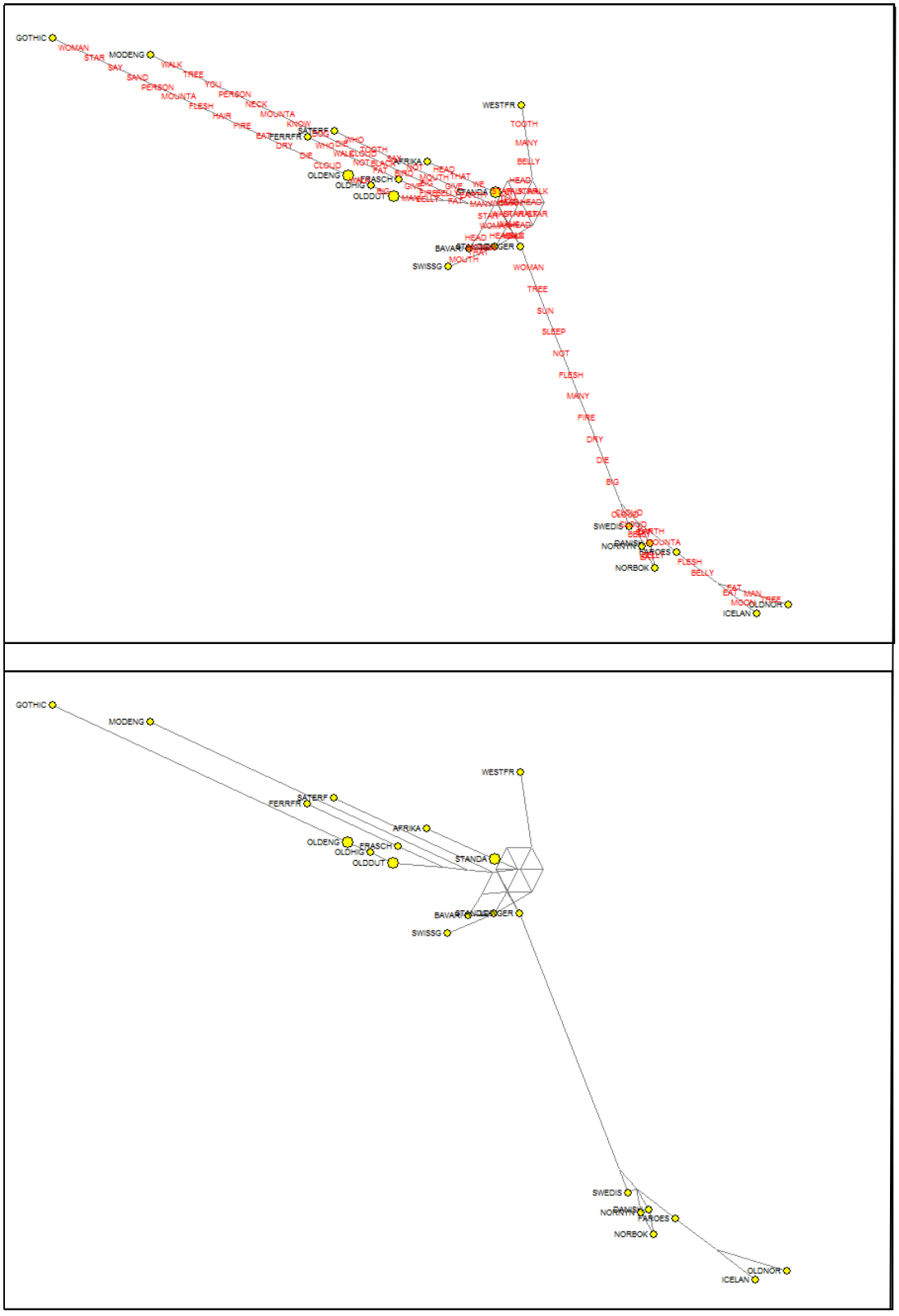
Saterland Frisian branched off from both the sequences which created the area of reticulation amongst the West Germanic varieties and the sequence of concepts which differentiated the North Germanic group by the concepts NOT and WOMAN. In addition to these it branched off based on BONE, EARTH, FEATHER, GIVE, MOUTH, SAY, TOOTH, and WHO. Of these, none was different to the previous simulation. It is likely that the difference in position of Saterland Frisian relative to other varieties in the network is due to the effects of changing lexical items in other languages; the resultant changes in coding would have made their overall similarities to each other change, leading to different groupings. Modern English branched off quite dramatically from the network based on seventeen concepts. These were YOU, BELLY, BIRD, BLACK, CLOUD, DIE, DOG, KNOW, LEAF, MANY, FLESH, MOUNTAIN, PERSON, ROOT, SKIN, SMOKE, TAIL, TREE, WOMAN. Of these, only FLESH underwent a replacement of term for this simulation: *flesh*, which has cognates in a number of the other West Germanic languages, such as Standard High German *Fleisch*, was replaced by *meat*. This does not appear to have done much to influence the placing of Modern English on its own, and has merely lengthened the branch on which the node sits. However, the substitution of *nek* for *hals* in Afrikaans and Standard Dutch does appear to have resulted in these two languages being positioned as attached to the Modern English branch at a point just where it emanates from the area of reticulation at the centre of the West Germanic section of the network. These both branch off due to their words for YOU, which are not related to either the Modern English or other Germanic forms. As Afrikaans *bas* was replaced by *skors* for BARK in this simulation, which is cognate with Standard Dutch *schors*, this was recorded as a shared retention by the algorithm. From this point Standard Dutch branched off due to the concept BONE, for which it had *bot* and Afrikaans branched due to the words for the concepts MANY, SKIN, THAT, WE. Of these, MANY and SKIN were replacements; for MANY, *baie*, from Malay *banjak*, was substituted for *veel*, and for SKIN *vel* was substituted for *huid*. Flemish was grouped away from Afrikaans and Standard Dutch, seemingly on the basis of having retained *hals* for NECK and *bast* for BARK.

West Frisian branched out from the reticulation from a point between the branch on which Low German and the other Frisian varieties were grouped, and the section which terminates in Flemish. West Frisian diverged from this point by the concepts BELLY, HEAD, MANY, SKIN and TOOTH. None of these had any substitution take place. The branch of the network from which the High German varieties, Low German, and the other Frisian varieties descend differed from the posterior sequences making up the reticulation by BONE, CLAW, and STAR.The High German varieties were distinguished based on SEED and TAIL, as in the previous simulation. Bavarian *Weiberz* was replaced by *Frau*, which led to it no longer being differentiated by the concept WOMAN. As Bavarian and Swiss German both had cognates which were not found in Standard High German for SIT, namely *hockä* and *hocke*, these were both differentiated from it by the concept SIT. As before, Swiss German differed from both Bavarian and Standard High German by the concepts THAT and MOUTH. Low German, Frasch Frisian and Ferring Frisian were differentiated from the posterior sequences by the concept WALK; all of these used a *g*-form rather than an *l*-form. Ferring Frisian and Frasch Frisian differed from the Low German sequence by WOMAN, STAR, ROUND, ROOT, GIVE, FIRE, BARK and NOT. Of these, none substituted by any other terms. Frasch Frisian branched off from Ferring by the concepts CLAW, FAT, MANY, SKIN, and WHO. MANY was shown here as a difference due to the substitution of Ferring Frisian *manigen* with *fölen*, which is a cognate of the terms for MANY in the sequences posterior to it. This resulted in Frasch Frisian *maning* having no cognate in a sequence immediately posterior to it. Frasch Frisian *hüd*, SKIN, was substituted for by *schan*, which is an early borrowing from Old Norse. These two instances of substitution increased the number of concepts by which it differed from Ferring Frisian. The case of MANY demonstrates how a substitution in one variety can lead to another in a closely related variety being recorded as a change by the algorithm; this is due to the absence of a cognate for that term in the related variety's sequence.

The section of the network containing the old West Germanic varieties and Gothic diverged from the rest of the network by the concepts FAT, LEAF, MAN, MANY, ROUND, SAY, SKIN and WOMAN. Of these, MAN, MANY, SAY and SKIN had new words from the previous simulation used for them. For MAN the *g*-forms were used, for MANY the *m*-forms were replaced by *f*-forms, for SAY the *s*-forms were replaced by *kw*-forms and for SKIN the *h*-forms were replaced by *f*-forms. Old High German, Old Frisian, Old Saxon and Old English were differentiated next based on the concept HAIR, for which *h*-forms were replaced by *f*-forms. Old High German branched off next due to its word for DRY, *durri*, a cognate of the Old Norse form, and BONE, for which *bein* was replaced by *knohha*. The languages above this were shown as diverging from Old High German based on BARK and TREE. Of these the terms for TREE in these languages were substituted, with *t*-forms (such as Old English *trēow*) being used instead of the *b*-forms (such as Old English *bēam*). A *t*-form was not attested in Old High German, for which the *b*-form was used again. However, the terms for BARK were not substituted in the languages above Old High German, but in Old High German itself, where *skorza*, cognate with the Afrikaans and Standard Dutch words, was used in place of *rinta*. This again displays how substitutions in one variety may be shown as changes in another, depending on how similar to other sequences the variety in

which these substitutions took place comes to be. Replacing *w*-forms for BELLY in Old English, Old Saxon, Old Dutch and Old High German with *b*-forms (such as Old English *būc*) caused Gothic and Old Frisian to branch off from them by the concept BELLY, as they only had *w*-forms attested for this concept. Substituting the Old English and Old Saxon terms for DIE (*stēorfan* and *stervan*) with *sweltan*, a less common cognate of the Gothic term, resulted in these diverging from the posterior sequences by DIE. Old English diverged from Old Saxon by the concepts BLACK, CLOUD, GIVE, KNOW, and NECK. All of these had additional terms substituted in place of their original terms; these were *blæc*, *scio*, *sellan*, *cnāwan* and *swēora*. Gothic was still an outlier and diverged from Old Frisian by fourteen concepts. None of these involved substitutions of any kind. The unusual grouping of these two on the same branch was a product of changing the terms for BELLY in the other old West Germanic varieties to terms which were neither attested in Old Frisian and Gothic; this, combined with some of the substitutions in Old English and Old Saxon, and the increase in the number of cognates between Old Dutch, Standard Dutch and Afrikaans (again caused by substitution), resulted in these two having the greatest sequence similarity to each other in this simulation.

The next simulation to be performed was done under the Infoclean semantically strict condition, in which items for which some languages did not have attested words were removed. The following concepts were removed from this simulation: ASHES, BARK, TO BITE, BONE, CLAW, EGG, FEATHER, TO FLY, GREEN, TO KILL, LEAF, LOUSE, NOSE, ROOT, ROUND, SEED, SKIN, SMOKE, TO SWIM, TAIL, WARM and YELLOW. For the remaining concepts, the words used in simulations one and seven were used. This resulted in the following network, which had seventy-eight concepts and twenty-five sequences (see next page).



Old and Modern Germanic Languages, Infoclean, Semantically Strict Simulation, ε=0
Language n=25, Concept n=78

This resulted in a network which very clearly divided into two main sections: a North Germanic section, and a section which was made up of all of the old and modern West Germanic varieties, as well as Gothic, which emanated from this section but was shown as a very divergent outlier. These two sections diverged from each other based on the concepts WOMAN, TREE, SUN, SLEEP, NOT, FLESH, MANY, FIRE, DRY, DIE and BIG. Within the North Germanic section the first division occurred with Swedish branching off based on the concept CLOUD, for which it had *moln*, which does not have a cognate of the same meaning within the rest of the North Germanic languages. All of the languages above this division had cognates for the concept CLOUD; these were all words with the core structure /skV$_{[+high,+front]}$/, and are the descendants of Old Norse *ský*. The next division above the Swedish one was a three-way division. One of these branches separated Faroese, Icelandic and Old Norse from the rest of the North Germanic languages. All three diverged from these other sequences on the basis of the concepts EARTH and MOUNTAIN, for which they all had *mold* and *fjall* respectively. Icelandic and Old Norse diverged more than Faroese from these other North Germanic sequences, with both sharing divergences from Faroese in the concepts FLESH and BELLY. For FLESH the term Icelandic and Old Norse used was *hold*, whereas Faroese *kjøt* was a cognate of the other, more posterior sequences, which all used terms with the structure /C$_{[-vo,+obst]}$V$_{[+fro,+ro]}$C$_{[-vo,-cont]}$/, such as Swedish *kött*, [ʃøtː]. For BELLY they had *magi*, which had cognates in the two Norwegian varieties, as opposed to Faroese *búkur*. Above this branch, Icelandic and Old Norse diverged. Icelandic diverged by EAT and MOON; for EAT it had *borða*, which is the product of a semantic shift from "table" to "eat", and which is not attested in any of the other North Germanic languages; for MOON it had the word *tungl*, which originally had the more general meaning of "celestial body" in Old Norse. Old Norse branched off based on FAT, MAN and TREE. For FAT it had the word *smjǫr*, the descendants of which now generally refer more specifically to grease or butter in the modern North Germanic languages. For MAN it had *verr*, which has been replaced in all of Old Norse's descendants by words with the core structure /mVC$_{[+alv]}$/, and which originally just meant "person". In this simulation Old Norse *baðmr* was used rather than *tré*; it is this latter form which has survived into all of the modern North Germanic languages with this meaning. Danish branches off from the section shared by all of the other North Germanic varieties by the concept EAT, for which it had *spise* rather than a descendant of Old Norse *eta*. Norwegian Nynorsk branched off from the common point by the concept BELLY, for which it has *mage*, which is not cognate with the Swedish, Danish and Faroese terms (which all have the core structure /bVk/ and which are ultimately borrowings from Middle Low German). Norwegian Bokmål was shown as differing from Norwegian Nynorsk by the concept EAT, as it used the same terms as Danish, *spise*, rather than *eta*. It was also shown as differing from Danish by the concept BELLY, for which it has the term *mage*, as in Norwegian Bokmål. Norwegian Bokmål differed from each of them by one term only under the Infoclean condition, resulting in it being shown as equidistantly related to both. This may reflect the fact that it has developed with significant Danish influence, despite being a variety of Norwegian; without additional information, however, the relationship shown by the

network could equally be interpreted as that of a variety of Danish which has developed under heavy Norwegian influence, or an instance of language convergence.

Low German is shown in the network as falling between the branch which separates the North Germanic languages and an area of reticulation from which the other modern West Germanic languages emerge. Its sequence is most similar to that of Standard High German; in this simulation it is grouped more closely to the High German varieties than any other on the basis of HEAD, for which it has *Kopp*, a cognate of Standard High German *Kopf*. It differed from the sequences immediately posterior to it and making up the reticulation by the concept WALK, for which it had *gån*, whereas the posterior sequences all used a term with the core structure /lVC$_{[-vo,+lab]}$/, such as Standard High German *laufen* and Standard Dutch *lopen*. The High German varieties were differentiated from the immediately posterior sequences making up the reticulation by the concepts HEAD (which was shared with Low German) and STAR. This is due to the fact that words with the core structure /kVC$_{[-vo,+lab,+obst]}$, such as Standard High German *Kopf*, have come to be the everyday words for HEAD, while in most of the other West Germanic languages words with the structure /hV(C$_{[+lab]}$)(V)(C$_{[+obst]}$)/, such as Standard Dutch *hoofd*, are used. All of the High German varieties use words with a root-internal nasal for STAR, such as Standard High German *Stern*, while the rest of the West Germanic languages use a nasal-less form, such as modern English *star*. It is thought that these variants developed in parallel (OED, 2015; see Appendix A for details). Bavarian diverged from Standard High German based on the concept WOMAN, for which it has *Weiberz*; Standard High German does have *Weib*, a cognate of this word, but it is no longer the everyday term and has pejorative connotations (Duden, 2015). Swiss German diverged from Standard High German based on THAT and MOUTH, for which it has *säb* and *Muul* rather than *das* and *Mund*.

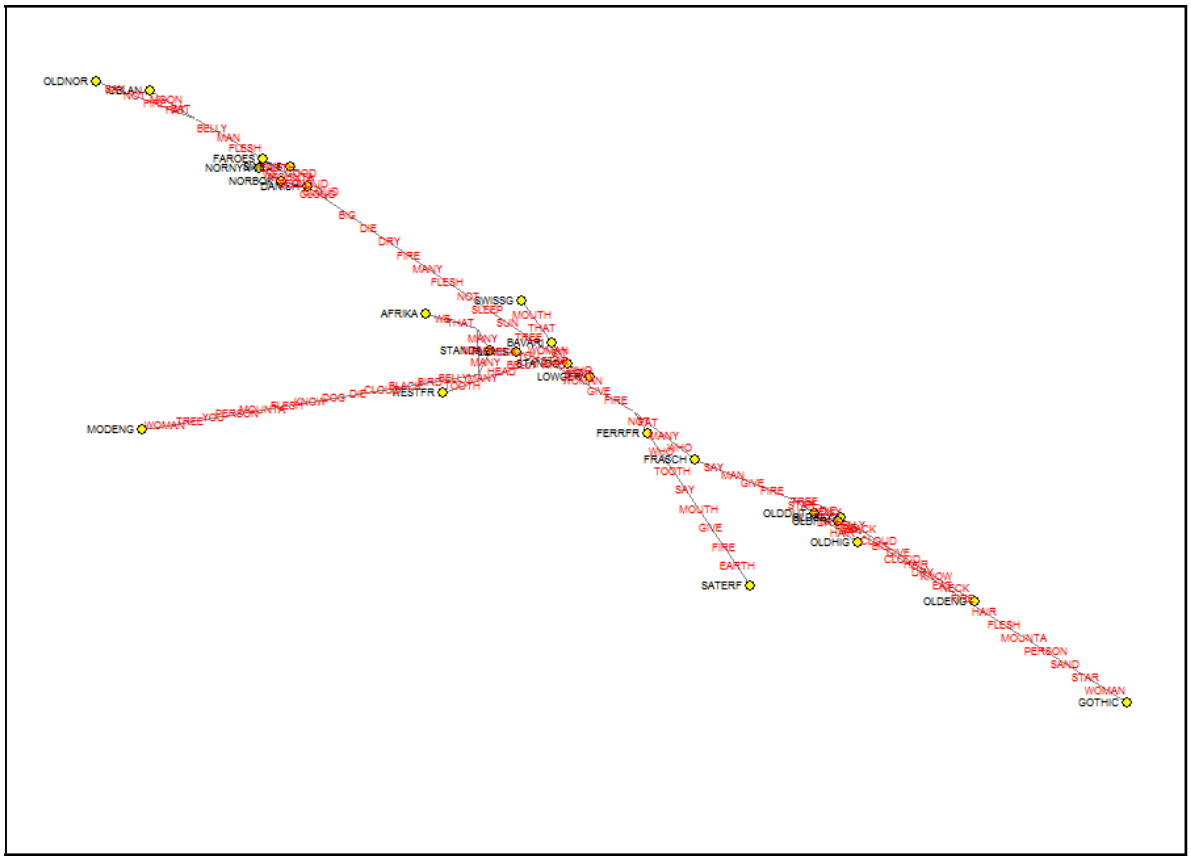
Under the Infoclean condition, Standard Dutch and Flemish were shown as one node, here marked STANDA; by removing concepts where certain languages were missing the sequences of these two became identical and were recorded by the algorithm as the same sequence occurring twice in the data. Afrikaans, Standard Dutch and Flemish branched off from the area of reticulation because of their words for YOU, *jy* and *jij*/*gij*, which are innovations. Afrikaans diverged from Standard Dutch and Flemish by HEAD, THAT, and WE. For HEAD it had *kop*, a cognate of the Low and High German words; for THAT it had the innovation *daardie*; for WE it had the semantically broadened *ons*, which was originally the first person plural accusative but which has now come to do service as a general first person plural pronoun. West Frisian differed from all of the other Germanic languages by HEAD, WALK, BELLY, MANY and TOOTH, for which it had *holle, rinne, liif, in soad* and *tosk*. *Rinne* (originally meaning "run"), *liif* (originally meaning "body") and *tosk* (originally "tusk") are all the products of semantic shifts. The etymology of *holle* is uncertain, but it has been suggested that it is related to High German *hüllen* "covering of the brain" (Hoekstra, personal correspondence). These terms have caused it to branch off on its own, although its overall sequence similarity is closest to that of Standard Dutch, rather than the other Frisian varieties.

On the basis of the concept WOMAN the old West Germanic varieties, Modern English and the modern East (Saterland) and North (Ferring and Frasch) Frisian languages branched out from the network. This was due to the fact that all of these had words descended from Proto-Germanic **wiƀan* for this concept. The first to split off from this branch was Saterland Frisian, which branched off based on EARTH, GIVE, MOUTH, NOT, SAY, TOOTH and WHO. Saterland Frisian had *fúul*, a cognate of the High German, Low German and Netherlandic words for MANY, which was why it was placed below the section of the branch which distinguishes all the sequences above it from those below by the concept MANY. All of the varieties above this section of the branch have words with the core structure /mVC$_{[+obst]}$/, such as Modern English *many*. Modern English branched out above this by the concepts BELLY, BIG, BIRD, BLACK, CLOUD, DIE, DOG, KNOW, MOUNTAIN, NECK, PERSON, YOU, TREE, and WALK. Of these, PERSON and MOUNTAIN are loans from Latin and Norman French, while BELLY, BIG, BIRD, BLACK, CLOUD, DOG, KNOW, YOU and WALK are native terms which have no cognates with these same meanings in any of the other Germanic languages. The next division separates the language sequences above it from those below it by the concept FAT; all of the sequences above this division had words with the core structure /smVr/, while those below had the structure /fVt/. Frasch and Ferring Frisian both branch off from this section based on their terms for FIRE and GIVE; for FIRE they had *iilj* and *ial*, respectively; for GIVE *düünj* and *du*. Ferring Frisian differed from the Frasch Frisian sequence by the concepts FAT, NOT, WALK, WHO. Despite a very high sequence similarity to Frasch Frisian, Ferring Frisian shared a cognate with a number of the sequences below the Frasch Frisian sequences, namely FAT, for which it had *feet*, a cognate of the other /fVt/ forms. Ferring Frisian's words for NOT and WALK also had cognates in other branches; Ferring Frisian *ei*, NOT, has a cognate in Saterland Frisian *ai*, while Ferring Frisian *gung*, WALK, has cognates in all of the North Germanic languages, and most of the Old Germanic languages (with the exceptions in this simulation being Old Dutch, Old High German and Old Frisian).

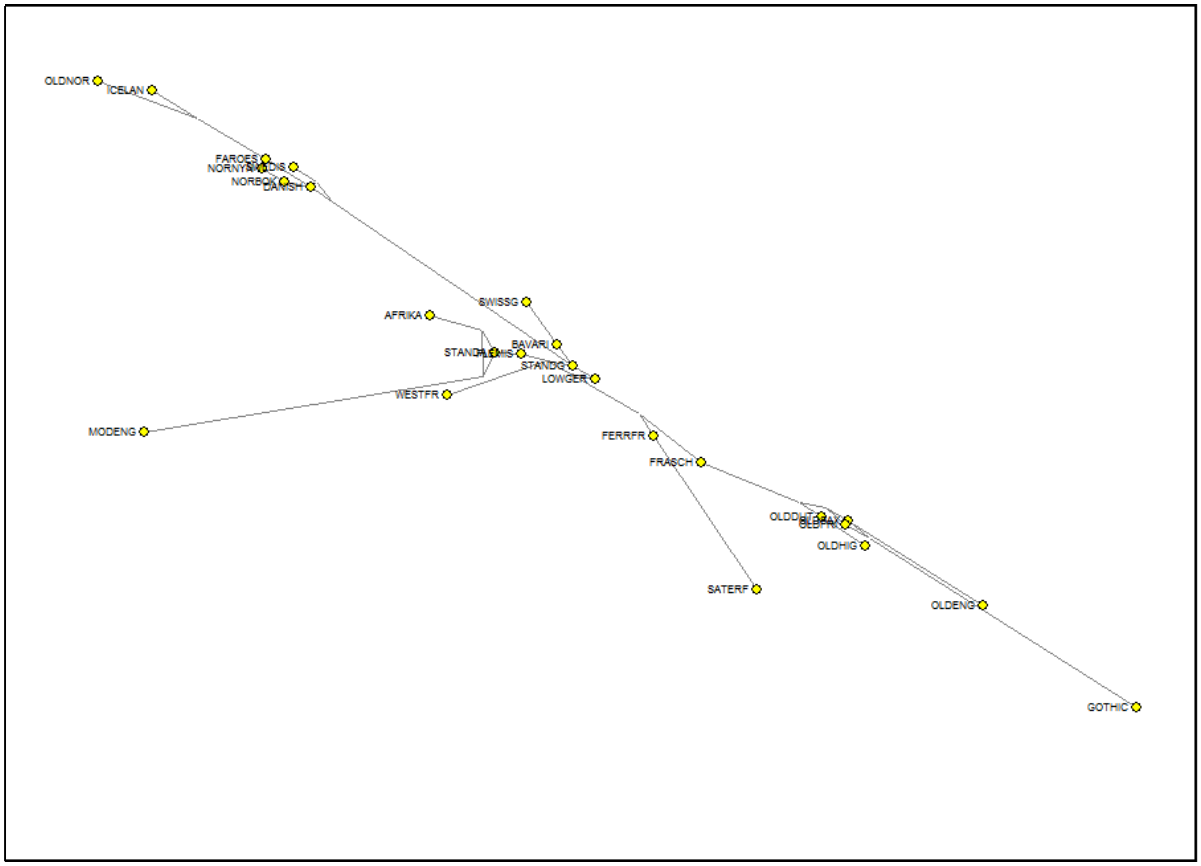
The old Germanic languages (with the exception of Old Norse), branched off based on the concepts BELLY and MAN. For BELLY they all had words with the core structure /wVm(b)/, such as Old English *wamb*, and for MAN words with the structure /C$_{[+vo,+lab,-plo]}$V$_{[-low]}$r/, such as Old English *wer*. Neither of these has survived into the modern languages with the same meanings. Old Dutch and Old Frisian were grouped together as one node (marked OLDDUT) because under the Infoclean condition all of the concepts included were ones they had cognates for; this resulted in the algorithm recording each one as the same sequence used twice. Old High German, Old English, Old Saxon and Gothic all branched out above them by the concept BIG; this was because all of these languages had words with the structure /mV$_{[+hi,+fro]}$C$_{[-vo,+obst]}$Vl/, such as Gothic *mikils* and Old High German *mihhil*, while Old Dutch and Old Frisian used *grōt* and *grāt*. Old High German had the word *loufan* for WALK; this caused it to be placed at the bottom of the branch marked WALK. Old English, Old Saxon and Gothic used *g*-forms (Gothic *gaggan*, Old English and Old Saxon *gān*/*gangan*). Old Saxon and Old English were shown as one large node (marked OLDENG); this was due to the fact that, with the words used for this simulation and the Infoclean condition which meant that a number of concepts were removed, 100% of their terms were

cognates and the algorithm therefore treated them as the same sequence that had been put in twice. Gothic was an outlier and differed significantly from the other languages in the part of the network it occupied. It differed by the concepts CLOUD, DIE, DRY, EAT, FIRE, HAIR, FLESH, MOUNTAIN, PERSON, SAND, SAY, STAR and WOMAN.

The thirteenth simulation was an Infoclean semantically strict condition which made use of words which were synonyms or near-synonyms in a number of the different languages. The simulation had the same concepts as above removed, resulting in the use of seventy-eight concepts and twenty-five sequences. The network below was generated from this data.



Old and Modern Germanic Languages, Infoclean, Semantically Strict Additional Words Simulation, ε=0 Language n=25, Concept n=78



Old and Modern Germanic Languages, Infoclean, Semantically Strict Additional Words Simulation, ε=0 Language n=25, Concept n=78

Within the North Germanic section of this network the differences were not great on comparison to those seen in the previous simulation. Swedish still branched off from a basal position in the section by the concept CLOUD; however, the length of its branch was increased by the use of *bra* for GOOD. The use of *karl* for MAN in Icelandic and Old Norse increased the length of the branch by which they diverged from Faroese by one concept (in addition to BELLY and FLESH). Faroese diverged from the rest of the North Germanic languages as in the previous simulation. As Old Norse *tré* was used instead of *baðmr*, Old Norse no longer differed from any other North Germanic language by TREE. The use of Old Norse *kveða* in place of *segja* for SAY meant that this was shown as a concept by which it diverged, as all the other North Germanic languages had a descendant of *segja* for SAY in this simulation. Additionally the substitution of *né* (which is cognate with the Gothic and old West Germanic forms) instead of *eigi* for NOT, and *funi* (cognate with Gothic *fon*) instead of *eldr* for FIRE, resulted in an increase in the length of the branch by which it diverged from the common branch it shares with Icelandic. Norwegian Nynorsk only diverged from Bokmål by the concept WE due to the replacement of *vi* with Nynorsk *me*, and the substitution of Danish and Norwegian Bokmål *spise* with *æde* and *eta*, respectively; as these are cognates of Nynorsk *eta*, they were no longer treated as concepts by which Norwegian Nynorsk diverged from Bokmål and Danish. The North Germanic languages were separated from the West

Germanic languages by the terms for CLOUD, BIG, DIE, DRY, FIRE, MANY, FLESH, NOT, SLEEP, SUN, TREE and WOMAN.

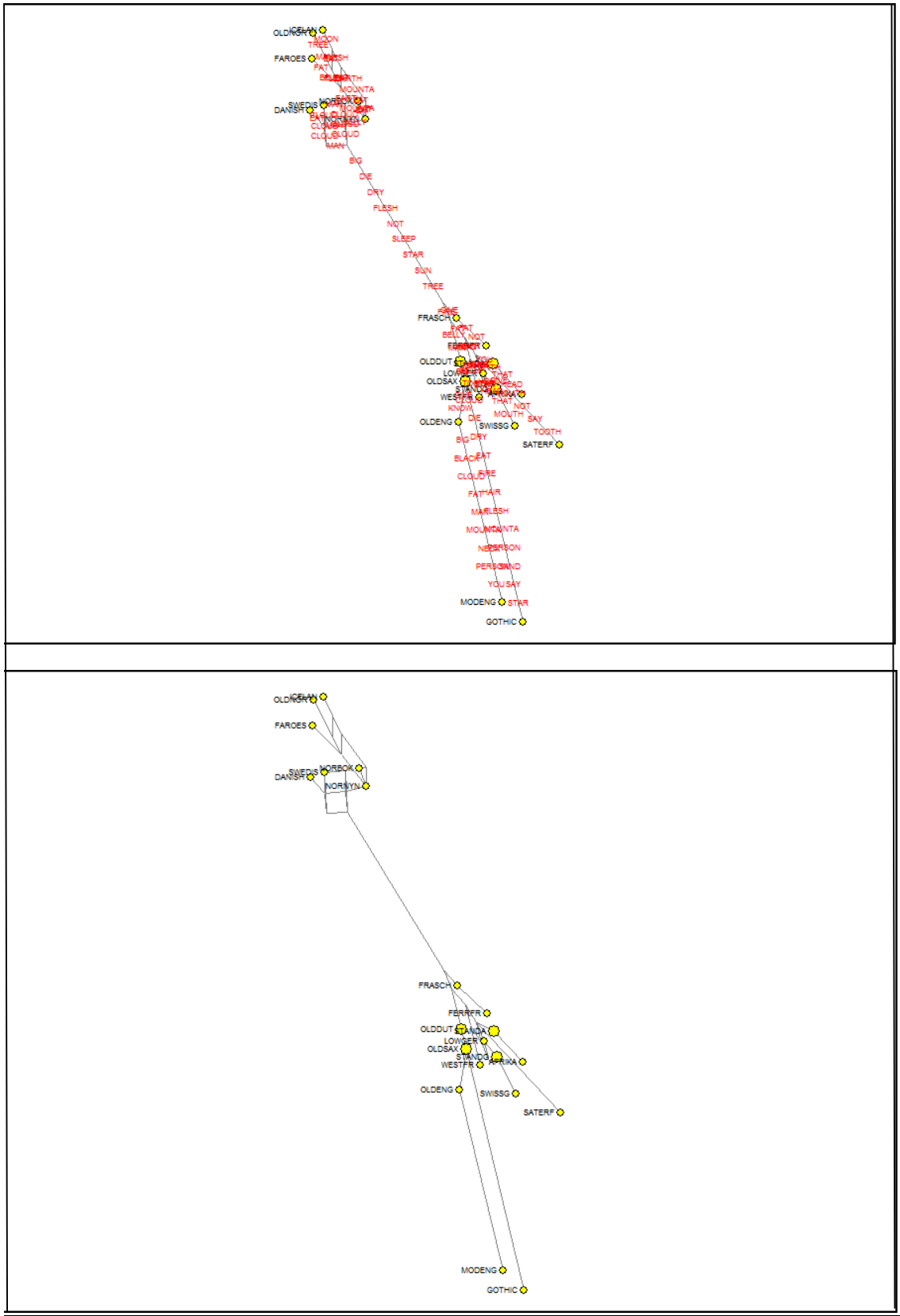
Within the West Germanic section of the network, the High German varieties differed from the other modern West Germanic varieties by the concept STAR, for which they had words with a nasal in the root. Bavarian and Swiss German differed from Standard High German by their terms for SIT, for which *hockä* and *hocke* were substituted in the respective varieties. Bavarian *Weiberz* was substituted for by *Frau*; this is the reason only SIT is recorded as a difference between Bavarian (and Swiss German) and Standard High German. Swiss German differed from Bavarian and Standard High German by THAT and MOUTH as in the previous simulation. Substituting *hals* for *nek* in both Afrikaans and Standard Dutch resulted in them diverging from Flemish by the concept NECK; it also resulted in Modern English being grouped as emanating from the same part of the network as them. All three differed from each other based on MANY; this would have occurred anyway with Modern English *many*, which does not have a surviving cognate in either of them; however, the difference in this concept between Standard Dutch and Flemish was due to the substitution of Afrikaans *veel* (cognate with Standard Dutch *veel*) with *baie*. Afrikaans further differed from Standard Dutch by WE and THAT, although these did not involve using any words not used in the previous simulation. Aside from MANY, Modern English diverged from these by BELLY, BIRD, BLACK, CLOUD, DIE, DOG, KNOW, FLESH, MOUNTAIN, PERSON, YOU, TREE and WOMAN. None of these involved substitution. West Frisian diverged basally from the Netherlandic varieties and Modern English by BELLY, HEAD, MANY and TOOTH; none of these involved substitution.

The branch on which Low German and the northern and eastern Frisian varieties are located diverges from the rest of the modern West Germanic languages by WHO; Low German occurs above this point due to the substitution of *weer* (cognate with the High German, Dutch, Flemish, West Frisian and Modern English words) with *wokeen*. This is a contraction of *welk een*, "which one", and is cognate with Saterland Frisian *hoker* (from Low Franconian *hok*, a contraction of *hwelik*; Hoekstra, personal correspondence) and Ferring Frisian *wäl* (also from *hwelik*; Hoekstra, personal correspondence). Low German branches off on its own based on STAR, for which it had *Schtirn*, which is a cognate of the High German forms based on the presence of a nasal in the root (see Appendix A). Ferring Frisian and Frasch Frisian were placed beyond Low German at the end of a branch indicating that they diverged from it by the concepts WOMAN, FIRE, and GIVE. None of these involved substitution of any terms. Ferring Frisian and Saterland Frisian branched off based on the concept NOT for which they had *ei* and *ai*, respectively; no substitution was involved here. Saterland Frisian diverged from Ferring Frisian by the concepts WHO, TOOTH, SAY, MOUTH, GIVE, FIRE and EARTH. None of these involved the substitution of terms. Frasch Frisian diverged from the common branch it shared with Ferring Frisian by FAT, MANY and WHO. None of these involved substituting terms.

The section of the network comprising the old West Germanic varieties and Gothic was shown as branching off from the Frasch Frisian node; this indicates that, of all the modern West Germanic language sequences, that of Frasch Frisian had the greatest sequence

similarity to the basal members of the old West Germanic languages (under the Infodelete "Other Words" setting). These old varieties diverged from Frasch Frisian by the concepts SAY, MAN, GIVE and FIRE. Of these, SAY and MAN had substitute terms used, namely *kw*-forms (such as Old English *cweðan*) instead of *s*-forms in the case of the former, and *g*-forms (such as Old English *guma*) in the case of the second. The use of *g*-forms for MAN did not in this instance have a different effect from the use of *wer*-forms in the previous simulation, as neither has surviving cognates with the same meaning in the modern languages. Old High German branched off from Old Dutch by the concepts HAIR and DRY; both of these involved term substitutions, with *fahs* being used in place of *hār* for HAIR (Old Dutch only had *hār*), and *durri* in place of *trokken* for DRY (Old Dutch had *drugi*). Gothic, Old Frisian, Old English and Old Saxon were in a section which branched off by TREE; this is due to the use of *t*-forms (such as Old English *trēow*, Old Saxon *trio*, Old Frisian *trê* and Gothic *triu*) in these languages; this was recorded by the algorithm as a change in the concept TREE from the posterior sequences, which all used *b*-forms (Old Dutch *bōm*, Frasch Frisian *buum*). Old English and Old Saxon branched off by the concept DIE; this was due to the substitution of Old English *stēorfan* and Old Saxon *stervan*, which had cognates in all of the other old West Germanic languages, with *sweltan*, which only has a cognate in Gothic. Despite this, their sequences were still more similar to the other old West Germanic varieties. Old English diverged from Old Saxon by the concepts BLACK, CLOUD, GIVE, HAIR, KNOW and NECK. Of these, BLACK, CLOUD, GIVE, KNOW, and NECK involved term substitutions. Old Frisian and Gothic were both consigned to their own branch because of the difference between their terms for BELLY and those of the other languages; the w-forms of the previous simulation (such as Old English *wamb*) were replaced by *b*-forms (such as Old Saxon *būk* and Old English *būc*) in all of the old West Germanic varieties except for Frisian (for which only *wamme* is attested); likewise Gothic, which is not West Germanic, only has *wamba* for BELLY. Gothic diverged significantly from the rest of the varieties in this section, and branched off from Old Frisian (with which it had the highest sequence similarity out of all the other sequences used) by the concepts DIE, BIG, CLOUD, DRY, EAT, FIRE, HAIR, FLESH, MOUNTAIN, PERSON, SAND, STAR and WOMAN. None of these involved term substitution.

The following simulation was an Infoclean semantically lax simulation. In this simulation a selection of terms which exist in the modern Germanic languages but have undergone semantic shifts were used in place of terms with the more strict meanings in the above simulations. This network was generated from a data-set of seventy-eight concepts and twenty-five languages (see next page).



Old and Modern Germanic Languages, Infodelete, Semantically Lax Simulation, ε=0
Language n=25, Concept n=78

The network still had two distinct groupings of languages, that of the North Germanic languages and that comprising both the old and modern West Germanic languages. Gothic branched out of the area occupied by the West Germanic languages, although it differed from them considerably. Modern English also differed considerably from them.

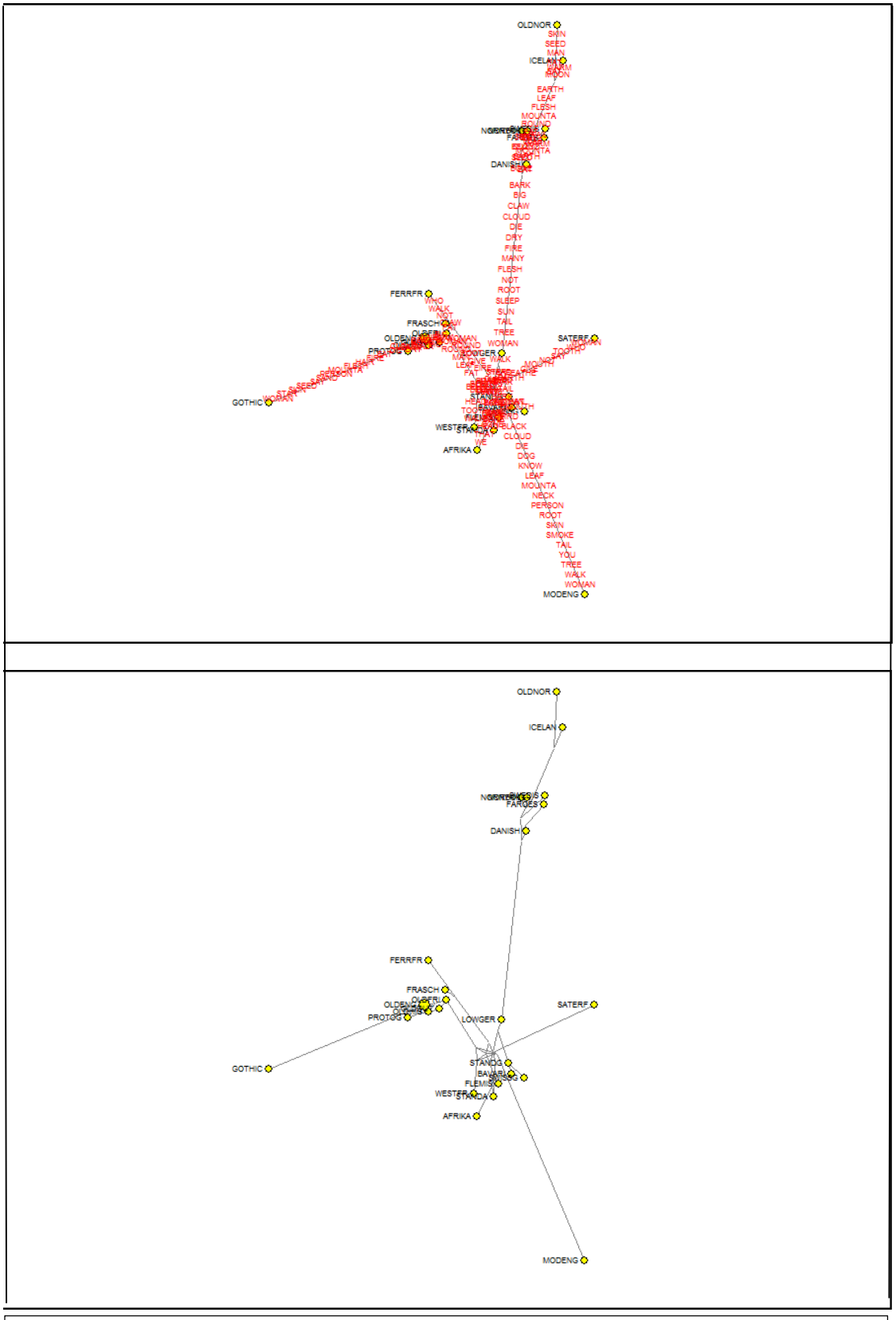
The substituting of Modern English *womb* in place of *belly* for the concept BELLY (which was assigned the same code as the words with the core structure /wVm($C_{[+vo,+lab]}$)/, such as Old English *wamb*), *fowel* in place of *bird* for BIRD (resulting in this being coded as a cognate of Old English *fugol*), *starve* (descended from Old English *stēorfan*) in place of (*to*) *die* for DIE, *hound* (a cognate of Old English *hund*) in place of *dog* for DOG, (*to*) *sell* in place of (*to*) *give* for GIVE (for which Old English *sellan* was also used), *can* in place of (*to*) *know* for KNOW (for which Old English *cunnan*/*cannan* was used), and *beam* (descended from Old English *bēam*) in place of *tree* for TREE appears to have changed the way Modern English was grouped. Modern English is shown in this simulation as descending directly from Old English. Despite differing from Old English by the concepts BIG, BLACK, CLOUD, FAT, MAN, MOUNTAIN, NECK, PERSON and YOU, the use of terms which have undergone semantic shifts from Old English to Modern English has resulted in an increase in the overall sequence similarity between them such that Old English is now shown as the Germanic language to which Modern English has the greatest similarity.The use of Standard Dutch *wijf* (cognate with Old Dutch *wīf*) in place of *vrouw* for WOMAN has not caused Standard Dutch to be more closely linked to Old Dutch. The same is true of Standard High German, in which *Weib* (cognate with Old High German *wīb*/*wīp*) was used in place of *Frau* for WOMAN.

The substitution of the modern North Germanic descendants of the Old Norse *smjǫr* (Danish *smør*, Norwegian *smør*, Icelandic *smjer* and Swedish *smör*) for their terms for FAT in the previous simulations, had no major effect on the grouping of the North Germanic languages other than to cause Faroese to branch off from the branch connecting it to Icelandic and Old Norse by one extra concept, FAT (for which it had *feitt*). By substituting *karl* for *man* and *mand* in Swedish and Danish, they were shown as branching off together based on the concept MAN. The use of *karl* in place of *verr* for MAN in Old Norse did not cause it to be grouped more closely to Danish and Swedish as this change was not sufficient to increase the overall sequence similarities between them for this to happen.

Subgroup 4: Old and Modern Germanic Languages with Proto-Germanic

These simulations used the data from the semantically strict simulations of the old and modern Germanic languages, as well as reconstructed Proto-Germanic items. The reconstructed Proto-Germanic items were added to the pre-existing datasets of simulations ten and twelve. As there were many concepts for which at least two possible ancestral items could be reconstructed, two word choices for each simulation were possible. There were thus two simulations under the Majority Wins condition, and two under the Infoclean condition.

These were referred to as Word Choice One and Word Choice Two. The following network was generated for the "Majority Wins, Word Choice One" simulation, which had twenty-six language sequences and ninety-nine concepts.

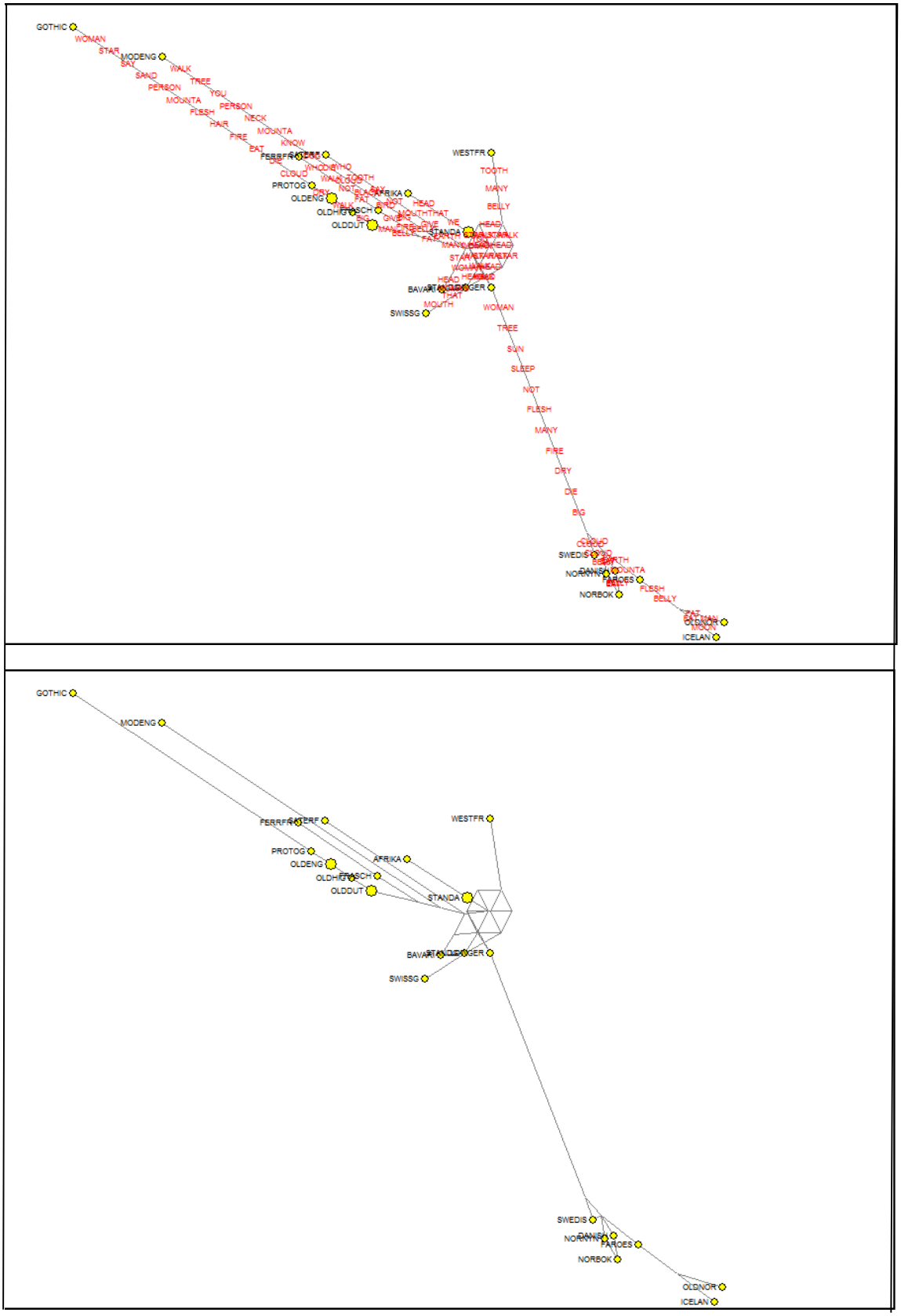


Old and Modern Germanic Languages, with Proto-Germanic, Word Choice One, Majority Wins, Semantically Strict Simulation, ε=0, Language n=26, Concept n=99

The attested varieties of Germanic were placed in the same places in the network as they were in simulations ten and eleven; the inclusion of reconstructed Proto-Germanic items did not significantly alter how they were organised relative to each other.

Proto-Germanic was placed in the section of the network in which all of the old Germanic languages except for Old Norse were placed. This section branched off from the section of the network in which most of the modern West Germanic languages were placed by the concepts BELLY, FAT, LEAF, MAN, ROUND and WOMAN. Old English and Old Saxon were grouped together as one large node marked OLDENG; this was because under this condition they had cognates for all of the concepts used, and were thus treated by the algorithm as the same sequence used twice. The Proto-Germanic node was placed at the end of a section of reticulations; the sequences marked as the most closesly related to it were Old English (and Old Saxon) and Old High German. The Proto-Germanic sequence differed from those of Old Saxon and Old English by the concepts BARK and DRY; this was because for BARK Proto-Germanic **skurtaz* was used (where Old Saxon and Old English had *rinda* and *rind*) and for DRY **þurzuz* (cognate with Gothic *þaursus* and Old Norse *þurr*) was the only reconstructed form supplied by Orel (2003); for DRY Old Saxon and Old English used *drokno* and *dryge*. The Proto-Germanic sequence also differed from the Old High German sequence by two concepts: DRY and WALK. DRY was marked by the algorithm as a difference because Old High German *trokken* (cognate with the Old English and Old Saxon forms) was used in this simulation, and not *durri*, which would have been marked as a cognate of Proto-Germanic **þurzuz*. For WALK, Proto-Germanic had **gēnan*, which was a cognate of Old English and Old Saxon *gān*, whereas Old High German had *loufan*; in Proto-Germanic the *l*-form meant "to leap", and it was only later on that a semantic shift caused this to come to mean "to walk" in Old High German, Old Frisian and Old Dutch. Gothic still differed significantly from the other Old Germanic languages.

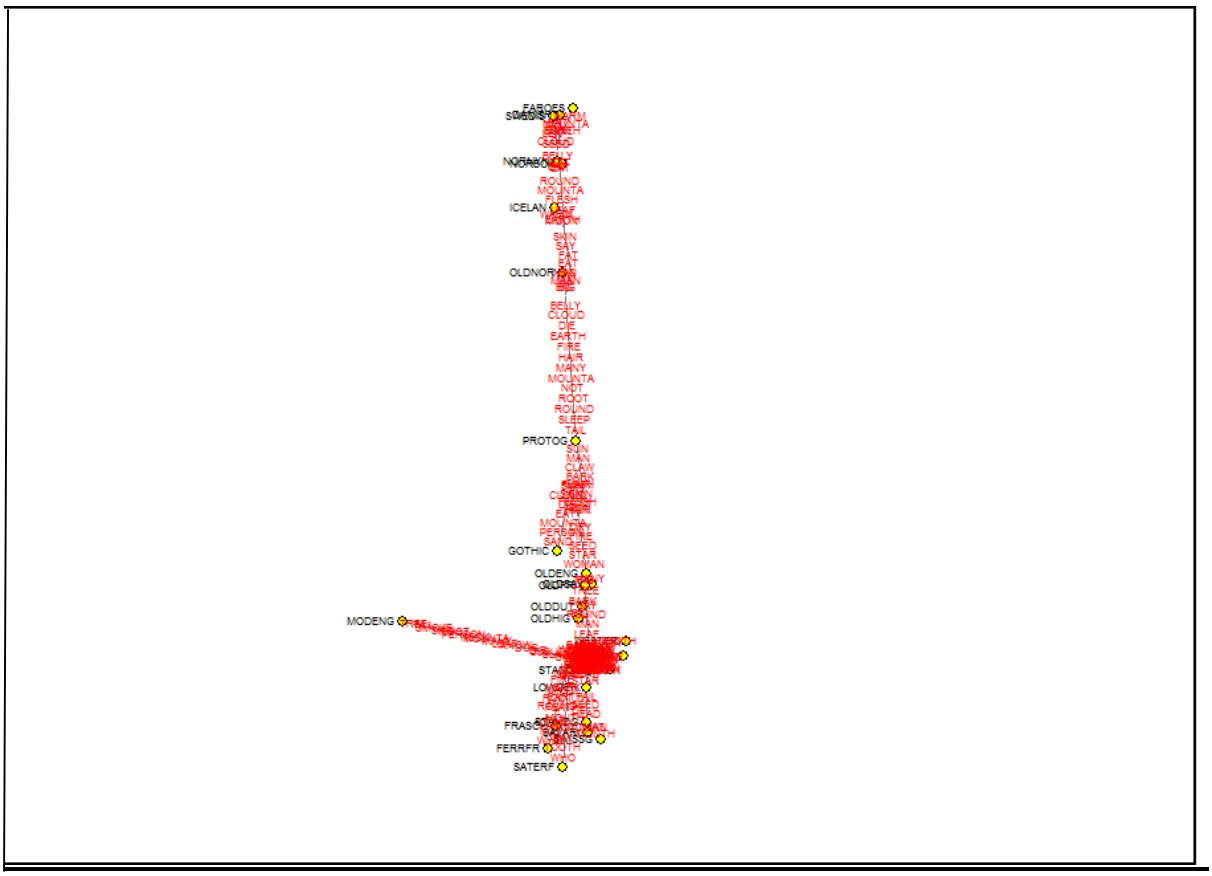
The next simulation used the same words as those above, but was performed under the Infoclean condition. This simulation was based on twenty-six language sequences and seventy-eight concepts. This generated the following network (see next page).



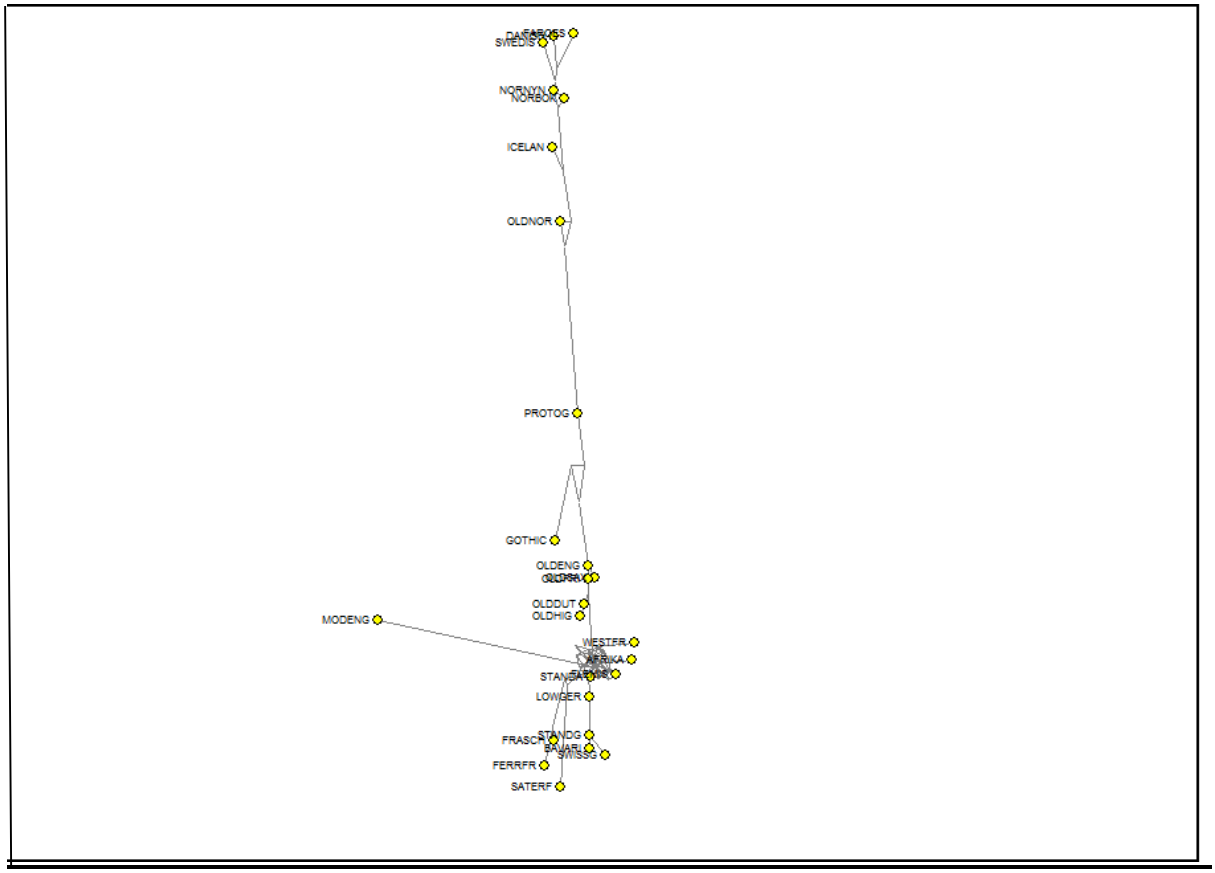
Old and Modern Germanic Languages with Proto-Germanic, Word Choice One, Infoclean, Semantically Strict Simulation, ε=0, Language n=26, Concept n=78

This network was not different from simulation 13 except for the fact that it has an additional node in it, that of Proto-Germanic. Proto-Germanic was placed in the section of the network in which sat the other old West Germanic varieties and Gothic. The removal of the concept BARK resulted in Proto-Germanic differing from Old English and Old Saxon (which were indicated by the large node marked OLDENG) by the concept DRY only; this was due to the fact that the reconstructed form in Proto-Germanic is **þurzuz* (cognate with Gothic *þaursus* and Old Norse *þurr*). The removal of BARK also had the effect of reducing the number of cognates between the Old High German and Proto-Germanic sequences in this simulation. This resulted in Old High German being placed below Old English and Old Saxon (the OLDENG) node on the branch, as it differed from the Proto-Germanic sequence not only by DRY, but also by WALK. Gothic differed significantly from the Proto-Germanic sequence, and was still grouped as an outlier of this section of the network.

The following simulation was a semantically strict Majority Wins simulation in which the concepts for which there were two reconstructed Proto-Germanic words were changed to reflect the second set of possible forms. This was called Word Choice Two. This had ninety-eight concepts and twenty-six language sequences. This setting generated the network below.



Old and Modern Germanic Languages with Proto-Germanic, Word Choice Two, Majority Wins, Semantically Strict Simulation, ε=0, Language n=26, Concept n=99

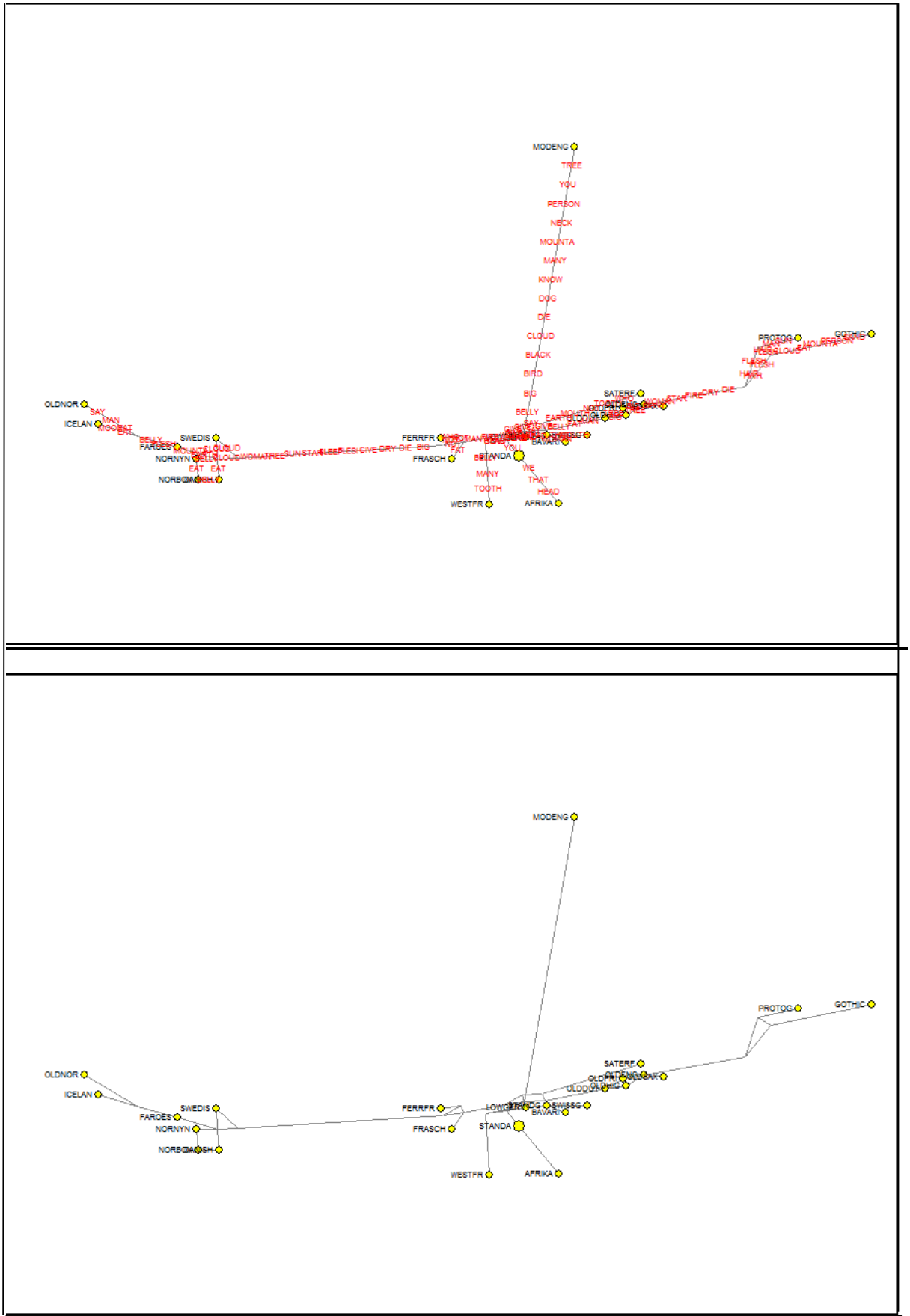


Old and Modern Germanic Languages with Proto-Germanic, Word Choice Two, Majority Wins, Semantically Strict Simulation, ε=0, Language *n*=26, Concept *n*=99

The use of different words for a number of the concepts had a significant impact on how Proto-Germanic was placed in the network. Proto-Germanic was located further away from Old English than it had been in the previous Majority Wins simulation; in this simulation its node sits in between Old Norse and the other old Germanic varieties. Old Norse and Proto-Germanic differed by the concepts BELLY, BIG, CLOUD, DIE, EARTH, FIRE, MANY, MOUNTAIN, NOT, ROOT, ROUND, SAY, SLEEP, TAIL, FAT and MAN. Of these, FIRE, MANY, FAT,MAN and SAY involved changing terms. For FIRE, **fon* (coded as a cognate of Gothic *fon*) was used instead of **fuwer~*fur~*fûir*; Old Norse *eldr* is a cognate of neither of these, and in this instance either form would have resulted in a difference between the Proto-Germanic and Old Norse sequences. For SAY, Proto-Germanic **kweþanan* was used instead of **sagjanan~*sagjan~*sagǣjan* (where Old Norse had *segja*). For MANY, Proto-Germanic **felu* was used instead of **managaz* (whereas Old Norse *margir* was kept). For FAT (**faito-*) and MAN (**gumon*), words which were not used in any of the old Germanic languages were used in Proto-Germanic for this simulation; this resulted in Proto-Germanic branching off by itself. As descendants of these forms are attested in a number of the descendent languages with the same meanings, this branching off represents an exaggeration of the probable real-life distances between them. This could be viewed as a potential problem which could result from the use of non-equivalent items; this will be discussed below.

The old West Germanic languages and Gothic differed from Proto-Germanic by the concepts CLAW, SUN, BARK, SKIN, and FLESH. All of these involved changes in the terms used in the last simulation: for CLAW Proto-Germanic **klôh-* was used instead of **klawa-~*klæwâ-*, for SUN a hypothetical form with an *l*-formant was used instead of **sunnon*, for BARK **barkuz* was used in place of **skurtaz*, for SKIN **fello* was used instead of **hūðiz~*hudiz*, and for FLESH **huldan* was used instead of **flaiskaz*. In the case of FLESH Gothic had the word *leik*, which Orel (2003) has not reconstructed for Proto-Germanic; this could thus be seen as an innovation unique to Gothic. Gothic branched off on its own by the concepts CLOUD, DIE, EAT, HAIR, MOUNTAIN, PERSON and SAND. The old West Germanic languages further differed from Proto-Germanic by the concepts DRY, FIRE, SEED, STAR and WOMAN. Of these, FIRE, SAY, SEED, STAR and WOMAN involved substitution of terms in Proto-Germanic. These were: **fon* for FIRE, **kweþanan* for SAY, **fraiwan* for SEED, **sternon* for STAR, and **kwenan* for WOMAN. Old English, Old Saxon and Old Frisian all had words with the core structure /rV$_{[+hi,-ro]}$nC$_{[+alv,+plo]}$/ (such as Old English *rind*) for BARK. This could be viewed as a West Germanic innovation (a word with this structure is also found in Old High German, although this was not used in this simulation). This caused them to branch off from the longer branches differentiating them from Proto-Germanic and North Germanic. Old High German and Old Dutch branched off together by the cognates for TREE they had, *boum* and *bōm*, respectively. These were not only not cognates of the forms used in the posterior old Germanic languages (all of the other old West Germanic varieties, Gothic and Old Norse), which all used *t*-forms (such as Old English *trēow*) but also Proto-Germanic, in which **trewan* was substituted for **boumaz*. Of the terms differentiating the old West Germanic varieties from Proto-Germanic, DRY, SEED, STAR and WOMAN had cognates between Proto-Germanic and Old Norse, while DRY, FIRE, SEED, STAR and WOMAN had cognates between Proto-Germanic and Gothic. It is certain that the increase in cognates between Proto-Germanic and Old Norse was sufficient for Old Norse to be grouped more closely to Proto-Germanic than before; this had the effect of causing the North Germanic section of the network to be reorientated so that Old Norse was closer to Proto-Germanic. As the modern North Germanic varieties share more cognates with Old Norse than any other Germanic variety (barring each other), they were not assigned to different parts of the network. Due to the fact that Icelandic has the most in common with Old Norse, this resulted in the inner arrangement of the North Germanic languages in being effectively reversed from that of the previous simulation. The greater sequence similarity between Proto-Germanic and Gothic compared to that between Gothic and the old West Germanic varieties also meant that Gothic was placed so as to be closer to Proto-Germanic; the result of this was that it branched off basally from the section in which the old West Germanic varieties were situated. As the overall sequence similarity between the old West Germanic languages and Proto-Germanic was still slightly higher than that between Proto-Germanic and Old Norse, Proto-Germanic was positioned slightly more closely to the basal members of the old West Germanic section.

The last simulation was performed under the semantically strict Infoclean condition, with the same words as those used above (with the exception of ones falling under removed concepts). This simulation had twenty-six language sequences and seventy-eight concepts.



Old and Modern Germanic Languages with Proto-Germanic, Word Choice Two, Infoclean, Semantically Strict Simulation, ε=0, Language n=26, Concept n=78

The removal of a number of concepts had a dramatic effect on the placing of Old Norse in relation to Proto-Germanic. Proto-Germanic was still located in a section of the network in which Gothic and the old West Germanic varieties were the closest old Germanic varieties to it. Gothic branched off from both Proto-Germanic and the old West Germanic varieties by the concepts HAIR, FLESH, CLOUD, EAT, MOUNTAIN, PERSON and SAND. Proto-Germanic and Gothic had cognates for WOMAN, STAR, FIRE, DRY and DIE; both differed from the old West Germanic varieties by these concepts. Old Norse was located much further away in the network; this was due to the removal of the concepts ASHES, BARK, TO BITE, BONE, CLAW, EGG, FEATHER, TO FLY, GREEN, LEAF, LIVER, LOUSE, NOSE, ROOT, ROUND, SEED, SKIN, SMOKE, TO SWIM, TAIL, WARM and YELLOW. Of these concepts, BARK, CLAW, and SEED underwent substitution which resulted in cognates between Old Norse and Proto-Germanic for these three concepts. This suggests that the changing of these three characters in the previous simulation was enough to alter the Hamming distance between Old Norse and Proto-Germanic such that their sequence similarities were great enough to result in the repositioning of the entire North Germanic section so that Old Norse could be placed more closely to Proto-Germanic. The use of **gumon* for MAN and **faxsan* for HAIR, when the descendant forms of both were not used in any of the other languages for this simulation resulted in the algorithm treating these as innovations unique to Proto-Germanic; this is the same as the non-equivalence identified above; this will be discussed later on. The repositioning of Old Norse resulted in SUN (for which Proto-Germanic had a hypothetical cognate of Old Norse *sol*) and FLESH (for which Proto-Germanic had **huldan*, cognate with Old Norse *hold*) being marked on the Proto-Germanic branch as differences causing it to branch out on its own. Proto-Germanic did indeed differ from the Gothic and old West Germanic sequences by these characters; however, the repositioning of Old Norse has created an arrangement which makes it harder to see that these are shared with Old Norse, which could increase the likelihood that these characters could be misinterpreted as innovations unique to Proto-Germanic. This problem necessitates careful examination of the data when trying to look at why nodes were placed the way they were.

Discussion

The simulations using only lexical data from the old Germanic languages consistently resulted in a tripartite division in which Gothic and Old Norse each occupied their own distinct branches of the network, and the old West Germanic languages formed a cluster. This held across all of the simulation conditions, suggesting that the differences in basic lexicon between the generally recognised divisions of East, West and North Germanic are significant enough for this tripartite division to remain stable even when words which are not attested across all of the languages or are not the only available terms for a concept are taken into account. The pattern created through the use of lexical data differed in some ways from classifications of the Germanic languages which have been based on phonological data. The main difference in classification is that the lexical data leads to a very distinct three-way division between North, West and East Germanic, without any apparent subgrouping of them

at a higher level. This is in marked contrast to the classification of Bahnick(1973; as cited in Nielsen, 1989), which used phonological and morphological data and which resulted in a grouping of the North and West Germanic languages together as members of a North-West Germanic subdivision of the Germanic language family. In this classification, Gothic is an outlier, with the rest of the attested Germanic languages shown as diachronically closer to each other. This division has come to be seen by Salmons (2012) as more plausible than a tripartite division with no subdivisions.

At a lower level there were significant differences within the West Germanic group. Most currently accepted classifications of the West Germanic languages which are based on sound-correspondences place Old High German at a basal position within this branch due to the High German Consonant Shift (Fortson, 2004), while the North Sea varieties (Old English, Old Frisian, Old Dutch and Old Saxon) form their own subgroup. Within this subgroup Old English and Old Frisian are frequently grouped together as constituents of an Anglo-Frisian subgroup (Barber, Beal & Shaw, 2012; Lass, 1997). In marked contrast to this, the use of lexical data resulted in Old English and Old Frisian being grouped quite far away from each other, with Old English generally occupying a branch of its own and Old Frisian being grouped with Old Dutch. Old High German was generally grouped as being closer to Old Dutch and Old Frisian than either of these was to Old English or Old Saxon; thus from the point of view of this lexical data there was neither a distinct Anglo-Frisian subgroup nor a distinct division between North Sea Germanic and Old High German. The changing of lexical items and the changing of strategies for the handling of missing data did not significantly alter this picture. Changing lexical items or simulation conditions and strategies tended to simply alter the distance by which varieties differed from each other; this resulted in some cases in two distinct varieties being treated as the same sequence used twice (for example, Old Dutch and Old Frisian in simulation 3 of simulation subgroup1, and Old English and Old Saxon in simulation 1 of subgroup 3), while in others they were grouped as separate, but closely related, varieties (such as Old Dutch and Old Frisian in simulation 1 of subgroup 1).

This shows one problem with the lexical approach: high levels of lexical similarity between two varieties may increase the likelihood that errors in the coding or choice of single words could result in the algorithm treating them differently. Even though the varieties may be grouped as similar, the difference between being shown as the same, and similar but different is a considerable one when attempting to construct a phylogeny. The impression given by this set of simulations is that of a division in the Old West Germanic languages between what could be thought of as a northern group (comprising Old Saxon and Old-English) and a more southern group (comprising Old High German, Old Dutch and Old Frisian). This division could indicate that Old Frisian was already coming under enough Old Dutch influence to result in the use of a number of terms common to both of them as opposed to terms which were in wider usage in Old English and Old Saxon. Considering that it is generally believed that significant Dutch influence was exerted on Frisian at a later period (Hoekstra & Tiersma, 1994) it is possible that Dutch influence began at an earlier period than was previously thought; as most of the basic vocabulary items shared by Old Dutch and Old Frisian appear to be continuations in both languages of Proto-West Germanic terms, it is possible that this

influence came down not so much to borrowing from one into the other, but to one influencing the diction and lexical selection of the other. This requires more detailed investigation. The fact that Old Norse is placed far away from the old West Germanic languages when lexical data is used indicates a high degree of divergence between Old Norse and the old West Germanic languages in terms of basic vocabulary. When it is borne in mind that classifications based on morphological and phonological data group these two together at a higher level, the interpretation that for some reason dramatic changes in word selection or possibly semantics took place not long after they diverged appears to offer itself. One way in which such a scenario could have played out is that different dialectal terms came to be selected in different regional varieties of a common North-West Germanic language, while the sound systems of these varieties remained more similar to each other than to East Germanic. This similarity may have been maintained because the tendency to pronounce certain phonemes in certain ways was already established across this north-western variety, and this may have played out in a tendency for certain sounds to change in largely the same way even in dialects of North-West Germanic which were geographically separated from each other (bringing to mind Schützeichel's (1976) concept of *Anlagetendenz* (Nielsen, 1989)). For instance, the tendancy to raise and front vowels in the syllable preceding another one with a high vowel or glide may have been sufficiently widespread so that it was able to carry on developing into *i-umlaut* even when geographical space was allowing for the selection of different lexical items in different dialects without their spread through the language area. This is something which requires further investigation.

The modern Germanic languages were divided up into two distinct groups, the North and West Germanic languages. In most of these simulations Low German was situated basally to the rest of the West Germanic languages. This indicates that, compared to the other West Germanic languages, Low German has the smallest number of differences from the North Germanic languages, although the actual number is still large enough for it to be unambiguously a West Germanic language. This suggests that a number of innovations in basic vocabulary which have occurred in the histories of the other West Germanic languages have not been shared by Low German. The High German varieties tended to branch off basally from the other modern West Germanic varieties, somewhat in line with the results of phonological classifications which take the High German Consonant Shift into account. In phonologically-based classification, however, Low German is generally grouped with Dutch, Frisian and English, rather than High German, which is the most basal branch of the West Germanic group in these classifications (Fortson, 2004; van der Wal & Quak, 1994). This indicates that the High German varieties share a number of lexical innovations with the other West Germanic varieties which are not shared by Low German. Modern English was grouped in all but one simulation as being most similar lexically to the North and East Frisian varieties (Ferring and Frasch Frisian in the case of the former, Saterland Frisian in the case of the latter). This appears to reflect an Anglo-Frisian subgrouping, although this subgrouping was not seen in the old Germanic language simulations. Additionally, West Frisian was grouped in most instances as being more closely related to the Netherlandic varieties of West Germanic (Standard Dutch, Flemish and Afrikaans). This could be interpreted as indicating that the North and East Frisian varieties are descended from dialects of Old Frisian which are

different from that used in the old Germanic simulations, and that they have not experienced much noticeable Dutch influence on their basic vocabularies. The fact that they are not grouped as closesly related to Danish, Low German or High German, which are the languages in the closest geographical proximity to these varieties, further suggests that the influence of these languages on the Frisian varieties' basic vocabularies has not been great. The placing of West Frisian in most of the simulations as being most closely related to the Netherlandic varieties suggests that the similarities in terms of basic vocabulary between these languages are significant; this could be due to the high levels of contact and exposure between these languages over several centuries (Hoekstra & Tiersma, 1994). However, in light of the placing of Old Frisian as closely related to Old Dutch in the old Germanic simulations, it is possible that this influence is older than previously thought, and that many of these shared terms have been inherited from Old Frisian by West Frisian.

Despite this, West Frisian was still fairly divergent, due to the presence of a number of terms which differed from the Netherlandic varieties in terms of semantics, such as *tosk* for TOOTH and *liif* for BELLY, as well as terms unique to it, such as *in soad* for MANY. Even though it was placed as most closely related to the North and East Frisian varieties, Modern English still differed significantly from all of the other West Germanic languages. This can be attributed to two things in these simulations: the presence of a higher number of loanwords in the basic vocabulary list which are not shared by the other West Germanic languages, and the presence of a number of terms which are innovations unique to English. The first group includes Latin, Norman French and Old Norse loans, such as *person*, *mountain*, *sky* and *leg*. The second includes either words which are only attested in English and which have an unknown etymology (such as *kill*), words which are the products of semantic shifts which did not occur in other Germanic languages (such as *cloud*), or a combination of the two (such as *dog* and *bird*). Afrikaans, Flemish and Standard Dutch were consistently grouped together, due to their high levels of lexical similarity. Where there were differences in how they were grouped these were due to changes in word choice. For instance, where Afrikaans and Flemish *nek* were chosen for NECK, they were grouped more closely to each other, but when Afrikaans *hals* was used, it was placed more closely to Standard Dutch.

This indicates that even small differences in word choice could have significant effects on the placing of nodes in the network. This problem would be likely to be intensified with smaller datasets, as fewer changes would be required to alter the overall similarity of two sequences. This is a potential danger when working with meagrely attested languages and language varieties. This changing of position was even more dramatic in the second simulation, in which the use of Netherlandic *nek*, a cognate of Modern English *neck*, resulted in Modern English being grouped with Standard Dutch, Flemish and Afrikaans, rather than the North and East Frisian varieties. This is quite a dramatic regrouping, and suggests that the placement of sequences may be very sensitive to the coding of single characters even in fairly large datasets. This is likely to be especially the case where sequence similarity is high, as was the case across the West Germanic varieties. This indicates that the use of lexical cognates with a program such as Network for the purposes of constructing language

phylogenies may be very open to the influence of even small errors in coding, and should be borne in mind when attempting this.

The placing of the modern North Germanic varieties was interesting in that it did not reflect the traditional West-East division of many classifications based on observed sound changes, instead placing Swedish and Danish (traditionally constituting the Eastern group) far away from each other, with Danish placed basally to the rest of the North Germanic languages, and Swedish sitting as an outlier. The reason for the basal placing of Danish is the number of basic vocabulary items which are of Middle Low German origin; these resulted in the algorithm attempting to place it more closely to Low German than any of the other North Germanic varieties. Faroese and the two official varieties of Norwegian (Nynorsk and Bokmål) were grouped as being fairly closely related to each other and to Danish. This was because they have been subjected to strong Danish influence over the course of several centuries (Askedal, 1994), with a number of Danish loans entering their basic vocabularies as well, or their choice of words for particular concepts converging with the Danish ones under this influence. Icelandic and Swedish were generally placed as outliers. This is in stark contrast to the traditional grouping based on diachronic phonological data, which groups Icelandic, Faroese and Norwegian together in a Western subgroup, owing to their descent from Old West Norse dialects, and Danish and Swedish in an Eastern subgroup, owing to their descent from Old East Norse dialects (Henriksen & van der Auwera, 1994; Faarlund, 1994; Fortson, 2004).

The simulations using data from both the old and modern Germanic languages provide a picture which is slightly more complex than the traditional three way split into East, West and North Germanic subgroups. As in the previous simulations, the North Germanic languages/varieties formed their own distinct group, with Danish, both varities of Norwegian and Swedish all grouped relatively closely to each other. Icelandic and Old Norse were outliers in this group, and appeared to differ from the other North Germanic languages by roughly the same number of lexical items. This is not surprising, as a number of Old Norse basic vocabulary items have not survived with the same meanings in the modern continental North Germanic languages or Faroese, but have survived in Icelandic. Icelandic has nevertheless undergone some innovations in its basic lexicon, resulting in its either branching off from a common branch with Old Norse (simulation 2 of subgroup 3) or occupying an entire branch of its own which emanates from a section of the network placed in such a way as to suggest Old Norse is a more distant relative (simulation 1 of subgroup 3). This placing was dictated by the choice of lexical items: in simulation 2, a larger number of cognates in Icelandic and Old Norse existed in the data, while in simulation 1 this number was lower, with Old Norse having more cognates overall with Faroese. A number of these words have or had meanings which are not easily distinguishable. For example, Old Norse *sæði/sáð* and *frjó* both mean “seed”, and it is difficult to judge which one had the more neutral meaning; the first is continued in Faroese *sáð* while the second is continued in Icelandic *fræ*, with each respective term being the most neutral and common term for SEED in its respective language.

This highlights a major potential difficulty with the use of lexical data, as instances such as this create difficulties with deciding on reliable classifications and groupings, and may not be easily solved without recourse to other data. One possible way to get around this would be to first do a frequency analysis to determine which term was the more common one in the ancestor language; this would be very open to being influenced by the quality and type of extant sources, however, and would still not be able to account for the fact that even if one form was more common than the other, the two still existed side-by-side with similar meanings. With the exception of Old Norse, the other old Germanic languages consistently occupied a branch which was shown as being part of a subgroup which also included the modern West Germanic languages; this could largely be due to the fact that the largest number of old Germanic varieties were West Germanic, and that as a result their sequence similarities were close enough to those of the modern West Germanic languages to result in their being placed within the group. As all of the older West Germanic varieties were more lexically similar to each other than to their descendants, they all occupied a single branch, while their descendants occupied multiple branches within this section. As Gothic repeatedly had higher sequence similarity with the Old West Germanic varieties than with any other modern or old varieties, despite being highly divergent, it was consistently assigned to this section, although its node lay a considerable distance from the others. The placing of the modern West Germanic varieties in relation to their ancestor varieties, with no modern varieties shown as arising directly from their known ancestors, can be tentatively interpreted as indicating that significant lexical change has occurred in all of them over time.

Additionally, based on the apparent clustering of many of the modern varieties in the network diagrams, the sequence similarity of a number of the modern varieties (such as Low German and Standard High German) appears to be greater than that between a number of the modern varieties and their ancestors. This suggests that there has been a degree of convergence between some varieties, particularly those which are geographically close to each other; this was nevertheless a fairly weak pattern and divergence was still significant. Such a branching pattern could be interpreted as a sign that a large number of lexical replacements have occurred in each individual language over time, but that the actual basic lexical replacements have either been due to borrowing amongst the languages, parallel development, or the influence of one variety's diction on that of another. This possibility is reinforced by the presence of complex reticulate relations between some varieties (as in simulation 3 of subgroup 3) which suggest complex relationships between the varieties. This is open to further investigation.

In all of the simulations involving reconstructed Proto-Germanic data items the Proto-Germanic node was placed closest to one or another of the recognised old Germanic languages or groups of languages. It was in these simulations that changes in word choice and/or coding strategy resulted in very noticeable differences in the networks. In the first two simulations in this subgroup the Proto-Germanic node was located most closely to the old West Germanic languages. This was done using the set of words marked Word Choice One; the majority of these reconstructed words were found in the data for the West Germanic languages, although a significant number were found in Gothic and Old Norse. This

positioning remained roughly the same under both the Majority Wins and Infoclean conditions, suggesting that the number of cognates with West Germanic terms in the Proto-Germanic data was large enough for coding strategy not to matter. A marked change was seen in the third simulation, which made use of a different set of words, marked Word Choice Two under the Majority Wins condition. In this simulation the Proto-Germanic node was placed in between the North Germanic section of the network and the rest of the Germanic languages. Its closest relative under this condition was Gothic, which was shown as being more closely related to the Proto-Germanic node than the old West Germanic varieties. The position of the Proto-Germanic node in simulation 4 of this subgroup was again closest to the Gothic node, although the North Germanic group was placed on the other side of the network to it, with the West Germanic varieties occupying an intermediate position; this simulation was based on Word Choice Two, but was performed under the Infoclean condition. This change suggests that under the Majority Wins condition, with Word Choice Two, the percentage of cognates shared by Old Norse, Gothic and Proto-Germanic was high enough to rearrange their positions, with Gothic having the greatest sequence similarity to Proto-Germanic. When items which were not attested in some of the languages were removed altogether, the sequence similarity between Proto-Germanic and Old Norse dropped enough for the Old Norse node to be placed far away again, while the similarity between Proto-Germanic and Gothic remained high. This strongly indicates that different ways of handling missing data can have a significant effect on the way nodes are placed within the network.

The simulations highlight a number of things about this method. Firstly, the choice of lexical items is critical to take into consideration, as even small differences in lexical item choice can have significant effects on the placing of nodes in the network. Secondly, different strategies for handling instances where data is missing can have an equally dramatic effect on the arrangement of the networks generated. In the first case, there are a number of things to take into consideration. Limiting the dataset to basic vocabulary has the advantage that the dataset can be kept to a reasonably manageable size. This is also likely to reduce the degree to which borrowing is likely to have occurred in the data, as items of basic vocabulary appear to be less likely to be borrowed than more peripheral lexical items, such as those relating to technology or warfare, which may be much more open to rapid change (Forster, Pölzin & Röhl, 2006). However, basic vocabulary items are not immune from borrowing and change (Campbell, 2004); English has borrowed a number of its basic vocabulary items, such as *person* (from Latin and Norman French (OED, 2015)), *mountain* (from Norman French (OED, 2015)) and *sky* (from Old Norse) (OED, 2015; Barber, Beal & Shaw, 2010). Depending on how intensive contact situations between languages may be, borrowing may still pose a significant problem even when basic vocabulary lists are used.

Additionally, the use of lexical data on its own throws up the problem of the amount of detail which can be incorporated into an analysis. Marking terms as cognates can be used to differentiate between varieties only in instances where whole items have changed. Beyond this, however, it does not allow for differentiation of cognates which are identical and those which have undergone sound changes. This means that more detailed, multileveled classifications cannot be undertaken using this method alone, simply because there is no way

of weighting different types of data against each other; treating cognates which have undergone sound changes by giving them different codes would result in them being treated in exactly the same way as instances where a completely different, non-cognate term appears in the data. This is a problem which presents a particular impediment to using this method to reliably show the relationships between languages as it means that classifications based on this method can only be done at a fairly vague resolution. This is the case even when careful data selection has taken place.

Another problem with the use of lexical cognates alone is that of borrowing between varieties within a subgroup or language family. For instance, Modern English *sky* is an Old Norse loanword; this is known from the fact that Modern English *sky* is pronounced with the cluster [sk] before the vowel (which in Old Norse was originally a high front rounded vowel (Faarlund, 1994; Fortson, 2004)); in Old English the cluster [sk] underwent palatalization to [ʃ] before front vowels (resulting in Old English *scio*, [ʃio]), while in Old Norse the cluster [sk] remained unchanged. Old English *scio*, "cloud" is the native cognate of Old Norse *ský*, and has not survived into Modern English. This relationship can be gleaned from looking at phonological data, but is obscured when only the cognate status of the term is used; it might be tempting to posit a solution to this problem such as assigning terms cognate status based only on whether they are terms native to a given variety as judged by phonological data, but this would still not allow the program to distinguish between these and terms which are non-native but still cognate with native terms and terms which are genuinely unique innovations (i.e.: likely to have arisen only in a certain variety) or borrowings from unrelated or very distantly related languages (which might not have sequences in the dataset). One way around this would be to establish a way of weighting different data items relative to each other such that instances where a native cognate with the same meaning for a concept has a higher weighting than a cognate which is an intrafamilial or intragroup loanword; an intrafamilial loanword which has a cognate in a particular language might then be given a greater weight than a loan from an unrelated language.

Innovations which are unique to certain branches or certain varieties could also be given great weights, as the development of a unique lexical item in a particular variety could be argued to indicate greater divergence of that variety from other varieties in a group of languages under examination (the presence of such an item in only one variety suggests that the variety in question has been separated from other related varieties in a way which has either prevented the diffusion of this item into neighbouring varieties, or has prevented the influence of other varieties from acting against this development; an example of such a case might be that of the semantic shift resulting in Old English *docga* replacing *hund* as the default word for DOG – this could be argued to have come about partially because Old English did not have the pressure from high levels of contact with other Germanic varieties for *hund* to be maintained as the default form at the time this shift took place. The shift could thus be viewed as an index of a period of separation. This is, however, conjecture, and there is ample evidence to show that this may not always hold for many languages; in fact isolation has in some instances led to the retention of archaic forms as isolated varieties do not come into enough

contact with other varieties to be influenced by changes happening in them (such as is the case for Icelandic).

It may be the case that there is a tendency for varieties which have diverged to greater degrees, and been separated for longer time periods, to have decreasing lexical similarity, but this trend could be weak. In most classifications of the Germanic languages which have been based on phonological and morphological data the North and West Germanic languages are classified as being more closely related to each other than either is to East Germanic; this was not clear from the simulations using lexical data, however, which consistently had a dramatic tripartite division of these groups, implying that all three separated in a three-way split from a common ancestor, with no clear intermediate grouping. This was because North and West Germanic varieties had large numbers of non-cognate items in their basic vocabulary datasets. On its own the use of lexical data results in their looking very different from each other from the outset; when phonological and morphological data are taken into account, the implication is that one of the groups' immediate ancestors underwent some dramatic changes in either the meanings of a number of basic vocabulary items, resulting in the replacement of some, or in the selection of different words with similar or related meanings. Another, less speculative, example highlighting the possible weakness of a trend for vocabulary differences to reflect degree of divergence, and possibly time of divergence, is the fact that Modern English is highly divergent from the continental West Germanic varieties in terms of vocabulary; on the basis of lexical items alone Modern English would be expected to occupy its own branch a considerable distance from the other West Germanic varieties, with Low German, Dutch, and High German likely to form a continental group of their own; however, from the point of view of diachronic phonology Dutch, Low German and Modern English would form their own group separate from High German, largely due to the effects of the High German Consonant Shift. To some extent some of these trends did appear to be reflected in the results of the simulations; in a number of the simulations with the modern Germanic languages the High German varieties did branch off basally from the rest, although the placing of other groups and varieties (notably West Frisian) was less consistent with well-established classifications. This is probably reflective of the fact that vocabulary in general is more open to outside influences and horizontal transmission (borrowing) than most other types of linguistic item.

The nature of vocabulary is such that on its own it is less suitable than other types of linguistic item for establishing ancestral relationships between languages. Vocabulary is more open to change, as speakers borrow new terms on a regular basis for new concepts, items and technologies; they coin new terms which may come to be borrowed by other speakers; and words undergo semantic shifts which may make analyses which require datasets of manageable sizes difficult. Part of the reason for this is that lexical items generally refer to concepts in the real world, such as physical items, entities and emotions, whereas closed-class morphological items, such as tense markers and articles, play grammatical roles and are as such frequently less open to outside influences and to shifts in meaning in the short term. While sounds do undergo changes, they are usually less open to outside influences and to shifts in meaning on their own simply because it is the combinations of sounds which make

up words and morphemes that are generally attached to concepts, not single sounds. Their changes also tend to be systematic, which usually cannot be said for lexical changes. It is this systematic behaviour and lower susceptibility to outside influences which has motivated the use of sound correspondences as the main way of establishing historical relatedness between languages (Lass, 1997; Campbell, 2004). It is generally agreed that many apparent exceptions to what are otherwise systematic sound changes can be explained by borrowing a particular word from a dialect or variety which did not participate in a particular change, or analogy (Campbell, 2004; Lass 1997).While it is not unknown for sounds or what seem to be sound rules to be borrowed (as in the case of Mamean languages which have borrowed a rule which palatalises velar consonants if the coda of their syllable is a uvular consonant from the K'ichean languages (Campbell, 2004)), this is less common than lexical borrowing. Additionally, the presence of certain morphological items or alternations in a paradigm which are unlikely due to parallel development or chance can also be used as a marker of relatedness, and the inclusion of these may serve to strengthen claims of relationships between two or more languages or language varieties (Lass, 1997). It may therefore be preferable to use data concerning sound correspondences, particular sound changes and the presence of certain morphological items or alternations in future research using this method. This would require identifying data which reflect shared innovations (which allow varieties to be subgrouped with each other), rather than shared retentions (which do not necessarily tell one much about how two varieties relate to each other; they do not provide evidence for being more closely related than other varieties, merely that for some reason these varieties did not lose an older feature) and then coding this data. For example, the presence of a particular type of sound change could be coded using the same amino acid codes for instances where the change is present; this could still be problematic, as instances where the change is not present at all would still need to be coded, and coding these instances where this change did not take place with the same amino acid code could over-estimate the extent to which other varieties which did not have the change are related; this is a possible avenue for further research.

One way this could be handled would be to have a clustering algorithm which could be used to compare data from parts of the dataset, allowing for progressively smaller scale comparisons. Another issue with this approach is the same as that for attempts to handle lexical data in a more detailed way than simply grouping cognates: weighting items relative to each other. Again, the problem with this is how to do it accurately, especially considering that in many instances certain sound changes are thought to be more or less likely than others, but how much more or less is unknown. This presents a challenge when using computational methods which require actual values to be used, as it may be uncertain to what extent using one value over another might influence the grouping of languages. Ascertaining this would require a greater knowledge of the relative probabilities of as many sound changes as possible across as many language families as possible; this requires more research.

This does not necessarily mean that the use of lexical data in quantitative simulations such as these is completely unsuitable for historical research purposes; however, it may not be suitable for establishing deep phylogenetic relationships between languages. Where it may be

useful in ascertaining how certain subsets of various languages' lexicons relate to each other. In addition to this, these results may be combined with other forms of data to reveal interesting relationships or indicate the possible presence of phenomena which are of further research interest. If the results of the simulations described above are to be tentatively accepted, they suggest that, despite phonological and morphological data pointing to greater relatedness between the North and West Germanic languages than between either and the East, there was significant change in many of the items of basic vocabulary they use, resulting in no subgrouping of them together in the network. Additionally, there was no distinct Anglo-Frisian subgroup based on lexical data in the simulations using the old Germanic varieties; Old Frisian and Old Dutch were grouped as closer relatives. This could be taken to suggest either that the influence of Dutch on Frisian began at an early date, or that the separation of Old English from Old Frisian resulted in semantic and lexical innovations in Old English which were not present in Old Frisian (or, for that matter, the other Old West Germanic languages). This is however something which can only be ascertained through further, finer grained research.

In terms of methodology, the use of lexical data may leave analyses very open to the quality, type and amount of attested information, particularly if older or not very well attested varieties are being examined. This is due to the fact that word choice may have a significant impact on the outcomes of simulations; this is something which would possibly be less of a problem if sounds correspondences or morphological items were used, as these would be less likely to be influenced by genre or the paucity of items used (if the languages examined were scantily attested or from a range of texts). This could possibly also allow for a larger dataset in instances where data is unattested or missing, as certain sounds might occur in a large number of instances in a text, even though the number of words is low. For newer languages or very well attested ones this would be unlikely to be as big a problem, although the use of phonological and morphological data might nevertheless be preferable if what is desired is establishing a phylogeny, as these are less open to borrowing and the effects of semantic shifts. Further research using these types of data is required to determine how effective they may be.

Different strategies for handling missing data items (in this case unattested words) had different effects on the placing of nodes within the network. Using the Infodelete strategy, in which instances where missing basic vocabulary items were coded as deletions of characters, had the effect of increasing the degree to which certain sequences were shown as differing from each other. This could be seen in subgroup one of the simulations, which used only data from the old Germanic languages. In the second of these simulations, which was performed under the Infodelete condition, the Old High German, Old Dutch and Old Frisian nodes were shown as differing from each other by a much larger number of terms than in the simulations which were not performed under the Infodelete condition. This was because the terms EGG, LOUSE, ROUND and SEED, which were coded as deletions in the Old Dutch and Old Frisian sequences, were not ignored for these two languages by the algorithm, but were instead treated as further "mutations". This caused the degree by which they differed from Old High German to increase. The Old Dutch and Old Frisian nodes also increased in the

degree by which they differed from each other, as the terms for LEAF, MAN, NOSE, ROOT, and SWIM were unattested and therefore coded as deletions in Old Dutch, while terms for ASHES and BARK were unattested in Old Frisian and also coded as deletions. All of these deletions at different sites within sequences of these languages lowered their overall sequence similarities significantly. This indicates that coding instances of missing data items carries a high risk of overestimating the degree of divergence between two or more language varieties; this would be especially likely where one variety is much better attested than another. This suggests that this strategy should not be used. Using a Majority Wins strategy may have the opposite effect, i.e. it overestimates the degree of similarity between two or more varieties. In the first simulation using only data from the old Germanic languages this strategy was used: unlike the Infodelete simulation, the differences between Old High German, Old Dutch and Old Frisian were much smaller, with less divergence having taken place between them. This simulation showed Old Frisian and Old Dutch as diverging from the ancestral line they share with Old High German by the concept BIG; Old Dutch differed from Old Frisian by BARK. This is a significant reduction from the Infodelete condition's estimation of their degree of divergence; the Majority Wins condition resulted in a network which indicates that their basic vocabularies are very similar, while the Infodelete condition created one which implies a lot of difference. This is because all unattested terms were assigned the same codes.

The Infoclean condition, which involved running the simulation with all missing data items completely excluded, resulted in an Old Germanic network which was very similar to the Majority Wins network; however, in this network Old Dutch and Old Frisian were grouped together as one node, indicating that the sequences were identical. This was due to the removal of all concepts for which they had no attested terms, with the result that they had no non-cognate terms in their respective sequences afterwards. Another instance where leaving out missing items all together resulted in a different grouping of languages in the network occurred in subgroup three, in which data from old and modern Germanic varieties was used. In the semantically strict Majority Wins simulation Old Norse and Icelandic diverged from different parts of the North Germanic section of the network, with Old Norse branching from a section shared with Faroese (and thus implying that Faroese had the greatest basic vocabulary similarity to Old Norse), while Icelandic diverged from the Norwegian varieties, with Norwegian Nynorsk being shown as having the greatest basic vocabulary similarity to it. In the semantically strict Infoclean simulation, Old Norse and Icelandic branched off from a common branch; in this case Icelandic was shown as Old Norse's closest relative in terms of basic vocabulary. Faroese occupied a position basal to them, indicating that it is still similar to them but not as similar as Icelandic and Old Norse are to each other. These differences in grouping are all significant, and indicate that different coding strategies designed to handle missing data can have dramatic effects on how the algorithm groups the languages. As the use of an Infoclean strategy does not rest on making possibly questionable assumptions about missing data items (as the Majority Wins strategy does) and does not carry the same risk of overestimating the degree of divergence between varieties as the Infodelete strategy does, this may be the safest of the three strategies to use, particularly if very little other data is available to inform coding strategies. However, it does carry the risk that with badly attested languages datasets could become quite small much more sensitive to the codes assigned to the

remaining items, as the results of a smaller dataset would be more likely to be thrown out by a small number of "outlier' pieces of information (such as words with no cognates in a given variety). The issues presented above may be a motivation for further research involving other data, such as phonological and morphological, to determine if they are better for use with quantitative measure such as these.

This does not necessarily mean that the use of lexical data is pointless; methods such as this could still prove valuable in exposing trends in how areas of vocabulary in various languages are related to each other. This may be especially so when these results are compared to those based on studies using other types of data. More research is required in this area.

Conclusion

This dissertation has brought up a number of issues involving the use of lexical data with a quantitative phylogenetic networking method such as median-joining for use in determining how languages are related to each other. Firstly, the use of lexical data, even data limited to basic vocabulary, is sensitive enough to differences in how words are coded, and how they are chosen, to have a significant impact on how languages are grouped, particularly at finer resolutions. Additionally, in instances where data may be missing the preferable way of handling this seems to be to remove these data items altogether, as coding the missing items as deletions carries a high risk of overestimating the degree to which varieties differ from each other in terms of basic vocabulary, and using a Majority Wins strategy may carry the opposite risk of underestimating the degree of divergence in vocabulary between varieties. This unfortunately carries the risk of dramatically reducing the size of the data set in instances where large numbers of items are not attested; this may reduce the reliability of results due to the greater effect of "outlier" items on sequence similarities. These points, together with the fact vocabulary items are more open in general to outside influences than other linguistic structures, may be a motivating factor for future research using phonological and morphological data instead. This also suggests that any results obtained in this dissertation are prudently taken with caution until more research with other types of data can be performed. Taken in relation to more traditional classifications of the Germanic languages, this research indicates that any tendency for lexical data to reflect degree of divergence in relation to other types of data may be very weak. This is likely due to the greater tendency of vocabulary to change under differing social and environmental conditions than many other types of linguistic structure. There is yet room for much more research on these quantitative methods.

Acknowledgements:

I would like to thank Professor Jarich Hoekstra of the Department of Frisian Studies at the Insitute of Scandinavian, Frisian and General Linguistic Studies (Institut für Skandinavistik, Frisistik und Allgemeine Sprachwissenschaft (ISFAS) - Abteilung für Frisistik) at Christian-Albrechts-Universität, Kiel, whose help with Frisian etymology was invaluable to this dissertation. I would also like to thank Pieter Duijff of the Fryske Akademy for his help with West Frisian etymologies. Additionally I would like to thank Anthony Rowley of the Kommission für Mundartforschung, Bayerische Akademie der Wissenschaften, Joseph Salmons of the Department of German, University of Wisconsin and Erin Noelliste of Indiana University Bloomington, for their information on various lexical and phonological features of southern High German varieties. For help with Icelandic etymologies I thank Professor Jón Axel Harðarson, Professor Magnús Snædal Rosbergsson and Professor Guðrún Þórhallsdóttir of the University of Iceland (Háskóli Íslands). For providing valuable information on the workings of median-joining algorithms I have Dr Peter Forster of Murray Edwards College, University of Cambridge, and Dr Arne Röhl of the University of Hamburg (Universität Hamburg) to thank. For help with the details of sound changes and word etymologies in Afrikaans, I thank Dr Andrew van der Spuy of the University of the Witwatersrand, South Africa, and Professor J. Conradie of the University of Johannesburg, South Africa.

References:

Allen, W. B. (1989). *Vox Latina: A Guide to the Pronunciation of Classical Latin*. Cambridge: Cambridge University Press. Pp. 14-15; 22-23.

Askedal, J. O. (1994). Norwegian. In König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. London: Routledge. Pp. 219-220; 267-268

Atkinson, Q. D., & Gray, R. D. (2005). Curious Parallels and Curious Connections-Phylogenetic Thinking in Biology and Historical Linguistics. *Systematic Biology, 54(4),* pp. 513-526.

*[8]Bahnick, K. (1973). *The Determination of Stages in the Historical Development of the Germanic Languages by Morphological Criteria*. The Hague.

Bandelt, H-J., Forster, P. & Röhl, A. (1999). Median-Joining Networks for Inferring Intraspecific Phylogenies. *Molecular Biology and Evolution, 16 (1),* pp. 37-48.

Barber, C., Beal, J. C., & Shaw, P. A. (2012). *The English Language, 2nd Ed.* Cambridge: Cambridge University Press. Pp. 66; 67; 88; 89; 90; 97-100; 110-111; 119; 143; 149; 201-204; 206-207; 224; 226; 242

Barnes, M. P. & Weyhe, E. (1994). Faroese. In König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. London: Routledge. P 192; 195.

Bremmer. R. H. (2009). *An Introduction to Old Frisian: History, Grammar, Reader, Glossary*. Amsterdam: John Benjamins. P. 123

Campbell, A. (1959). *Old English Grammar*. Oxford, Oxford University Press. Pp. 50-51.

Campbell, L. (2004). *Historical Linguistics: An Introduction, 2nd Ed.* Edinburgh University Press. Pp. 16-19; 79; 103-104; 201-210; 303-305;

Croghan, V. & Holmqvist, I. (2010). *Teach Yourself Complete Swedish*. London: Hodder & Stoughton. P.87.

Crystal, D. (2005). *How Language Works*. London: Penguin-Allen Lane. P. 376

Crystal, D. (2007). *Words, Words, Words*. Oxford: Oxford University Press. Pp. 90-91.

Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press. Pp. 191-192.

De Schutter, G. (1994). Dutch. In König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. London: Routledge. Pp. 447-448;

Duden. [Online Resource]. Accessed at: www.duden.de. [Accessed: 2015].

[8] All references which are preceded by an asterisk are secondary references.

Faarlund, J. T. (1994). Old and Middle Scandinavian. In König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. London: Routledge. P 38-39; 40; 43.

Forster, P., Toth, A., & Bandelt, H.-J. (1998). Evolutionary Network Analysis of Word Lists: Visualising the Relationships between Alpine Romance Languages. *Journal of Quantitative Linguistics, 5(3),* Pp. 174-187.

Forster, P., & Toth, A. (2003). Toward a phylogenetic chronology of ancient Gaulish, Celtic and Indo-European. *PNAS, 100(15),* pp. 9079-9084. [Accessed Online]: www.pnas.org/cgi/doi/10.1073/pnas.1331158100

Forster, P., Polzin, T., & Rohl, A. (2006). Evolution of English Basic Vocabulary within the Network of Germanic Languages. In: Forster, P., & Renfrew, C. (Eds.): *Phylogenetic Methods and the Prehistory of Languages*. McDonald Institute for Archaeological Research. Pp.131-137.

Fortson, B. W. (2004). *Indo-European Language and Culture: An Introduction, 2nd Ed*. Chichester: Blackwell Publishing. Pp. 247; 300-301; 301-302; 306-308; 312; 313; 315; 317; 318; 323-324; 330; 331-332; 338-378.

Fulk, R. D. (1998). The Chronology of Anglo-Frisian Sound Changes. In: *Approaches to Old Frisian Philology*. Bremmer, R. H. Jr., Thomas, J., & Vries, O. (Eds). Amsterdam: Rodopoi, Pg. 185.

De Geïntegreerde Taalbank. [Online Resource]. Accessed at: www.gtb.inl.nl. [Accessed: 2015]

*Grimm, J. (1840). *Deutsche Grammatik IV*. Göttingen

Hagège, C. (2011). *On the Death and Life of Languages*. Yale University Press. Pp. 75-106.

Heggarty, P. (2006). Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully be Applied to Language Data- and to Dating Language?. In Forster, P., & Renfrew, C. (Eds.): *Phylogenetic Methods and the Prehistory of Languages*. McDonald Institute for Archaeological Research. Pp. 183-194.

Heggarty, P., Warren, M. & McMahon, A. (2010). Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B, 365.* Pp. 3829-3843.

Henrikson, C. & van der Auwera, J. (1994). The Germanic Languages. In: König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. P. 7; 5-9

Hoekstra, J. & Tiersma, P. M. (1994). Frisian. In: König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. Pp. 505-506; 528-529

*Holtzmann, A. (1870). *Altdeutsche Grammatik*. Leipzig

Kooij, J. G. (1987). Dutch. In Comrie, B. (Ed): *The World's Major Languages*. Beckenham, Kent: Croom Helm. Ltd. Pp. 139-157.

Lass, R. (1997). *Historical Linguistics and Language Change.* Cambridge: Cambridge University Press. Pp. 34; 125; 134; 153; 169-171.

Lehmann, W. P. (1994). Gothic and the Reconstruction of Proto-Germanic. In: König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. London: Routledge. Pp. 21-22; 23

Lundskær-Nielsen, T. & Holmes, P. (2011). *Danish: An Essential Grammar, 2nd Ed.* Abingdon: Routledge. Pp. 48; 91.

*Makaev, E. A. (1965). *Jazyk drevnejšix runičeskix nadpisej*. Moscow

*Maurer, F. (1952). *Nordgermanen und Allemannen.* Bern-München

McMahon, A. & McMahon, R. (2006). Why Linguists Don't Do Dates: Evidence from Indo-European and Australian Languages. In Forster, P. & Renfrew, C. (Eds): *Phylogenetic Methods and the Prehistory of Languages*. McDonald Institute for Archaeological Research. Pp. 153-160.

Millikan, R. G. (2004). *Varieties of Meaning: The 2002 Jean Nicod Lectures*. Massachusetts: MIT Press. Pp. 16-20.

Nielsen, H. F. (1989). *The Germanic Languages: Origins and Early Dialectal Interrelations*. Tuscaloosa: University of Alabama Press. Pp. 28-29; 31-32; 67-69; 72; 73-75; 95

Nielsen, H. F. (2000). *The Early Runic Language of Scandinavia: Studies in Germanic Dialect Geography*. Heidelberg: Universitätsverlag C. Winter, GmbH. Pp. 10; 57-58.

Ordbog over det Danske Sprog. [Online Resource]. Accessed at: www.ordnet.dk/ods . [Accessed: 2015]

Orel, V. (2003). *A Handbook of Germanic Etymology*. Leiden: Brill.

Oxford English Dictionary. [Online Resource]. Accessed at: www.oed.com. [Accessed: 2015]

Renfrew, C., & Forster, P. (2006). Introduction. In: Forster, P., & Renfrew, C. (Eds.): *Phylogenetic Methods and the Prehistory of Languages*. McDonald Institute for Archaeological Research. Pp. 1-8.

Salmons, J. (2012). *A History of German: What the past reveals about today's language.* Oxford: Oxford University Press. Pp. 84; 192

*Schleicher, A. (1860). *Die deutsche Sprache*. Stuttgart.

Stern, H. R. (1984). *Essential Dutch Grammar*. Dover Publications, Inc. Pg. 36.

Strandskogen, Å.-B & Strandskogen, R. (1995). *Norwegian: An Essential Grammar.*Abingdon: Routledge. Pp. 74-75

Svenska Akademiens Ordbok. [Online Resource]. Accessed at: www.g3.sprakdata.gu.se/saob/

Swadesh, M. (1951). Diffusional cummulation and archaic residue as historical explanation. *Southwestern Journal of Anthropology, 7*, pp. 1-21

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society, 96*, pp. 453-463

Thráinsson, H. (1994). Icelandic. In: König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. London: Routledge. Pp. 142-148; 157

Van der Wal, M. & Quak, A. (1994). Old and Middle Continental West Germanic. In: König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. London: Routledge. Pp 73; 74; 91-92.

Van Kemenade, A. (1994). Old and Middle English. In: König, E. & van der Auwera, J. (Eds): *The Germanic Languages*. London: Routledge. Pp. 114; 116; 118

Voyles, J. B. (1968). Gothic and Germanic. *Language, 44*, pp. 720-746.

*Wrede, F. (1919). Zur Entwicklungsgeschichte der deutschen Mundartenforschung. *Zeitschrift für deutsche Mundarten.* 19: 270-83.

Appendix A

The Data: Swadesh 100-word entries for the Germanic languages

The data for this analysis consists of lexical items for concepts used in the Swadesh 100-word list in the following old Germanic languages: Old Norse, Old English, Old Saxon, Old Frisian, Old West Low Franconian (Old Dutch), Gothic and Old High German. The modern descendants of these languages (with the exception of Gothic) that have been used in this study include Norwegian Bokmål, Norwegian Nynorsk (sometimes referred to as Neo-Norwegian in the literature), Swedish, Icelandic, Danish and Faroese; Modern English, Modern Low German, Modern Dutch, Flemish and Afrikaans; three varieties of High German (Modern Standard High German, Bavarian and Swiss German), and four varieties of Modern Frisian (West Frisian, Ferring Frisian, Frasch, and Saterland Frisian). Additionally, reconstructed Proto-Germanic forms are included. The selection of the varieties used in this dissertation was motivated by a desire to have a number of different varieties which would allow for testing of the network program under a number of conditions, rather than for establishing a comprehensive classification of the various Germanic languages and dialects. Certain varieties, such as Yiddish and Pennsylvania German, were not included as there were difficulties in obtaining reliable datasets for them timeously, and their inclusion was not felt necessary for testing the networking program.

The coding of the words as cognates was done on two levels. The first simply looked at whether languages had the same forms or not for a given concept on the Swadesh 100-word list. Thus, regardless of whether or not a form in one language may have been borrowed from another language, if they were clearly the same form, or were based on the same form, the were marked as cognates. For example, the Modern English word *leg* is known to be a borrowing from Old Norse *leggr*; despite the fact that Old English *scanca* has survived into Modern English as *shank*, this will not be used instead of *leg* because it is no longer the most commonly used form, and has undergone a slight semantic shift. In this instance, *leg* and *leggr* will simply be coded as cognates. This is in line with the coding done in most papers up to now on the use of computational phylogenetic methods in linguistics which use lexical data. The group of simulations performed under this strict coding scheme are collectively referred to as the Semantically Strict simulations. The second way of coding the data took the history of an individual word into account. Following the example given above, the word *shank* would be used in place of *leg*, as it is the descendent of the Old English form; this condition is referred to here as the Semantically Lax condition. This coding strategy looks at the survival of morphological forms regardless of their meaning. In instances where an older form has not survived at all into a more modern form of one of the languages, the most used form is selected instead, as the older form is no longer part of the language. By doing this, the effects of using different criteria to choose lexical information for such a study can be gauged. In cases where a form is unattested in one of the languages, there are several possible strategies to pursue. One of these is to mark this as a deletion. This is, however, extremely problematic, as the program will still treat this as a change, which may result in an exaggeration of the distance of relatedness between languages, especially if one of them has a large number of items missing.The second strategy was to assign the missing characters the

most widespread code in the data set. This can also be quite problematic as it could result in a group of languages as being more closely related than they actually are, simply because there are missing lexical items. The third strategy which could be pursued is that of simply removing the whole set of characters where there is missing data entirely. In this way, these missing characters are not even included in the analysis, preventing either of the two problems mentioned above from cropping up if the overall dataset is very complete and is large. However, if large numbers of character sets must be removed, or if the whole data set is quite small, this could still result in over- or underestimation of relatedness, with similar effects on the end result of the simulation. Each one of these strategies will be used in the simulations as well, with the deletion condition called Infodelete, the majority wins condition called Majority Wins and the simple removal of problematic items called Infoclean.

For the concept ALL, under both the Semantically Strict and Semantically Lax conditions, every one of the words was coded as being cognate. They all descend from Proto-Germanic **allaz*, with all having largely preserved its structure, with either a normal or geminated lateral approximant, either intervocalically or word finally, and a rounded vowel of some type beginning the word. The only exceptions to this are Faroese and Icelandic, where the geminated lateral approximants became stop-approximant combinations ([dl] and [tl], respectively; Barnes & Weyhe, 1994). In some cases, the word is monosyllabic, and in others disyllabic; in the first case this is simply due to syncope of the final vowel, while in the second the additional vowel has been retained, although it is often reduced to a schwa. In all instances, however, these forms can all be seen to descend regularly from the Proto-Germanic form, with reduction of the form having occurred in all instances, and with no evidence of borrowing (in instances where there are two forms such as Old Frisian, Ferring Frisian, Saterland Frisian and Frash Frisian, the doublets are due to dialectal variation, not borrowing). As ALL is part of the concepts which are generally considered to have closed class words, the probability that the word for ALL has been borrowed is low.

The concept ASHES had cognate forms in all of the languages used, with the exceptions of Old Saxon and Old Frisian, for which these words are not attested (although it seems unlikely they would have used terms which were very different from the other languages, given that cognate forms are found across all of them). The modern Frisian forms all begin with either a front vowel of some form (Saterland, Frasch and Ferring Frisian) or, in the case of West Frisian, a glide-front high vowel combination. These are all forms which could plausibly have arisen from a front vowel, in line with the phenomenon known as "Anglo-Frisian brightening", in which a Germanic short **a*, [a], was fronted and raised in all environments except when followed by a nasal (Campbell, 1959). This was generally realised in Old Frisian as [e] or [ɛ], and in Old English as [æ] (hence *æsce*, [æʃɛ]). In all of the attested forms the structure of the word is ($C_{[+gli]}$) $VC_{[+ sib]}$($C_{[+obst,+vel]}$)(V).[9] These forms all descend from Proto-Germanic **askon*, with deletion of the final nasal having taken place (Orel, 2003). As there are two characters missing, this concept was coded using the strategy mentioned above, with

[9] Note: a number of the features used in this analysis are not part of the standard set. The reason for this is that some non-standard features were deemed more useful to the determination of cognates than some of the standard features.

an Infodelete, a Majority Wins, and an Infoclean condition. All attested forms were coded as cognates in the semantically lax and semantically strict conditions.

The concept BARK had four separate attested forms, and was additionally unattested in Gothic and Old Frisian. The form *rind(a)/rinta/rin/rinj* is attested in many of the West Germanic languages, such as Old English, Old Saxon, Old High German, the modern High German varities, and Ferring and Frasch Frisian. These were all coded as cognates based on the presence of the structure $C_{[+postalv/alv,+trill]}V\ C_{[+alv,+nas]}$, with the additional syllable or palatalization following the nasal of secondary importance. The second form, *borkr/bark/buark/Boarke/børkur/börkur* is attested in Old Norse, Swedish, Norwegian[10], Danish, Modern English, Ferring Frisian, Saterland Frisian, Low German, Faroese and Icelandic. All these forms possess the same core structure, $C_{[+vo,+labl,+plo]}V(C_{[+postalv,+trill/app]})C_{[+vel,+plo]}$. Based on this, under the semantically strict condition, these were coded as cognates (B). The third form, *bast/bas* occurs in West Frisian, Flemish, and Afrikaans. These forms all present as a monosyllabic word of the form $C_{[+vo,+lab,+stp]}\ V_{[+low]}\ C_{[-vo,+alv,+fric]}$. Based on this, under the semantically strict condition, these were coded as cognates (C). The lexemes *skors/schors/skorza* occurred in Afrikaans, Dutch and Old High German; because of their almost identical forms, they were coded as cognates (D) for the semantically strict set of simulations. In Old High German, *skorza* and *rinta* are both recorded, and there does not seem to be any major difference in meaning between them. A similar situation was encountered with Afrikaans, which has both *bas* and *skors*. As two different characters items cannot be coded simultaneously unless an additional character is inserted into the sequence, in both cases the synonyms were initially treated as if each was the only character, and more than one simulation was run. Orel (2003) reconstructs two lemmas for BARK in Proto-Germanic, **skurtaz* and **barkuz*. The first was treated as the ancestor of the D forms and assigned the code (D), the second was assigned the code B. Under the semantically lax condition, the Modern English word *rind* was used, as it descends from Old English *rind*, although its meaning has changed from "bark" to "a hard outer layer or skin on a variety of objects"; Modern English *bark* is a borrowing from Old Norse *börkr* (OED, 2015).

Concept number four was problematic in some sense as it could easily be either BELLY (referring to the part of the abdomen in which the intestines are found) or STOMACH (the internal organ); the reason for this is the terms for these are sometimes not always used exclusively for one of these meanings in some languages (for instance in Modern English it is quite common to use "stomach" to refer to the belly). It was therefore decided to run two semantically strict simulations, one using BELLY and the other STOMACH with the respective senses above. Under BELLY, one set of cognates in the semantically strict simulation condition had the core structure /$C_{[+lab,-nas]}$Vm(b)/, where the parenthesised voiced bilabial stop is widespread, but not universal. This structure occurred in Gothic *wamba*, Old English *wamb*, Old Saxon, Old High German and Old Dutch *wamba*, and Old Frisian *wamme*. These all descend from Proto-Germanic **wamƀō* (Orel, 2003) and as such were

[10] For the sake of brevity, if the variety of Norwegian is not specified, it means a form occurs in both Nynorsk and Bokmål.

coded as cognates. The Old Norse word for a human BELLY was *magi*; this has been continued with the same meaning into Norwegian *mage* and Icelandic *magi*; these were assigned their own codes. A number of words had the structure /bVC$_{[+obst]}$/, as in Ferring Frisian *bük*, Saterland Frisian *Buuk*, Frasch Frisian *bük*/*buuk*, Low German *Buuk*, Standard Dutch and Flemish *buik*, Afrikaans *buik*, Standard High German and Bavarian *Bauch*, Swiss German *Buuch*, Danish *bug*, Faroese *búkur*, and Swedish *buk*. Old High German *būh*, Old Dutch *būk* and Old English *buc* are attested with meanings ranging from BELLY to "something swollen" (Duden, 2015; SAOB, 2015); while this word fell out of use in English, in the continental West Germanic languages it replaced words such as *wamba* with the meaning BELLY. In Danish and Swedish the words *bug* and *buk* are thought to be borrowings from Middle Low German (SAOB, 2015); in Faroese the word *búkur* entered the word via Danish at a later date. As these words have become the normal words for BELLY in Danish, Swedish and Faroese, they were coded as cognates of the West Germanic forms under the semantically strict condition. Modern English *belly* is descended from Old English *bælg*/*bælig*, which originally meant "bag, sack"; it was due to a semantic shift, possibly in reference to the belly bulging or acting as a container that it took on the modern meaning (OED, 2015). West Frisian has the word *liif*; this was assigned its own code. Alongside *bük*/*buuk* Frasch Frisian has the word *lif*; in an additional semantically strict simulation this was coded as a cognate of the West Frisian form. If the concept BELLY is replaced with STOMACH (the internal organ), a common root structure seen is /mV$_{[+low]}$C$_{[+vo]}$/, as in Old English and Old Frisian *maga*, Old High German *mago*, Old Norse *magi*, Standard Dutch, Flemish and Afrikaans *maag*, Standard High German and Bavarian *Magen*, Swiss German *Magen*, Danish *mave*, Faroese *magi*, Icelandic *magi*, and Norwegian and Swedish *mage*. Based on these the Proto-Germanic **mazen-* (SAOB, 2015) or **maʒōn* (Orel, 2003) is reconstructed. These were coded as cognates under the semantically strict condition. Modern English *stomach* is a loanword from Old French *estomac*/*stomaque*; however, it has completely replaced the older Germanic form, and so was simply assigned its own code for the semantically strict simulation; however, the continuation of Old English *maga* does survive with a different meaning, namely "stomach of an animal, throat or gullet of voracious animal", in Modern English *maw*. This was included in a semantically lax simulation. Finally, as there was no word for STOMACH (internal organ) recorded for Gothic, Old Saxon and Old Dutch, simulations were run in which these missing pieces of data were treated as deletions (Infodelete), coded with the most common code (Majority Wins) and where the character as a whole was simply removed (Infoclean).

There were four different lemmas for the concept BIG in the languages studied. This actual word *big* only occurred in Modern English with the meaning "big" (the Danish word *big*, [biˌʔ] means "to build"). The forms *mikils/micel/mikil/mihhil* are found in Gothic, Old English, Old Saxon and Old High German; these all have a voiced bilabial nasal as the first consonant in the word, the vowel [i] following, an obstruent which has descended from the voiceless velar plosive [k] (in the case of Gothic and Old Saxon this sound has remained; in Old English [k] was palatalised to [ɟ] before front high vowels; in Old High German the fricative represented by <hh>, probably [x] or [ç], is attested sporadically where other Germanic languages have a [k]). These were coded as cognates under both the semantically

strict and lax conditions. The lemmas *groz/grōt/grāt/grat/groot/grut/groß/grooss* made up the majority of forms seen in the data, with these occurring in Old High German, Old Dutch, Old Saxon, Old Frisian, all of the modern Frisian varieties, Modern Dutch, Flemish, Afrikaans and all of the Modern High German varieties. All of these lemmas have similar forms, with either a voiced velar stop or a voiceless velar fricative (in the case of Old Dutch and its modern descendants, Standard Dutch, Flemish and Afrikaans; this sound, [ɣ], is believed to have descended from Proto-West Germanic **g*/**ɣ* (De Schutter, 1994)), followed by either an alveolar trill or a uvular trill (this sound is thought by some to have spread out from several centres of prestige and replaced an original native alveolar trill in Modern Standard Dutch, and Modern High German; however this is uncertain (Salmons, 2012)) a long, low vowel, and either a voiceless alveolar stop, [t], or, in the High German varieties, a voiceless alveolar fricative, [s] (the change of the voiceless velar plosives to fricatives was part of a wholesale consonant shift known as the High German Consonant Shift, which changed the manners of articulation of a large number of stops in the pre-Old High German period). In addition to this, there is no evidence that they were borrowed by one language from another; they were therefore coded as cognates for both the semantically strict and lax sets of simulations. Old Norse has the word *stórr*, which is continued in its descendants, with Danish, Swedish and Norwegian *stor*, Faroese *stórur*, and Icelandic *stór*. As these all have the same meaning and largely the same form, they were coded as cognates. The concept BIG brought in some difficulties with regard to the fact that a number of words can be used as synonyms with it; although there may be a slight semantic difference between them, it is difficult to say just how great this difference is, and so the choice of a particular form may be somewhat arbitrary. The Modern English word *great*, when used to refer to size, tends to suggest something larger than *big*, but this is by no means always the case; it can also be used metaphorically to refer to something or someone very good or influential, which is not the case for *big*. As *great* is of the same form and meaning (when referring to size) as the other Modern West Germanic forms given above, this was also factored into a simulation. Interestingly, Orel (2003) only reconstructs one word for BIG in Proto-Germanic, **mekilaz*. This was accordingly coded with the *mikil* forms.

The concept BIRD was represented overwhelmingly by forms descended from Proto-Germanic **fuglaz*. These forms occur throughout the Germanic languages, although the situation with English is slightly more complicated than the other Germanic languages; this will be discussed shortly. The canonical structure of all of these forms goes along the lines of $C_{[-vo,+lab\text{-}dent,+fric]}V_{[+back]}C_{[+vel]}\ C_{[+lat\text{-}app]}$ or $C_{[-vo,+lab\text{-}dent,+fric]}V_{[+back]}C_{[+vel]}VC_{[+lat\text{-}app]}$ (in Afrikaans the lack of the intervocalic velar consonant is due to a process whereby intervocalic velar fricatives were deleted, giving such forms as *voël*, [fuːəl], from Middle Dutch *vogel*, [foːxəl], and morphological alterations such as *oog*, [ʊəx], "eye", *oë*, [uːə], "eyes"). The presence of an additional syllable in both the Faroese and Icelandic words is due to a tendency in both languages to insert an epenthetic vowel between the masculine nominative marker *–r* and the noun or adjective it was attached to; this was the normal pronunciation by the 17th century and was standardised when both languages' orthographies were standardised. With these forms all conforming to a particular structure and having the same meaning, they were coded as cognates. The English forms present an interesting example of semantic change and how it

can result in forms which were found throughout the history of the language effectively swapping meanings over time. Modern English *bird* is the neutral, everyday word used when talking about birds; however, its Old English ancestor, *bridd*, was not the more neutral form, having the much more specific meaning of "a young bird". In Old English the most common and neutral word representing the concept BIRD was *fugol*; this has been continued in Modern English as *fowl* (the intervocalic voiced fricative [ɣ], which the <g> represents, was lenited to the glide [w] during the early Middle English period (van Kemenade, 1994)), which has the more specific meaning of "a wild bird" often in the context of hunting. Thus Modern English has a cognate of the **fuglaz* forms, and Old English an additional word for BIRD, albeit with a more specific meaning; these forms were used in the semantically lax simulations, although under these conditions neither becomes preferable since both forms are present anyway. This issue will be discussed in more detail in the "Discussion" below.

Concept seven in the list is TO BITE, which has cognate forms across the Germanic languages, all being descended from Proto-Germanic **bitanan*. All character entries under this form were thus entered as cognates, under all conditions, as there is no evidence of borrowing between languages for this character, and all entries have the same basic structure (voiced bilabial plosive-vowel-voiceless alveolar stop or fricative; beyond this the endings vary from language to language, despite all descending from the Proto-Germanic **–anan* infinitive ending).

The characters for the concept BLACK were largely the same across the languages sampled. Each language had a term which was a descendant of the Proto-Germanic word **swartaz*. In almost all of the languages concerned (except for Modern English) the most common word for BLACK has either a voiced or unvoiced alveolar fricative ([s] or [z]) followed in most instances (Danish, Norwegian Bokmål and Ferring Frisian being the exceptions) by either a voiced labio-velar approximant or a voiced labio-dental fricative ([w] or [v], respectively; where the old West Germanic languages regularly preserve Proto-Germanic **w*, [w], Old Norse has fortified this to [v]) a vowel which is generally [+back], with an alveolar trill or approximant (except in non-rhotic varieties of English, where the approximant /ɹ/ is regularly deleted in coda position, with compensatory vowel lengthening) and an alveolar plosive. In the Modern High German varieties the post-alveolar fricative [ʃ] takes the place of [s] or [z]; this is due to a change which occurred in the Middle High German period in which voiceless alveolar fricatives followed by another consonant were palatalised: /s/⟶[ʃ]/_C.[11] The presence of the affricate [ts] (rather than a plosive; Standard High German *Schwarz*, [ʃvaʀts]) at the end of the word in the Modern High German varieties is due to the High German consonant shift, which resulted in a shift in the manner of production of stops in certain positions to fricatives, and of fricatives in intervocalic and coda position to stops. The lack of a glide or fricative following the voiceless alveolar fricative in the onset of the characters in Danish and Ferring Frisian is simply to lenition of this phone; in Norwegian Bokmål there are two forms given, *sort* and *svart*, with the first reflecting the strong Danish influence on Bokmål, and the second the form descended from varieties less influenced by Danish (much

[11] When the voiceless alveolar fricative came before the voiceless velar plosive [k], the plosive was deleted, although it is likely this is a later development.

like Nynorsk). As a result of these regularities, all of these forms were coded as cognates. In Modern English the word *black* is used to denote BLACK; however, the archaism *swart* ([swɔːt]) does exist, and is a continuation of the Old English form *sweart*. In Old English the word *blæc* is attested and had the same meaning as *sweart*. This naturally raises the question as to which form to use. For the first set of simulations done, the most common forms were used (*sweart* occurs more often in manuscripts than *blæc* does, and so was treated as the normal regular form of the word); additional simulations were later done using the less attested or more archaic words. *Black* and *blæc* were coded as cognates in these instances, as were *swart* and *sweart*.

The concept BLOOD was represented by cognates throughout the Germanic languages. Based on the forms in the attested languages, the Proto-Germanic form**blodan* has been posited as the probable form of the ancestor of this word. Gothic has the word *bloþ* ([bloːθ]), while Old Saxon and Old English have *blod* ([bloːd]), and Old Frisian *blōd* (probably pronounced the same way as the Old English/Old Saxon word, with the macron merely being an orthographic convention). Old High German and Old Dutch have the form *bluot*, which differs from the other forms in having a diphthong rather than a pure vowel as the nucleus of the word; this is due to a regular process by which Proto-Germanic word internal **o* was diphthongised at an early date, giving the diphthong [uo] or [ʊo]. This diphthong was later simplified to [uː], giving the Modern High German form *Blut*, [bluːt], and Modern Dutch *bloed*, [blut]. In Old Norse there is the word *blóð* ([bloːð]), where the voiced interdental fricative following the vowel occurs systematically throughout the language and corresponds with a voiced alveolar stop in Old High German, Old Saxon, Old English and Old Frisian, and a voiceless interdental fricative in Gothic. This correspondence occurs in other words in the list, namely those for RED, GOOD and HEAD; because of this, and the fact that these forms are attested throughout the older Germanic languages (making it somewhat unlikely that they originated in one language and were subsequently borrowed by all of the others, leaving no trace of any forms from before the borrowing), all of these forms were coded as cognates in the Old Germanic languages (when it came to the modern languages, HEAD and GOOD presented a slightly more complicated picture; this will be discussed further on). For BLOOD, all of the modern Germanic languages had words descended from those in the older Germanic languages; these were all coded as cognates in all conditions.

All of the old Germanic languages, with the exception of Gothic, have a word for BONE which has the general form /bVn/: in Old English there is *bān*, in Old Saxon, Old Dutch and Old Frisian there is *bēn*, and in Old High German and Old Norse *bein*. Because of this similar structure, in the semantically strict condition, these were coded as cognates, as were any words in the modern Germanic languages with this structure. However, Old High German also has the word *knohha*, with the same meaning as *bein*; a number of the modern Germanic languages instead use a descendent of the word *knohha* as their word for bone, for example: Ferring Frisian uses *knook*, Saterland Frisian *Knoke*, Frasch Frisian *knooke*, Low German *Knocken*, Modern Standard High German *Knochen*, Bavarian *Knocha*, Swiss German *Chnochä*, and Danish *knogle*. Under the semantically strict condition these were all coded as cognates; the inclusion of *knohha* meant that an additional simulation was run, with this word

used instead of *bein*. Additionally, Modern Standard Dutch uses the word *bot* to refer specifically to a bone, while *been* can also mean "leg"; if one goes by the criterion that the most common word used for a given concept in a language is the one that is to be used in the Swadesh list, then *bot* has to be used. However, *been* did once mean "bone", is descended from *bēn*, and is still used in Dutch; due to this, it seems somewhat artificial to exclude *been* completely from the linguistic analysis, and there is the possibility that this strategy could over-exaggerate the lexical differences between the languages concerned; thus, the word *been* was included in the semantically lax simulation for Dutch. The same problem is encountered in a number of the varieties of Modern High German, where the word *Bein* has generally come to mean "leg"; because of this, *Bein* was used instead of *knohha* forms in the semantically lax condition. Modern English *knuckle* is cognate with *knohha*; this was also included in a semantically lax simulation. The reconstructed Proto-Germanic form for BONE given by Orel (2003) is **bainan*. When this was used, it was coded as a cognate of the other /bVn/ forms.

All of the languages analysed had a word for the concept BREAST/CHEST which began with a [b] and had a voiceless alveolar fricative present. However, five of them had a word in which the [b] was not followed by trill of some kind, but a vowel and then a trill, for example Old Frisian *–borst*, West Frisian *boarst*, Frasch Frisian *burst*, Standard Dutch and Flemish *borst*, and Afrikaans *bors*; in all of these instances, the lack of a /br/ cluster at the beginning of the word (which is found in the forms from all of the other Germanic languages) can be explained in terms of metathesis, whereby one phone in the word is simply shifted, often to simplify a consonant cluster. In all cases this has been known to happen, and it is less likely that a different word for BREAST/CHEST has undergone changes that have made it look similar to the /br/ forms than it is that a single process of metathesis has taken place. Therefore these forms were still treated as cognates. However, a problem which does arise is the fact that in Modern English the word *breast* has the near synonym *chest*, which has taken on the more general and neutral meaning, with *breast* either referring to a female mammary gland, or, when used to refer more neutrally to the thoracic region, simply being viewed as slightly archaic. In the semantically strict condition the form *chest* was thus used, and in the semantically lax condition, *breast*. A more complicated problem, which in this instance did not cause major problems but theoretically could be a source of confusion, is the fact that the Old Frisian form *–borst* exists alongside *brust*. As this study is only looking at lexical cognates, however, the potential problems this, and the differences in the vowel, throw up are largely ignored as both forms are still cognates from an overall point of view.

Like the BREAST/CHEST case above, all of the languages examined had cognate words for the concept TO BURN (used intransitively). In all of the languages, the words used for this generally had the form [b(*r*)V(*r*)n], where the brackets indicate that an *r*-segment (or something descended from this *r*-segment[12]) occurs in either of those places. Again, like the BREAST example, this is due to metathesis of this *r*-segment: it has moved away from the syllable onset and become part of the coda. The reasons for this are not always clear, as is

[12] Non-rhotic varieties of Modern English, such as Received Pronunciation, have undergone compensatory vowel lengthening due to the loss of this rhotic segment in syllable coda position.

evident from the fact that this appears to happen sometimes in one word but not another where it would be expected (compare Old English *breost*, from Proto-Germanic **breustan*, and Old English *beornan*, from Proto-Germanic **brinnanan*). However, the overall structure of all of the words included in the analysis for TO BURN is similar enough that they are more than likely descended from an ancestral form (in this case ultimately **brinnanan*) and were thus coded as cognates under all conditions.

The character CLAW had three sets of cognate forms across the languages, as well as not being attested in one language. A widespread form had the structure /klV(w)(V)/, as in old English *clawu*, Old Saxon *clâuua*, Old High German, Old Dutch and Old Frisian *klāwa*, Modern English *claw*, West Frisian *klau*, Saterland Frisian *Klaue*, Frasch Frisian *klau*/*klaa*, Standard Dutch and Flemish *klauw*, and Afrikaans *klou*. These descend from Proto-Germanic **klawâ-*~**klæwâ-* (OED, 2015) or possibly the verb **klawjanan* (Orel, 2003). These were coded as cognates. Despite having a similar core structure, namely /klV$_{[-low,+back]}$/, the North Germanic forms are not descended from the same Proto-Germanic words as the West Germanic words. The North Germanic forms, such as Old Norse *kló*, Danish *klo*, Faroese *klógv*, Icelandic *kló*, and Norwegian and Swedish *klo*, are believed to be descended from Proto-Germanic **klôh-*, related to the verb stem **klâ-* (OED, 2015); these were thus assigned their own codes. The third set of forms had the root structure /k*r*VC$_{[+vo,+alv,+obst]}$/, as in Ferring Frisian *kral*, Low German *Krall*, Standard High German *Kralle*, Bavarian *Kroin* (showing addition of the feature [+nasal] to the final [l]), and Swiss German *Kralle*. This word is attested in High German from the 16th century (Duden, 2015); the presence of this word in Ferring Frisian probably marks it as a loanword into Frisian; however, it appears to have replaced the original word for CLAW. These words were assigned the same codes.

For the concept CLOUD the two most common lexical items were words with a form along the lines of Old English *wolcen* and words with the same form (largely) as Old Norse *ský*. The former occurred in Old English, Old Saxon, Old High German, Old Dutch, and Old Frisian, as well as in all of their modern descendants, with the exception of Modern English, which uses the word *cloud*. In all of these cases, the cardinal stem form of the word is either a labiovelar glide (in the old Germanic varieties) or a labiodental fricative (in the modern varieties; this fricative is the result of a fortification process which the continental Germanic languages underwent, beginning in the early middle ages, in which /w/⟶ /v/ or /ʋ/), followed by a vowel with the value [+back], a lateral approximant [l] and a voiceless velar plosive (in the case of Old English this was palatalised because of the following front vowel, leading to the presence of the affricate [ʧ] where the other languages have [k]). It additionally seems likely that the /-Vn/ ending of the word is to be viewed as part of the stem, with this having been lost via syncope in the later forms of the language (the fact that forms lacking either /-Vn/ or /-n/ are found in Old High German and Old Dutch, in the former alongside the full form, suggests that this was a change which was already underway at an early stage of the attestation of a number of the Germanic languages). In the Bavarian dialect of Modern High German, the form differs somewhat in that there is no approximant between the stop and the vowel, a diphthong instead of a monophthong and a nasal at the end of the word, where some varieties have a vowel (Bavarian *Woikn*, [vɔɪkn], but Modern Standard High

German *Wolke*, [vɔlkə]); this is appears to be due to a set of sound changes which occurred in a variety ancestral to Bavarian, in which a word final unstressed vowel regularly became the nasal [n], and the lateral approximant [l] was realised as a high-vowel like segment when in the coda of a syllable, leading to diphthongisation of the vowel (compare also Bavarian *Kroin*, "claw", *geib* , "yellow", with Standard High German *Kralle* and *gelb*). Owing to these regularities, as well as the fact that there is no evidence of borrowing of this word between varieties, all of these forms were coded as cognates under all conditions. Interestingly, the form *scio* is attested with the meaning "cloud" in Old English, although it is not as common as *wolcen*; this is unlikely to be a borrowing from Old Norse, as it shows both the palatalization of velar stops before high front vowels characteristic of Old English ([sk]⟶ [ʃ]/_$V_{[+hi][+fro]}$) and the breaking of the West Saxon variety of Old English ([iː] ⟶ [io]).[13] This suggests that *scio* is a relic form which survived in Old English and Old Norse, but not in the continental West Germanic languages. *Scio* was substituted for *wolcen* in one round of simulation. Gothic has the form *milhma*, which is not attested anywhere else in Germanic, unless one regards Swedish *moln* as being cognate with it (although this is uncertain); these were coded as cognates in the study. Old Norse and its descendants (with the exception of Swedish) all have forms which begin with a voiceless alveolar fricative, followed by a voiceless velar plosive and a high vowel, such as Old Norse *ský*, [skyː], Danish and Norwegian *sky*, [sky], and Faroese *skýð*[ʃiː]. Icelandic has *skýj*, [skiː]. Based on the shared structures of these forms, and the fact that no instances of borrowing could be detected, they were coded as cognates under all conditions. Lastly, Modern English *cloud* is descended from Old English *clūd*, "rock, hill" (presumably the semantic shift was based on the shape of the clouds being likened to a hill) and does not have a synonym or near synonym that has survived from an older stage of the language. *Cloud* therefore stands on its own in the data. As *clūd* did not mean CLOUD in Old English, it was not included as an alternative to *wolcen* or *scio* in any simulations.

The concept COLD was represented by cognates throughout the Germanic languages. In all cases there are no unexpected forms or phonetic values in any of the languages. The general form of the word is the voiceless velar plosive [k] followed by a vowel which is usually [+back], followed by the liquid [l] and an alveolar plosive. The exceptions to this can all be explained by sound changes which have occurred independently in the languages: Old English has the word *ceald*,[ʧeald], with the affricate [ʧ] due to the palatalization of [k] before the diphthongised [a], which took on a high vowel quality at the beginning of the diphthong (Modern English has [k] because it is descended from an Anglian dialect which did not participate in this sound change as extensively as the West Saxon dialect, which is the most thoroughly attested variety of Old English (Barber, Beal & Shaw, 2012); Modern Dutch, Flemish and Afrikaans all have *koud*, with the consonant cluster [ld] after the vowel being simplified by the deletion of the liquid in Old Dutch (van der Wal & Quak, 1994); Bavarian's *koid* is due to a process whereby the liquid [l] was realised as a high vowel when in the coda of a syllable, resulting in a diphthong; Swiss German has *chalt*,[xalt], which

[13] Modern English *sky*, which means "sky, not "cloud", is a borrowing from Old Norse, as evinced by the [sk] cluster. Due to the fact that it is a borrowing, it was not included under this entry under any condition, despite its form.

appears to be because of a sound change which resulted in the stop [k] becoming a fricative when before a vowel or an approximant; Swedish *kall*, [kalː] is due to the deletion of the alveolar plosive and compensatory lengthening of the neighbouring liquid (SAOB, 2015). The Gothic *–s* and Old Norse *–r* at the end of the word are simply nominative case markers and are not part of the stem. As there is no evidence of borrowing, these were all coded as cognates under all conditions. Proto-Germanic **kaldaz* (Orel, 2003) is proposed based on this form.

TO COME was another concept which was represented by words which were all cognates across the languages. Most of the forms examined had either a voiceless velar plosive followed by a labiovelar glide or a velar plosive followed by a rounded vowel, then a bilabial nasal making up their stem, for example Gothic *qiman*, [kwiman], Old High German *kweman*, [kwɛman], Old English *cuman*, [kʊman], Old Norse *koma*, [koːma]. The difference between a rounded vowel (particularly something like [u]) and the labiovelar glide [w] is not great phonetically, and so this alternation between different languages is not so large as to justify coding the respective lexical items as being non-cognate; it is at least plausible that these alternations stem from reanalysis of an ancestral form which had either a glide or labialised velar plosive before the vowel, or [+round] vowel which was produced near the velum. The lack of a rounded vowel in Ferring Frisian could be result of deletion of the glide, followed by a shift in the vowel at some point after. However, the stem structure /kVm/ is common to all of the languages for the concept TO COME, leading to the decision to code them as lexical cognates in all conditions. Orel (2003) has proposed the Proto-Germanic **kwemanan* for this concept.

The concept TO DIE had a number of different forms across the languages concerned, often with synonyms or near synonyms in a given language. One word had the stem structure voiceless alveolar fricative-labiovelar glide-front vowel-voiceless alveolar stop, seen in Gothic *(ga)swiltan*, and Old English and Old Saxon *sweltan*. The high vowel of the Gothic item corresponds with a lower but still front vowel in a number of other Old English and Old Saxon words, for example Gothic *qiþan, giban, qino, rign* versus Old English *cweþan,* Old Saxon *kwethan, gevan,* Old English *cwene, ren*, Old Saxon *regin* (meaning, "to say/speak", "to give", "woman" and "rain", respectively). This correspondence suggests that borrowing is unlikely, as a borrowed item would more probably have the same vowel in the borrowing language as in the language being borrowed from, especially seeing as the vowels concerned occur in the phoneme inventories of all three languages ([i], [ɛ]) (Lehmann, 1994; van Kemenade, 1994) . Thus, it is probable that these are genuine inherited forms; because of this and the fact that they have the same stem structure, they were coded as cognates. However, Old English and Old Saxon had another word for TO DIE which, *steorfan* and *stervan*, respectively. These have the same stem structure as Old High German *sterban*, Old Dutch *stervan*, Old Frisian *sterva*, Ferring Frisian *sterev*, West Frisian *stjerre*, Saterland Frisian *stierve*, Frasch Frisian *stärwe*, Low German *schtarbm*, Standard Dutch and Flemish *sterven*, Afrikaans *sterf*, Standard High German *sterben*, Bavarian *schteam*, and Swiss German *schtäärbe*. All of these forms begin with a consonant cluster made up of either a voiceless alveolar fricative and a voiceless alveolar plosive, or a voiceless palatal fricative and a

voiceless alveolar plosive (in the case of the Modern Low and High German varieties) followed by a vowel, usually some form of rhotic consonant, and either a labial fricative or plosive (the reason this segment is not left out is that labiovelar fricatives and bilabial plosives are known in diachronic linguistics to frequently alternate when pronounced intervocalically (Campbell, 2004), which, coming before the Germanic verbal ending *–an*, it was).[14] Due to this, as well as the fact that the sound correspondences between them are regular (the High German forms with a plosive at the end of the stem, rather than a fricative, are due to the High German consonant shift, and would have originally been fricatives in Pre-Old High German; the diphthong in Old English is due to a process of diphthongisation and corresponds with a monophthong in many of the other old Germanic languages) and the fact that they have the same meaning, they were all coded as cognates in the semantically strict condition. Modern English complicates matters somewhat, however. The most common word for TO DIE in Modern English is *(to) die*; this was borrowed from Old Norse *deyja* when the Danelaw was still in existence and much Anglo-Saxon territory came to be controlled and settled by Danes. The Old Norse word is also ancestral to all of the forms seen in the modern North Germanic languages, for example Standard Danish *dø*, Faroese *doyggja*, Icelandic *deyja*, Norwegian Nynorsk *døy*, Norwegian Bokmål *dø*, and Swedish *dö*. These all have the same stem structure, /dV/ and have the same meaning. Thus under the semantically strict condition all of these terms must be coded as cognates, including the English one. However, Modern English still has the word *starve*, which is descended from *steorfan* but has undergone a slight semantic shift, from "to perish, die" to "to lack food, die due to lack of food". Thus, under the semantically lax condition, *starve* was used instead of *die*, as it carries on an older form despite having a different meaning, and was coded as a cognate of the *stervan* forms. Two words for TO DIE in Proto-Germanic are reconstructed by Orel (2003), **sweltanan* and **sterbanan*; these were coded as cognates of the *sweltan* and *stervan* forms, respectively, in two separate simulations.

Item eighteen on the Swadesh 100-word list, DOG, had several attested forms; however, one form in particular was found in all of the Germanic languages, and also turned out to be the most common form used in all of them except one. This word had the stem form /hV$_{[+ro]}$nC$_{[+alv][+stp]}$/. Thus in Gothic there is the word *hunds* (with *–s* simply the masculine nominative singular marker), in Old English, Old Saxon and Old Frisian *hund*, in Old High German and Old Dutch *hunt*, and in Old Norse *hundr* (the *–r* morpheme is the masculine nominative singular marker). This form is continued in the modern languages, for example Ferring Frisian *hünj*, Frasch Frisian *hün*, West Frisian *hûn*, Saterland Frisian *Huund*, High and Low German *Hund*, Standard Dutch, Flemish and Afrikaans *hond*, Icelandic and Faroese *hundur*, and Danish, Norwegian and Swedish *hund*. Each conforms to the same stem structure very highly, with variation mainly restricted to whether the alveolar plosive at the end of the stem is voiced or not and the quality of the stem vowel. The Frisian varieties which lack an alveolar stop after the nasal formerly had one, but lost it sometime in the Middle or Early Modern Frisian periods (usually taken to be from c. 1550-c.1820; Bremmer, 2009). Thus these forms were all coded as lexical cognates. Old High German also has the word

[14] Compare Latin *Tiber*, [tiːbɛr], with Italian *Tiverre*, [tivɛrːe], Latin *taberna*, [tabɛrna], Italian *taverna*, [tavɛrna], "tavern".

rudo attested with the meaning DOG; this word is not attested in any of the other Germanic languages and so in this analysis must be treated as an innovation. This is obviously not a cognate of *hund* and so was assigned a different code in a second simulation. Modern English is the odd one out in the data set, with the word *dog* being the more neutral, common word for DOG. The word *dog* is descended from Old English *docga*/*dogga*, which had a more specific meaning than it does today (it is generally thought to have denoted a particular type of hunting dog). *Dog* is clearly a very unlikely cognate of *hund*, and was therefore assigned a different code in the semantically strict condition. However, Modern English has the word *hound*, which is a descendant of *hund*; its meaning is more specific than *dog*, often implying a dog used in hunting or sport itself, much like Old English *docga*. Thus, under the semantically lax condition, both of these words were substituted for the words used in the semantically strict condition and coded accordingly. The reconstructed Proto-Germanic word for DOG is **hundaz* (Orel, 2003).

The concept TO DRINK was represented by cognates across all of the Germanic languages, all having a stem with the form $C_{[+alv][+plo]}rV_{[+fro]}\,CC_{[+vel][+plo]}$. All have some *r*-like sound in second position of the stem, such as a trill (either uvular, velar, or alveolar) in the case of all of the continental Germanic languages, as well as Icelandic and Faroese, or a liquid [ɹ] in the case of most standard varieties of English, such as Received Pronunciation or General American (Barber, Beal & Shaw, 2012; Crystal, 2007). The consonant given above as simply C takes one of two different forms. In Old Norse and its descendants, the North Germanic languages, it is simply part of the velar stop which follows the vowel (for example, Old Norse *drekka*, Danish *drikke*, Swedish *dricka*). This resulted from a process in Pre-Old Norse in which the nasal in nasal-stop clusters (such as [ŋk]) assimilated to the following plosive entirely, resulting in a geminated stop where the other Germanic languages have a nasal-stop cluster (Fortson, 2004; Faarlund, 1994). As these words are all cognates, they were assigned the same code under all conditions. Orel's (2003) reconstructed form is **drenkanan*, which was given the same code in the Proto-Germanic simulation.

Two forms are attested for the concept DRY. One group conforms to the stem structure $C_{[+alv,+plo]}rV(C_{[+vel]})$. This set of forms is found mainly in the languages traditionally classified as West Germanic, for example, Old English *dryge*, Old Saxon *drokno*, Old High German *trokken*, Modern English *dry* (missing the stem final consonant), West Frisian *droech*, Dutch *droog*, Low German *dröch*, and Modern High German *trocken*. As can be seen from comparing the Modern English word with its Old English equivalent, the lack of a velar consonant is due simply to the intial lenition then loss of this segment in the Middle English period (c.1150-c.1450) (OED, 2015; Baber, Beal & Shaw, 2012). As there is no evidence of borrowing between these languages regarding this word, these words were coded as cognates under the lexically strict condition. The second form from the data had the stem structure $C_{[+alv]}Vr$, for example Gothic *þaursus*, Old Norse *þurr*, Faroese *turrur*, Danish *tør*, Norwegian *tørr*, and Swedish *torr*. Additionally, the word *durri* is attested in Old High German alongside *trokken*; both appear to mean DRY, with no particular difference in meaning between them. These forms were assigned a different code from the *trokken* forms and were all coded as being cognate with each other. In the case of Old High German *durri*,

the chances of this being borrowed from Old Norse were considered slim, as there are no similar forms in Old Saxon, which lay geographically in between Old High German and Old Norse speaking areas; there is also the vowel at the end, which in light of the lack of a vowel in Old Norse makes this explanation somewhat problematic. It seems more likely that this is a relic form which has been retained in Old High German and Old Norse, but lost (or is at least unattested) in Old Saxon, Old English and Old Frisian. Thus, two simulations were performed, with one of each of the Old High German forms in each.

The concept EAR was represented by cognates across all of the languages concerned. All of the words for EAR had the core structure /V*r*(V)/, where the bracketing of the vowel indicates that the word final vowel was not found in all of the languages, but was still widespread. Old English has the word *eare*, Old High German and Old Dutch have *ōra,* Old Saxon *ora*, Old Frisian *āre*, and Old Norse *eyra*. The words descended from the old West Germanic varieties (all of those listed above except for Old Norse) have all (with the exception of Bavarian) lost the vowel found at the end of the older forms via syncope. Thus Modern English has *ear*, Modern Standard High German *Ohr*, Swiss German *Oor*, Low German *Ohr*, Dutch, Afrikaans and Flemish *oor*, Ferring Frisian *uar*, West Frisian *ear*, Saterland Frisian *Oor*, and Frasch Frisian *uur*. Bavarian has the form *Ooa*, [oːɐ], although it is uncertain if this is due to the deletion of the intervocalic trill [r] in *ōra*, with the result being vowel hiatus, or the deletion of the second vowel of *ōra* followed by the elision of the trill first to a liquid and then a vowel. The likelihood is greater that the first explanation is the better of the two, as it requires only one change to get the modern form (without evidence to the contrary, the simpler explanation is often the better). However, Bavarian does show a tendency to diphthongise vowels before certain consonants, often with deletion of the consonant in question, as can be seen in correspondences such as Standard High German *wer* versus Bavarian *wea*, "who", Standard High German *Herz* versus Bavarian *Heaz*, "heart". Nevertheless, it can clearly be seen that the Bavarian word for EAR is a cognate of the others. Across all of the descendants of Old Norse the older disyllabic structure of the word for EAR has been preserved: Danish has *øre*, Faroese *oyra*, Icelandic *eyra*, Norwegian Bokmål *øre*, Norwegian Nynorsk *øyra*, and Swedish *öra*. These are all cognates, both with each other and with the West Germanic forms, and were coded as such. Gothic presented a form somewhat different in that where the other Germanic languages have some form of rhotic element as the consonant in the word, Gothic has a fricative, *auso*, [aʊso]. This correspondence can be seen in a number of Gothic words, such as *hausjan*, "to hear", with [s], versus Old English *hieran*, Old Saxon *horian*, Old High German *hōren*, and Old Norse *heyra*. This is due to a process of rhoticisation which affected the descendant of inter- and post-vocalic Proto-Germanic *z in all of the well-attested old Germanic languages except for Gothic (Nielsen, 1989). Based on this, it is clear that Gothic *auso* should be coded as a cognate of the other terms. Orel (2003) gives **auzon* as the reconstructed Proto-Germanic word for EAR.

Character twenty-two on the Swadesh 100-word list is EARTH/SOIL. This was a concept which was problematic due to the fact that a number of words in the Germanic language can be used to refer, with varying slight differences in semantics, to earth or soil. The approach taken here was that the most commonly used or neutral words should be used if possible (in

line with the general requirements laid out by Swadesh for inclusion in his list). In the semantically strict condition, this resulted in three principal forms being used. The first of these had the core structure /V(*r*)C$_{[+stp/+fric]}$ (V)/, where the segments in parentheses are found in some, but not all, of the languages which have this particular form. This structure is represented by Old English *eorþe*, Gothic *airþa*, Old Saxon and Old Dutch *ertha*, Old Frisian *erthe*, and Old High German *erda*. These have been continued as Modern English *earth*, Low German *Eer*, Modern Dutch and Afrikaans *aarde*, Flemish *aerde* (the difference in spelling is largely orthographic convention), Ferring Frisian *eerd*, West Frisian *ierde*, Frasch Frisian *jard*, Standard High German and Swiss German *Erde*, and Bavarian *Erdn*. The second main form occurred in Old Norse and two of its descendants, namely Icelandic and Faroese. This form was *mold*. This was assigned a separate code from the above forms. Lastly, Saterland Frisian had the form *Gruunde*. This appears to have undergone a semantic shift, from "ground, something on which things stand" to "earth, soil". This was assigned its own code. Under the semantically lax condition, things became more complicated. Modern English *soil* has largely the same meaning as *earth*, being only somewhat more restricted in what it refers to; this was therefore substituted in place of *earth*.

The item TO EAT was represented by five different forms in the languages examined. The first, and most widespread, had the core stem structure /V$_{[+fro]}$C$_{[+obst,+alv]}$(V)/. This form can be seen in Old English, Old Saxon and Old Dutch *etan*, Old Norse *eta*, Old High German *ezzan*, Old Frisian *iten*/*itan*/*ita*, Modern English *(to) eat* (where <ea> represents [iː] and the final vowel has been lost), West Frisian *ite*, Saterland Frisian *iete*, Frasch Frisian *ääse*, Low German *edden*, Modern Dutch and Flemish *eten*, Afrikaans *eet* (also showing syncope of the final vowel), Modern Standard High German *essen*, Bavarian *essn* (displaying syncope of the final vowel, with the final nasal being an infinitive marker; this has a parallel in Afrikaans *sien*, "see" and *gaan*, "go", where the final nasal in each word is derived from the Middle Dutch infinitive marker), Swiss German *ässe*, Faroese *eta* and Norwegian Nynorsk *eta*, and Swedish *äta*. Old High German has a fricative [zː] where the other varieties of Old Germanic (except for Gothic) have an alveolar plosive [t]; this is simply due to the effects of the High German Consonant Shift, which caused a large number of the West Germanic voiceless stops to become either fricatives or affricates, depending on where they fell within the word (often stops at the beginning of a word became affricates, while those occurring word-internally became simple fricatives; more on this below); the apparent change in the consonant's length is not clear, however. This has been continued in Old High German's modern descendants (here Standard High German, Bavarian and Swiss German), which all have [s]. Frasch Frisian has *ääse*, [eːsə], which also shows the shift from a stop to a fricative; this is thought unlikely to be a borrowing from a High German variety as it is hard to reconcile the differences in vowel length between the two (High German has a short vowel where Frasch Frisian has a long one) and for most of its history the Frasch Frisian area has been bordered by Low German areas, with High German influences likely to have only appeared later, after the mediaeval period (from roughly the 1500s).[15] Additionally, Frasch Frisian shows this

[15] It is probable that even after this most influence on Frasch Frisian would still have been from Low German, with any major influence from High German sources mainly coming about in the 1800s, especially after the unification of the German states and the establishment of a High German variety as the national standard.

placement of a fricative where a stop would be expected in a number of other words, but sometimes has a fricative where High German has a stop (for example, HG *Blatt* versus FF *blees*, "leaf") or a stop where High German has a fricative (for example, HG *Wasser* versus FF *wååder*, "water"); this suggests that Frasch Frisian fricative is the result of a separate sound change which affected the language at a later period, although its conditioning environment is not completely clear. These forms were all coded as cognates under all conditions. Gothic had the form *matjan*; as a verb this was not attested anywhere else. In Swedish, however, the word for food is *mat* (which is related to English *meat*), and it is possible that the two forms are related, with a semantic shift having resulted in the stem /mat-/ being carried on in one as a verb and the other as a noun. Under the semantically strict condition Gothic was thus coded as the only language with the stem /mat-/ in the verb TO EAT. Ferring Frisian had the word *litj* for TO EAT. If the liquid [l] is not considered, the remaining *–itj* certainly seems probable as a descendant of Old Frisian *ite*, especially as Ferring Frisian has a tendency to have a palatalised consonant where other Frisian varieties have a consonant-vowel combination at the end of the word (for example, West Frisian *bite* [biːtə] versus Ferring *bitj*, [bitʲ], "to bite"). The epenthesis of the liquid [l] before the first vowel does however seem an unusual change; it was decided to code this word as a cognate of the *etan*/*itan* forms purely on the basis of the structure of the rest of the word, treating the liquid as simply an unusual example of epenthesis. This was somewhat problematic, as it is difficult to gauge the relative probabilities that three features will be shared across several languages against the addition of a sound in an unusual position. Danish and Norwegian Bokmål had the word *spise* for TO EAT; this word is a loanword from Middle Low German, and has become the normal, neutral word meaning TO EAT; these were coded as cognates of each other. Both languages still have words derived from Old Norse *eta*, these being Danish *æde* and Norwegian *ete*; despite being used less often than *spise*, there does not appear to be a major semantic difference between them; these were therefore used in an additional Semantically Strict simulation. These were coded as cognates of each other, but clearly form a group different from those above. Lastly Icelandic had the word *borða*; this is not attested as a verb in any of the other Germanic languages, and thus stood on its own under the Semantically Strict condition.

The next character in the list was the noun EGG. In Gothic this took the form of the word *ada*; this was attested nowhere else and was thus assigned its own code. The North and West Germanic languages all had forms which, according to Orel (2003) are descendants of Proto-Germanic **ajjaz*. These could themselves be divided up into two forms, one of which ends in a voiced velar plosive [g], and one which lacks the velar plosive. The first form is found in Old Norse, Icelandic, Faroese, Norwegian and Modern English *egg*, Danish *æg* (although the process which has led to the phenomenon of *stød* in Danish has resulted in the velar [g] coming to be pronounced as a glottal stop) and Swedish *ägg*. The second of the sub-forms, which lacks the word-final stop, can be seen in Old English *æg*, [æj], Old Saxon and Old High German *ei*, Ferring Frisian *ai*, West Frisian *aai*, Saterland Frisian *Oai/Ai*, Frasch Frisian *oi*, Low German and Standard High German *Ei*, Standard Dutch and Flemish *ei*, Afrikaans *ei/eier*, Swiss German *Ai* and Bavarian *Oa*. Despite the existence of these two groups, they are all cognates, with the forms with the velar stop simply having undergone

fortification of the Proto-Germanic geminated glide **jj* (Proto-Germanic **jj* became Old Norse [gː]. Modern English *egg* is a borrowing from Old Norse, but there is no surviving descendant of Old English *æg* in Modern English, and so no alternate form was included under the concept EGG in the semantically lax condition. These were thus all coded as cognates. A word for EGG appears to not be attested in Old Dutch; while it is probable, based on the Modern Dutch, Flemish and Afrikaans forms, that Old Dutch had something along the lines of *ei*, because it is unattested it is necessary to follow the strategies mentioned above for handling missing information. Thus a simulation was run where the entire concept of EGG was removed, one where the ambiguous setting was used to assign the most likely form, and one where the missing information was treated as a deletion.

The concept EYE was represented by cognates across all of the Germanic languages. Gothic has *augo*, Old English *eage*, Old High German *ouga*, Old Saxon *oga*, Old Dutch *ōga*, Old Frisian *āge*, and Old Norse *auga*. In all of the modern descendants of these languages (with the exception of Gothic, which has no descendants) these words have been carried on in one form or another, without convincing evidence of borrowing. They all conform to the general structure /VC$_{[\ \{DescPGmc\ *g\}]}$(V)/[16], where the consonant is a descendant of Proto-Germanic **g*. In Old English, this **g* came to be palatalised before a front vowel, resulting in [j]; the final vowel of *eage*, [e] or [ɛ] (this was later reduced to a schwa), is the result of a process of raising low vowels which occurred in the Anglo-Frisian languages, hence Old English *eage*, Old Frisian *āge*, but Old High German *ouga*, Old Dutch *ōga*, and Old Saxon *oga* (Barber, Beal & Shaw, 2012). This resulted in Proto-Germanic **g*, [g], becoming [j] in Old English. A similar process is thought to have occurred in Old Frisian (Fortson, 2004). The Modern English word *eye*, [aɪ], is descended from this or a closely related form, with syncope of the final vowel and a change in the quality of the diphthong. A similar change seems to have affected Danish and Norwegian Bokmål, where the final vowel of Old Norse *auga*, [a], appears to have been raised to [e] or [ɛ] before being reduced to a schwa, resulting in the voiced velar stop being realised as a palatal glide [j], giving *øje* in Danish and *øye* in Norwegian Bokmål. The fricative of Standard Dutch, Flemish and Afrikaans *oog* ([ɣ] in the first two, [x] in the Afrikaans) descends from Old Dutch [ɣ], which had come to replace [g] in all environments before its first attestation (De Schutter, 1994). Despite the differences in sounds, these forms are all cognates, and as such were given the same code in all conditions. Orel (2003) reconstructs the word **augon* in Proto-Germanic.

The next concept on the Swadesh 100-list was FAT (the substance). One of the sets of cognate words had the structure /smV*r*/, where *r* represents a rhotic segment, as in Gothic *smairþr*, Old English *smeoru*, Old Saxon, Old High German and Old Dutch *smero*, Old Frisian *smere*, Old Norse *smjǫr* and Frasch Frisian *smeer*. These are descended from Proto-Germanic **smerwa* (OED, 2015) or **smerwon* (Orel, 2003). These were coded as cognates. In all of the modern Germanic languages used in this study, with the exception of Frasch Frisian, the default word for FAT has the form /fVt/, as in Modern English *fat*, Ferring Frisian *feet*, West Frisian *fet*, Saterland Frisian *Fat*, Low German *Fett*, Standard Dutch, Flemish and Afrikaans *vet*, Standard High German, Bavarian and Swiss German *Fett*, Danish

[16] C$_{[\ \{DescPGmc\ *g\}]}$ refers to a consonant which is a reflex of Proto-Germanic **g*.

fedt, Faroese *feitt*, Icelandic *fita*, Norwegian Nynorsk *feitt* and Norwegian Bokmål and Swedish *fett*. In these cases the original word seems to have been an adjective referring to the bulk of something, and only later took on the meaning FAT (the substance) (OED, 2015; SAOB, 2015; Duden, 2015); the OED (2015) indicates that this use in English is first attested in 1539, and the SAOB (2015) indicates this use from 1537. This suggests that this semantic development probably happened in parallel; nevertheless, these words are all cognates, all originally descending from Proto-Germanic **faito-* "fat" (adj) or **faitido-*, the past participle of the Proto-Germanic verb **faitjan*, "to fatten' (OED, 2015). These were thus assigned the same codes. In the modern North Germanic languages words descended from Old Norse *smjǫr* have been continued with the meaning "butter", such as Danish *smør*, Norwegian *smør*, Icelandic *smjer* and Swedish *smör*; these were included in a semantically lax condition.

The concept FEATHER was represented predominantly by words with the form /fV$_{[+fro]}$(C$_{[+intdent/alv]}$)V(*r*)/, as in Old English *feþer*, Old Saxon and Old Dutch *fethera*, Old High German *fedara*, Old Frisian *fethere*, Old Norse *fjǫðr*, Modern English *feather*, Ferring Frisian *feeler*, West Frisian *fear*, Frasch Frisian *fääder*, Low German *Feller*, Standard Dutch and Flemish *veer*, Afrikaans *veer*, Standard High German *Feder*, Bavarian *Feeda*, Swiss German *Fäädre*, Danish *fjer*, Faroese *fjøður*, Icelandic *fjöður*, Norwegian Nynorsk *fjør*, Norwegian Bokmål *fjær*, and Swedish *fjäder*. These are all descended from a word in Proto-Germanic reconstructed as **feþrâ* (OED, 2015) or **feþrō* (Orel, 2003); all of these words are related to each other lineally and were thus assigned the same code. Saterland Frisian *Fugge* is of an unknown etymology, but is definitely thought not to be a cognate of the above terms (Hoekstra, personal correspondence); it was therefore assigned its own code. A Gothic word for FEATHER could not be found, and so this missing data item was treated as either a deletion (Infodelete) or an ambiguous item (which under Majority Wins would be assigned to the most common form in the languages where the word is attested in the Infoambig condition); in a third simulation the character FEATHER was removed (Infoclean).

The concept FIRE (noun) had a number of different forms across the Germanic languages. A widespread form had the core structure /fV*r*/, where *r* represents a rhotic segment. This form can be seen in Old English *fyr*, Old Saxon and Old High German *fiur*, Old Dutch *fuir*, Old Frisian *fiūr*, West Frisian *fjoer* (with an epenthetic glide between the labiovelar fricative and the vowel), Saterland Frisian *Fjúur*, Low German *Füer*, Modern Dutch and Afrikaans *vuur*, Flemish *vuer*, Modern Standard High German *Feuer*, Swiss German *Füür*, Modern English *fire* (where loss of rhotics in the syllable coda position in Received Pronunciation has resulted in the pronunciation [faɪə], although in a number of other varieties the rhotic is still present in coda position (Barber, Beal and Shaw, 2012)) and Bavarian *Faia* (where the rhotic has also been lost in many instances in syllable coda position; compare *Feeda*, "feather", and *Ooa*, "ear", above). As there is no evidence that there has been borrowing involved, or that another form has been replaced by one of these, they were all coded as cognates. The other widespread form under the concept FIRE has the core structure /V$_{[+fro]}$l(d)/, where the voiced alveolar plosive is bracketed because it occurs in most, but not all, forms. This structure can be seen in Old Norse *eldr*, Icelandic and Faroese *eldur*, Danish *ild*, Swedish and Norwegian Nynorsk *eld*, Norwegian Bokmål *ild*, Ferring Frisian *ial* and Frasch Frisian *iilj*. It is likely

that the two varieties of Frisian have borrowed this word from either Old Norse or one of its descendants, as the *eld* form is not attested in any of the old West Germanic languages, and the areas in which these varieties are spoken are both in close proximity to Danish speaking areas, or areas through which the Danes traded. In fact, the missing plosive in the two Frisian varieties may even point to Danish, with the loss of the word-final plosive in Danish being reflected by them. However, as no alternative form is attested in these two varieties, the coding of these words as cognates under all conditions was adhered to. In addition to the word *eldr*, Old Norse had the poetic word for fire *funi*, which is a cognate of Gothic *fon.* It seems unlikely to be a borrowing, as Old Norse was not required to maintain CV syllable structure, which would bring into the question the reason for Old Norse speakers' epenthesising a vowel to the end of the word. Gothic *fon* is also unlikely to be related to the *fyr* forms, as rhotacisation in North and West Germanic only affected sibilants (Fortson, 2004). Thus in the semantically strict condition Gothic *fon* was coded separately from the other forms, and under the semantically lax condition Old Norse *funi* was used as its cognate.

Item twenty-nine on the Swadesh 100-word list was FISH. This was represented by cognates across all of the languages. The core structure of all of these was /fV$_{[+fro,+hi]}$C$_{[+sib]}$(k)/, where the voiceless velar plosive in brackets indicates that this sound is widespread but not universal among the forms attested. This can be seen in Gothic *fisks*, Old English *fisc* (where <sc> represents a [ʃ]), Old Saxon, Old High German, Old Dutch and Old Frisian *fisk*, and Old Norse *fiskr*. These have been carried on in the descendants of these languages (except Gothic, which has no descendants), such as Modern English *fish*, Icelandic *fiskur*, and Modern Standard High German *Fisch* (the [sk] cluster of Old High German was replaced in the Middle High German period by [ʃ]; this change affected other clusters involving [s] and another consonant, although to different degrees, as can be seen by the lack of [sk] clusters in native High German words, but alternations such as *sterben*, [ʃtɛʀbn̩] and *ist*, [ɪst]). These were therefore coded as cognates under all conditions. Orel (2003) reconstructs **fiskaz* for FISH.

The concept TO FLY was largely represented by the same word across all of the languages examined. The core structure of this stem was /flV(C)/. This can be seen in Old English *fleogan*, Old High German *fliogan*, and Old Dutch *fliegan* (where the *–an* is simply an infinitive marker), Old Frisian *fliāga*, Old Norse *fljúga*, Modern English *to fly*, Ferring Frisian *flä*, West Frisian *fleane*, Frasch Frisian *fliinj*, Low German *fleegn*, Standard Dutch and Flemish *vliegen*, Afrikaans *vlieg*, Modern Standard High German *fliegen*, Bavarian *fliang*, Swiss German *fliege*, Standard Danish *flyve*, Faroese *flúgva*, Icelandic *fljúga*, Norwegian Nynorsk *flyga*, Norwegian Bokmål *fly*, and Swedish *flyga*. As there is no clear evidence for borrowing here, and all of these words have the same form, it seems clear that they are cognates; as such they were assigned the same code.

The next concept was FOOT. This was represented by clear cognates across all of the languages concerned, with all words having the stem structure /fV$_{[+ro]}$C$_{[+alv]}$/. Gothic has *fotus* (where the *–us* appears to be a continuant of Proto-Germanic **-z* or **-uz* as a masculine nominative marker), Old English, Old Frisian and Old Saxon *fōt*, Old High German *fuoz*, Old Dutch *fuot*, Old Norse *fótr* (where the *–r* is, as in Gothic, a continuant of Proto-Germanic **-z*,

only in this case having undergone rhotacisation), Modern English *foot*, Ferring Frisian *fut*, West Frisian *foet*, Saterland Frisian *Fout*, Frasch Frisian *fötj*, Low German *Foot*, Standard Dutch, Flemish and Afrikaans *voet*, Standard Modern High German *Fuß*, Bavarian *Fuass*, Swiss German *Fuess*, Standard Danish *fod*, Icelandic and Faroese *fótur*, and Norwegian and Swedish *fot*. Orel (2003) reconstructs **fotz~*fotuz*; this reconstruction indicates some uncertainty concerning the origin of the second vowel in Gothic. In Icelandic and Faroese this vowel is completely unrelated to the vowel in Gothic, and is due to a process of vowel epenthesis which took place in both languages around the 1500s, in which a number of syllable final obstruent-trill clusters (including ones caused by the addition of the masculine nominative marker) became unacceptable coda structures, and so had the vowel [u] (later [y]) inserted between the consonants (Faarlund, 1994); compare Old Norse *sandr* with Icelandic *sandur*, "sand", and Old Norse *hundr* with Icelandic and Faroese *hundur*. The alveolar fricatives seen in the High German varieties (Old High German, Modern Standard High German, Bavarian and Swiss German) are due to the High German Consonant shift, which resulted in the changing of the manner of a large number of consonants; in this, many of the stops in word final position became voiceless fricatives. The Frasch Frisian form, *fötj*, which has [tʲ] as its final consonant, is one of many words in this variety that show palatalization of the final consonant in the word; this appears to have been a largely random process, as many of the words in which this has occurred do not appear to have any conditioning factor. As these are all cognates, and no borrowing could be discerned, they were assigned the same code.

Like FOOT, the concept FULL was represented by similar forms across all of the languages. The stem structure of these forms was /fV(l)/, where the brackets placed around the lateral approximant indicate that it is widespread amongst the forms, but is not found in all of them. However, in the forms where it does not occur its absence can be explained by changes which have removed it from an ancestral form which did have the approximant. Thus in Bavarian, which has *foi*, the absence of the liquid can be explained by the fact that there has been a strong tendency in Bavarian to change [Vl] clusters into the dipthong [oɪ]; this can be seen when one compares Bavarian to Standard High German, with Bavarian *foi* versus Standard High German *voll*, Bavarian *koid* versus Standard *kalt*. In Icelandic and Faroese, the [lː] of Old Norse became [tl] or [dl] sometime in the 1600s, giving *fullur*, where the <ll> is pronounced as one of the above plosive-liquid clusters. The other languages have preserved this liquid, as can be seen in Gothic *fuls* ([fʊls]), Old English *full* ([fʊlː]), Old Saxon *ful* ([fʊl]) Old High German and Old Dutch *fol* ([fol]), Old Norse *fullr* ([fulːr]), and all of their descendants (with the exception of Gothic) except for those given above. The reconstructed form of this word given in Orel's (2003) dictionary of Proto-Germanic is **fullaz*. These were all coded as cognates.

The next item on the Swadesh 100-word list is TO GIVE. This was represented by five different words in the data. The most widely attested forms of the word for this concept have the core structure $C_{[+vel/pal]}V(C_{[+lab]})$, where the element in parentheses indicates that this is not found in the stems in all of the languages concerned; it is however common enough to warrant its inclusion. This stem structure can be seen in Gothic *giban*, Old High German

geban, Old Saxon and Old Dutch *gevan*, Old Frisian *jeva*/*jān*, Old Norse *gefa*, Old English *gifan*, Modern English (*to*) *give*, West Frisian *jaan*, Low German *gebm*, Standard Dutch and Flemish *geven*, Afrikaans *gee*, Standard High German *geben*, Bavarian *gebm*, Swiss German *gee*, Danish *give*, Faroese *geva*, Icelandic *gefa*, Norwegian Nynorsk *gje*, Norwegian Bokmål *gi*, and Swedish *giva*. The reconstructed Proto-Germanic form is **gebanan* (Orel, 2003). The alternation in the stem-initial consonant between [+velar] and[+palatal] is due to the fact that in some of the languages the word has a palatal glide [j], while others have either a voiced velar plosive [g] (e.g. Modern English *give* [gɪv], German *geben* [geːbən]) or a fricative [x] (Old Dutch *gevan* [ɣeːvan]), for example Old English *gifan*, [jivan], Old Frisian *jeva/jān*, [jevə/jaːn], Swedish *giva*, [jiːva]. This is simply due to a number of palatalization processes which affected Old English and Old Frisian, and, separately and at a later date, Swedish and Norwegian; in both cases velar plosives before front vowels were palatalized (Fortson, 2004; van Kemenade, 1994; Lass, 1997). Modern English *give* is borrowed from Old Norse, but as an Old English ancestor of this form has not survived, it was kept. The exceptions to the above structure are Ferring Frisian *du*, Frasch Frisian *düünj*, and Saterland Frisian *reke*. In the case of the first two it seems likely that they are derived from the verb "to do", and have undergone a semantic shift. These were thus assigned the same code. The Saterland form was assigned its own code. Additionally, Old English had *sellan* as a synonym for *gifan*; this word survives in Modern English as *to sell*, having undergone semantic narrowing over time. Due to this, two simulations were performed with different forms in Old English; under the semantically lax condition *to sell* was used for Modern English.

The concept GOOD was uniformly represented across the Germanic languages by words with the structure $C_{[+vel,+obst]}VC_{[+alv]}$, as in Gothic *goþs*, Old English, Old Saxon and Old Frisian *gōd*, Old High German and Old Dutch *guot*, [gʊɔt] and [xʊɔt], and Old Norse *góðr*, Modern English *good*, Ferring Frisian *gud*, West Frisian *goed*, Saterland Frisian *goud*, Frasch Frisian *gödj*, Low German *good*, Standard Dutch, Flemish and Afrikaans *goed*, Standard High German *gut*, Bavarian *guad*, Swiss German *guät*, Danish *god*, Faroese *góður*, Icelandic *góður*, Norwegian *god* and Swedish *god*/*gott*. The reconstructed Proto-Germanic form is **godaz* (Orel, 2003). In the case of Swedish there were two forms, however, namely *god/gott*, which correspond with the above forms, and *bra*, ultimately from French *brave*, via Middle Low German, which has become extremely common and is often used as something of a synonym of *gott* (SAOB, 2015). Due to this, two simulations were performed under the semantically strict condition.

The concept GREEN was likewise almost uniformly represented by words with the same core structure across the Germanic languages. The stem structure of these words was /$C_{[+obst,+vel]}$Vn(V)/. This can be seen in Old English and Old Frisian *grēne*, Old Saxon *grōni*, Old High German *gruonaz/gruoni*, Old Dutch *gruoni*, and Old Norse *grœnn*, Modern English *green*, Ferring Frisian *green*, West Frisian *grien*, Saterland and Frasch Frisian *grain*, Low German *grön*, Standard Dutch, Flemish and Afrikaans *groen*, Standard High German *grün*, Bavarian *grea*, Swiss German *grüen*, Danish *grøn*, Faroese *grønur*, Icelandic *grœn*, Norwegian Nynorsk *grøn*, Norwegian Bokmål *grønn* and Swedish *grön*. A word for the concept GREEN does not appear to be attested in Gothic, however; this necessitated the

adoption of the three strategies mentioned above: perform a simulation with missing data items assigned the most common codes for items under the that particular concept, code missing items as deletions, and simply remove all missing items. Orel (2003) gives **groniz* as the reconstructed Proto-Germanic ancestor form of the attested words.

The concept HAIR (single) on the Swadesh 100-word list was represented by three forms across the Germanic languages. The most common, which was attested in twenty-two of the twenty-three varieties examined, had the structure /hV(*r*)/, where the *r* in parentheses indicates that the rhotic segment occurs in most of the words, but not all (Received Pronunciation does not have syllable-final rhotics, although American English does). This form can be seen in words such as Old English *hār,* Old Saxon, Old High German and Old Dutch *hār*, Old Frisian *hēr*, and Old Norse *hár*, as well as in all of their descendants. Swedish *hårstrå* is a compound which has come to be used to distinguish a single hair from the collective hairs which *hår* can refer to (SAOB, 2015); as it contains the element *hår-* it was still coded as a cognate of the above words. These were all assigned the same code, as no evidence of borrowing could be found. Additionally, two words with the core structure /fVC$_{[-vo,+vel,+obst]}$s/ were attested; these are Old English *fēax*, and Old High German *fahs*. These, or words based on them, are not attested in the living descendants of these languages. As there is no clear indication that they had any significant semantic difference from *hār* and *hār*, two separate, semantically strict simulations had to be performed. *Feax* and *fahs* were coded as cognates. Finally, Gothic had the word *tagl*, which is not attested in any of the other languages with the meaning "hair", although it is cognate with the Old English *tægl*, "tail". Under the concept HAIR (single), *tagl* was given its own code. Orel's (2003) reconstructions for HAIR in Proto-Germanic are **hēran* and **faxsan*. This resulted in two simulations with the Proto-Germanic data being run.

The next concept on the list was HAND. This was represented by words with the same structure across all of the languages studied. The structure of these words was /hVn(C$_{[+alv,+plos]}$)/, where the parentheses indicate that this final consonant is not found in all of the forms. This structure can be seen in Gothic *handus*, Old English and Old Saxon *hand*, Old High German and Old Dutch *hant*, Old Frisian *hond*, Old Norse *hǫnd*, and in all of their descendants. In Ferring Frisian, West Frisian and Frasch Frisian the word is missing the alveolar plosive after the nasal; Ferring Frisian has the word *hun*, [huːn], and West Frisian has *hân*, [hoːn]; in these two cases this can be easily explained by syncope of the final stop. However, the palatalization of the final nasal in Frasch Frisian *hönj*, [hønʲ] is difficult to explain as a process of sound change acting on the final stop; as this palatalization of final consonants occurs in other words where no deletion of a stop has taken place, such as *iinj*, [iːnʲ], "one", *rüüdj*, [ryːdʲ], "red" and *rötj*, [røtʲ], "root", it seems more probable that it reflects a process separate from the deletion of the homorganic stop, which, if ease of pronunciation is anything to go by, probably happened later on. Also, the vowel which occurs between the Gothic masculine nominative singular marker *–s* is hard to explain as part of this marker in light of the fact that numerous other words which have a masculine nominative singular marker lack this; this suggests that this is the descendant of a vowel which was part of the stem in Proto-Germanic but has been deleted in the other Germanic languages. Orel (2003)

gives **handuz* as the Proto-Germanic form. As these differences in core structure are not major, and do not detract from the fact that these words are cognate with the others under the concept HAND in the dataset, and that there is no evidence which could be used to argue for the borrowing of the *hand*-form at the expense of something else, they were assigned the same code.

The concept HEAD had three representative forms in the dataset. The most widespread of these had the core structure/hV($C_{[+lab]}$)(V)($C_{[+obst]}$)/, where the elements in brackets were not found in all of the forms, but were nevertheless common, and where at least one bracketed consonant had to be present. The full form can be seen in Gothic *haubiþ*, Old English *hēafod*, Old Saxon *hōvid*, Old Dutch *hōvit*, Old High German *houbit*, Old Frisian *hāved*, Old Norse *hǫfuð*, Swedish *huvud*, Danish *hoved*, Icelandic *höfuð*, and Norwegian Nynorsk *hovud*. Contracted versions of this form can be seen in Modern English *head*, Modern Dutch and Flemish *hoofd*, Afrikaans *hoof*, Modern High German *Haupt*, Ferring and Frasch Frisian *hood*, Saterland Frisian *Haud*, Faroese *høvd*, and Norwegian Bokmål *hode*. These have arisen by various largely independent processes of syncope of either the [+labial] consonant, the second vowel, or the final obstruent. These forms all ultimately descend from a common ancestor, namely Proto-Germanic **habudan~*houbidan* (Orel, 2003), cognate with Latin *caput* (OED, 2015). The second most attested form for HEAD had the structure /kVp(f)/, as seen in Old High German *kopf*, Modern Standard High German and Swiss German *Kopf*, Saterland Frisian *Kop*, Low German *Kopp*, Afrikaans *kop*, and Bavarian *Koopf*. This form is ultimately a borrowing from Latin, and originally had the meaning "drinking vessel, bowl, container"; this was borrowed into Old High German and initially appears to have taken on the meaning "skull", before shifting to mean more generally "head" (this happened in the early Old High German period) (Krause, 2001). The form seems to have spread northwards, entering Middle Low German and Saterland Frisian at later dates. In Afrikaans the word is descended from Middle Dutch *cop*, which originally referred to an animal's head (GTB, 2015). In all these cases the *kop* form has not entirely displaced the older form, although in a number of them (such as Afrikaans and Standard High German) it has become the form more commonly used. Under the semantically strict condition, in which the most commonly used and neutral word for a concept is used, this would be the set of forms used for these languages; however, the semantic difference between the *h*-forms and the *kop* forms is not great, and in fact the main difference in their meanings is that the *h*-forms simply cover a wider semantic field than the *kop* forms. Also, the *h*-forms are still very widely used with the meaning "$\text{head}_{[\text{body part}]}$", and so it would seem inaccurate to include them under the semantically lax condition for these languages; they were thus included under a second semantically strict simulation. The last word under the concept HEAD is West Frisian *holle*. This was coded separately from the other *h*-forms as it is unrelated to them; Hoekstra (personal correspondence) suggests a possible connection to Standard German *hüllen*, "cover of the brain, cranium".

The item TO HEAR was represented by cognates in all of the Germanic languages, with the core structure /hVC/ (in instances where there is a vowel or vowel-nasal segment after the last consonant, these are verb infinitive markers). Gothic has *hausjan*, which displays

Gothic's characteristic non-participation in the process of the rhoticisation of Proto-Germanic *z. In all of the other varieties the reflex of this sound is some form of rhotic consonant (with the exception of RP English, which does not have rhotics in syllable coda position, although the older form in which these were pronounced is reflected in the spelling of many words), thus: Old English *hieran*, Old Saxon (*gi-*)*horian*, Old High German and Old Dutch *hōren*, Old Frisian *hēra*, Old Norse and Icelandic *heyra*, Modern English *to hear*, Ferring Frisian *hiar*, West Frisian *hearre*, Saterland Frisian *here*, Frasch Frisian *hiire*, Low German *horn*, Modern Dutch *horen*, Flemish *hooren*, Afrikaans *hoor*, Modern Standard High German *hören*, Bavarian *hearn*, Swiss German *khööre*, Danish and Norwegian Bokmål *høre*, Faroese *hoyre*, Norwegian Nynorsk *høyre* and Swedish *hora*. The Swiss German word, *khööre*, [k^{h}øːrə], has an aspirated velar plosive rather than a glottal fricative due to the combining of the past tense prefix *ge-* with the original initial [h] of the root; this was subsequently reanalysed as the present tense of the verb (Salmons, personal correspondence). These were all listed as cognates and assigned the same code. The reconstructed Proto-Germanic word is **hauzjanan*~**hausjanan* (Orel, 2003).

The concept HEART was, like TO HEAR, represented by cognate terms across all of the languages concerned. The core structure of the words for this concept was /CV(*r*)$C_{[-vo,+alv,+obst]}$/. Gothic has the form *hairto*, Old English *heorte*, Old Saxon and Old Dutch *herta*, Old High German *herza/herzi*, Old Frisian *herte*, Old Norse, Faroese and Icelandic *hjarta*, Modern English *heart*, Ferring and Frasch Frisian *hart*, West Frisian *hert*, Saterland Frisian *Haat*, Low German *Hart*, Modern Dutch, Flemish and Afrikaans *hart*, Modern Standard High German and Swiss German *Herz*, Bavarian *Heaz*, Danish and Norwegian Bokmål *hjerte*, Norwegian Nynorsk *hjarte*, and Swedish *hjärta*. Orel's (2003) reconstructed form is **herton*. These were all coded as cognates, as in most cases they can be seen to have descended quite regularly from a single ancestral form; where there are differences between the forms, these can be explained by different processes of sound change acting on the form in different stages of the languages under consideration. For example, the diphthong spelled <eo> in the Old English form is due to the so-called breaking of single vowels (van Kemenade, 1994; Fortson, 2004) which was prevalent in Early West Saxon, the dialect of Old English which is best attested and from which this data is drawn. In the case of Old High German and its descendants having an affricate where other varieties have a simple stop for the post-rhotic consonant in the stem of the word, this is due to the High German Consonant Shift, which resulted in, among other things, consonants which were voiceless plosives becoming voiceless affricates (Fortson, 2004). In most instances, it is very unlikely any borrowing took place; if it did, there is hardly any evidence to show it took place.

Like the above two concepts, HORN (the body part of certain animals, rather than the instrument) was represented by cognate forms across all of the languages. The core structure of the word was largely the same in the languages concerned, with most having the structure /h$V_{[+back]}$(*r*)n/. This structure is plainly evident in Gothic *haurn*, Old English, Old Saxon, Old Dutch, Old Norse and Old High German *horn*, Old Frisian *herne*, Modern English, Frasch Frisian, Danish, Icelandic, Faroese, Norwegian and Swedish *horn*, Modern Standard High German and Swiss German *Horn*, Bavarian *Hoan*, Ferring Frisian *hurn*, West Frisian *hoarn*,

Low German *Hurn*, and Modern Dutch and Flemish *hoorn.* Afrikaans is different as it has part of the stem *hor-*, but has replaced the stem-final alveolar nasal [n] with *–ing*, [ɪŋ̟]; this is the result of a sound change in which the final alveolar nasal became syllabic, before having an epenthetic vowel inserted between it and the trill and being backed so that it was produced as a velar (van der Spuy, personal correspondence). Afrikaans *horing* is a cognate of the other forms and was coded as such. The Saterland Frisian word was also something of a curiosity, having a voiced alveolar plosive and an epenthetic vowel where most of the other languages posess a rhotic segment; this could have possibly come about through the emergence of a tap pronunciation of the pre-nasal rhotic, with this subsequently having a schwa inserted between it and the nasal, possibly due to this being an easier pronunciation; speakers may later have come to interpret this flap as a quick alveolar plosive, with the result that this pronunciation took over. The reconstructed Proto-Germanic word is **hurnan* (Orel, 2003)

The concept I (the first person singular) was represented very uniformly by related forms across all of the Germanic languages. Three main forms occurred in the data: one /$V_{[+fro]}C_{[-vo,+obst]}$/, another /jV($C_{[+vo,+vel]}$)/, and two lemmas /$V_{[+fro]}$/. Despite these differences, these forms all descend from a common ancestor, and are the results of later sound changes which have affected the different varieties of Germanic. If the most common features of these forms are taken, the ultimate core structure would be /$VC_{[+vel/pal]}$/. This structure can be very clearly seen in the first person singular pronouns of most of the West Germanic varieties: Old English has *ic*, [ɪʧ], Gothic, Old Saxon, Old Dutch and Old Frisian *ik*, [ɪk], Old High German *ih*, [ɪç], Old Norse *ek*, [ɛk]; West, Ferring and Frasch Frisian *ik*[ɪk], Saterland Frisian *iek* [ik]; Modern Dutch and Flemish *ik* [ɪk], Afrikaans *ek* [ɛk]; Modern Standard High German *ich* [ɪç], Swiss German *iich* [iç], and Low German *ick* [ɪk]. The second set of forms, with a glide before the vowel, are represented by Icelandic *ég*, [jɛg] (Icelandic <é> represents the sounds [jɛ]; (Thráinsson, 1994)), Danish *jeg*, [jaɪ], Norwegian Bokmål *jeg*, [jeɪ], and Swedish *jag*, [jaː]. In all of these cases the presence of the prothetic glide strongly resembles the outcome of the process of *a*-breaking, in which a short /e/in the root syllable of a word would be realised as [ia], and later [ja], if the following syllable contained the vowel /a/ (Faarlund, 1994). Orel (2003) reconstructs the first person singular pronoun of Proto-Germanic as **eka*; however, the final vowel was lost before breaking took place; the current glide came about by diphthongisation of the vowel after it became lengthened (Axelson, personal correspondence). The forms which consist of a vowel only are Modern English *I* and Bavarian *i*. These forms originally had a syllable final voiceless velar plosive, which came to be pronounced less and less when the pronoun was unstressed; these unstressed "weak" forms came to replace the older "strong" forms, leading to them coming to be the only forms of the pronouns used in English (OED, 2015); a similar process appears to have occurred in Bavarian. The English pronoun also underwent the Great Vowel Shift, which lead to its diphthongisation, from [iː] to current [aɪ] (Lass, 1997; Barber, Beal & Shaw, 2012). As all of these forms are descended from a single common ancestor, they were assigned the same codes and treated as cognates under all conditions.

The concept TO KILL had eleven forms across the data; additionally, a number of different forms occurred in some of the languages without clear semantic distinctions. Gothic had the

word *usqiman*, which did not appear to have any clear cognates in the other languages, and was assigned its own code. One set of words had the root structure /slV($C_{[+vel]}$)/, as in Old English *ofsléan*, Old Saxon and Old High German *slahan*; these all mean TO KILL, although they imply that the killing was done through an act of violence, especially with the use of a weapon, striking something so that it dies (OED, 2015). These are thought to be derived from Proto-Germanic **slaxanan* (Orel, 2003), and were assigned the same code. Another set of words with the meaning TO KILL had the structure /kwVl(ː)Vn/, as in Old English *cwellan*, Old Saxon *quellian* and Old High German *kwellan*; these are thought to be descended from Proto-Germanic **kweljanan* (Orel, 2003). These were likewise assigned the same codes; due to difficulties in determining semantic differences between these words and those above, more than one semantically strict simulation was required. Old Frisian *deia* stood on its own and was assigned its own code. The common Old Norse word for TO KILL was *drepa*; this is the ancestor of Danish *dræbe*, Faroese *drepa*, Icelandic *drepa*, and Norwegian *drepa*. These were all assigned the same code. Old High German additionally had the word *tōden*, whence Standard High German *töten*, and Swiss German *tööte*; relate to this are Standard Dutch *doden*, Flemish *dooden* and, via Middle Low German (SAOB, 2015) Swedish *döda*. These were assigned the same codes. Modern English (*to*) *kill* is of uncertain etymology, although it is thought not to be related to any of the attested Old English forms (OED, 2015); this led to it being assigned its own code. A number of languages expressed TO KILL by combining the adjective *dead* with the verb (*to*) *make*, as in Ferring Frisian *duadmaage*, Saterland Frisian *doodmoakje*, Frasch Frisian *düüdj mååge*, Low German *dood moken*, and Afrikaans *dood maak*. Despite having a word based ultimately on the same root as that in *doden*, it was decided to code these differently to capture the fact that a different strategy is used in these languages. West Frisian *deadzje* was run in one simulation as being a cognate of *doden*, and in another as not. In Modern English, the words *slay* and *quell* exist, the former generally restricted to poetic or literary use (OED, 2015) and the latter with the meaning "to crush or put down (an uprising)"; these were included in two semantically lax simulations. Standard High German *schlagen* generally means "to strike' (Duden, 2015), while *quälen* means "to torture". These were also include in two semantically lax simulations. A word meaning TO KILL for Old Dutch was not found; therefore three simulations under the conditions Majority Wins, Infodelete and Infoclean were performed.

The concept KNEE was represented across all of the languages by cognates. The core structure of all of these words was /(k)nV/. This could be seen in Gothic *kniu*, Old English *cneo*, Old Saxon, Old High German and Old Dutch *knio*, Old Frisian *knī*/*knē*, Old Norse *kné*, Modern English *knee*, Low German *Knee*; Standard Dutch, Flemish and Afrikaans *knie*, Standard Modern High German *Knie*, Bavarian *Knia*, Swiss German *Knoi*; Danish *knæ*, Faroese *knæ*, Icelandic *hné*, Norwegian *kne* and Swedish *knä*. The /k/ of the core structure is placed in parentheses to indicate that it is not universal amongst all of the forms, but is found in most of them. Modern English *knee* [niː] no longer has the velar plosive, having simplified the initial [kn] to [n] throughout the language; this occurred in the Early Modern period (ca. 1450~1650) (Barber, Beal & Shaw, 2012). The Frisian varieties have words which begin with the structure /knV/, but all have additional segments following this structure: Ferring Frisian *knöbian*, West Frisian *knibbel*, Saterland Frisian *Kniebel*, and Frasch Frisian

knaibling. Despite these additional segments, which are later additions to the word, the first syllable of each one corresponds to the core structure given above, which motivated the coding of these words as cognates of the words in the other varieties of Germanic. Orel (2003) reconstructs the Proto-Germanic term as **knewan*. All of these terms were treated as cognates under all conditions.

The next concept on the Swadesh 100-word list was TO KNOW. This raised a slight problem, as many of the Germanic languages have two separate words for TO KNOW, which have a slight semantic difference. If the knowing involves knowing a factual piece of knowledge, such as what the capital of a particular country is, a different word is used to when the knowing involves knowing a person or some animate object personally; this can be seen in Modern Standard High German, Dutch, Flemish and Afrikaans, where this slight difference in semantics is conveyed by the use of two different words: German *wissen* versus *kennen*, Dutch and Flemish *weten* versus *kennen*, Afrikaans *weet* versus *ken*. The approach taken here was that the word with the more widely applicable meaning should be used; this meant that the word used for factual and general information was the one chosen. In this case this resulted in two main forms being attested. The most widely attested had the core structure /$C_{[+lab,-stp]}V_{[+fro]}C_{[+alv]}$/. This can be seen in Gothic, Old English, Old Saxon and Old Dutch *witan*, Old Frisian *wita*, Old High German *wizzan*, Old Norse *vita*, Ferring Frisian *ved*, West Frisian *witte*, Saterland Frisian *wiete*, Frasch Frisian *waase*, Low German *weetn*, Standard Dutch and Flemish *weten*, Afrikaans *weet*, Modern Standard High German *wissen*, Bavarian *wissn*, Swiss German *wüsse*, Danish *vide*, Faroese, Icelandic and Norwegian *vita*, and Swedish *veta*. The second form attested has the core structure /(k)nV/. This can be seen in Modern English *to know*. These forms are those which have to be included when the assumption that the most common and neutral word is used; this was done for the first simulation. If this assumption is ignored and words which are less common but have the same, or a very similar, meaning are used, one also finds that Old English *cnawan* and Modern English *to wit* must be included. This was done in an additional simulation.

The concept LEAF (noun) was represented by two main forms across the Germanic languages. The most widely attested form had the structure /blV$C_{[+alv]}$/. This can be seen in Old Saxon *blad*, Old High German *blat*, Ferring Frisian *bleed*, West Frisian *bled*, Saterland Frisian *Blääd*, Frasch Frisian *blees*, Low German *Blatt*, Standard Dutch, Flemish and Afrikaans *blad*, Modern Standard High German, Bavarian and Swiss German *Blatt*, Danish and Norwegian *blad*, and Faroese *blað*. The second attested form had the structure /lV$C_{[+lab]}$/. This was attested in Gothic *laufs*, Old English *lēaf*, Old Saxon *lōf*, Old High German *loub*, Old Frisian *lāf*, Old Norse *lauf*, Modern English *leaf* and Swedish *löv*. In both cases, all of these words were coded as cognates under the semantically strict condition. Icelandic presented something of a problem, as its word for LEAF is a compound made up of both of the other terms, *laufblað*. The approach taken with it was to run one simulation where it was coded as a cognate of the *Blatt* forms and one where it was coded as a cognate of the *leaf* forms. This may be viewed as somewhat unsatisfactory, on the grounds that it does not take into account the fact that the compounding has occurred, that it is made up of both forms, and that it requires two separate simulations; this could be ameliorated in one of two ways: either

to have two LEAF characters and coded each as a cognate of a different form in each slot (but have the problem of increasing the size of the data set), or have a separate coding scheme for the type of change which has taken place, for example, "COMPOUNDexo". Neither of these approaches was taken here due to constraints on space and time. Additionally, there appears to be no attested term for LEAF in the Old Dutch corpus; because of this, the three approaches to dealing with missing data were used, namely running a simulation under the Majority Wins, Infodelete and Infoclean conditions. The reconstructed Proto-Germanic term given by Orel (2003) was **louban*. This was coded as a cognate of the *leaf* forms. While it is known that the *Blatt* forms originally came about through a semantic shift, in which a term meaning "blade, something with an edge" came to replace the older *leaf* form, the English word *blade*, which is cognate with these terms, was not included as it does not mean LEAF at all and seems to have had a more general sense in Old English (OED, 2015).

The concept TO LIE (physically place one's body down) was represented by cognate terms across all of the languages. The structure of the root for this concept was /lV(C)/, where the consonant in parentheses indicates that it was an extremely common segment in the terms in the dataset, but was not universal. This consonant was generally a velar obstruent; this can be seen in Gothic *ligan*, Old Saxon *liggian*, Old High German and Old Dutch *liggen*, Old Norse *liggja*, Standard Dutch and Flemish *liggen*, Modern Standard High German *liegen*, Bavarian *liang*, Swiss German *ligge*, Low German *lingn*, Danish *ligge*, Faroese, Icelandic and Norwegian Nynorsk *liggja*, Norwegian Bokmål *ligge* and Swedish *ligga*. The velar nasals of Bavarian and Low German can possibly be explained by the lack of stress on the vowel of the infinitive marker, resulting in it being elided, with simplification of the resulting [gn] cluster in Bavarian occurring via elision of the velar and spreading of the value [+velar] to the alveolar nasal, while in Low German the velar was not elided but took on the quality [+nasal]. In Old English there is the post-alveolar fricative [ʤ], represented by the digraph <cg> in *licgan*. This arose by palatalization of an original velar. Old Frisian shows a similar change in which the original West Germanic velar has been forwarded to [dz], giving *lidza*. The palatalisation of West Germanic velars in certain environments (particularly before front vowels, but there do appear to be exceptions) in Old English and Old Frisian has been taken by a number of scholars to suggest a particularly close affinity between the two (Fortson, 2004; Lass, 1997). In West Frisian, Saterland Frisian and Frasch Frisian, the [dz] of Old Frisian appears to have a reflex which has come about by simplification of the affricate to either a fricative or a simple stop: West Frisian *lizze*, Saterland Frisian *laze* and Frasch Frisian *lade*. Ferring Frisian has undergone a development similar to that of Modern English, in which the obstruent at the end of the root syllable has been completely elided, with dipthongisation of the vowel occurring at a later date: compare Ferring Frisian *lai* [lai], with Modern English *lie*. Afrikaans *lê*, [lɛː] may have come about through the elision of the intervocalic fricative [ɣ] of Dutch *liggen*, with compensatory lengthening of the vowel and lowering of its position in the mouth; alternatively it may be the result of a tendency in 17th century Dutch to confuse *liggen* and *leggen*, combined with elision of the second syllable's segments (Conradie, personal correspondence). Orel (2003) gives the reconstructed form **legjanan*. As all of these terms can be seen to be cognates, with regular sound changes accounting for the differences between them, they were assigned the same code.

Like TO LIE, the concept LIVER (the internal organ) was represented by forms which were all cognate, with the same root structure, /lVC$_{[+lab]}$(V)(*r)*/. This structure can be seen in Old English *lifer*, Old Frisian *livere*, Old Norse *lífr*, Modern English and Ferring Frisian *liver*, West Frisian *lever*, Saterland Frisian *Líeuwer*, Frasch Frisian *liwer*, Low German *Lever*, Dutch and Flemish *lever*, Afrikaans *lewer*, Modern Standard High German *Leber*, Bavarian *Leeba*, Swiss German *Lääbere* (the last three showing the results of the High German Consonant Shift, which resulted in [v] becoming [b]), Danish, Norwegian and Swedish *lever*, Faroese *lívur* and Icelandic *lífur*. These were all coded as cognates. Orel (2003) has **libaro* as the hypothetical ancestral form of these words. The word for LIVER does not appear to be attested in Gothic, Old Saxon, or Old Dutch; because of this, the strategy of running three simulations under three conditions, Infodelete, Infoclean and Majority Wins, was used to gauge the different effects these conditions would have on missing data.

The next concept was LONG (indicating physical size or space), which was again represented throughout the data by cognate forms. The core structure of these forms was /lV(C$_{[+nas]}$)C$_{[+vel]}$/, where the parenthesised segment indicates that having two consonants is common, but not universal. The two variants can be divided up into those which have a nasal followed by a velar plosive, and those which have only a velar plosive. The first sub-group can be seen in Gothic *laggs* (Gothic <gg> often represents [ŋg]; the *-s* morpheme is a grammatical gender marker and not a part of the root), Old English, Old Saxon and Old High German *lang*, Old Dutch *lank*, Old Frisian *long*, Old Norse *langr*, Afrikaans *lank*, and Faroese and Icelandic *langur*. The second group can be seen in Modern English *long*, Ferring Frisian *lung*, West Frisian *lang*, Saterland Frisian *long*/*loang*, Frasch Frisian *lung*, Low German, Dutch, Flemish, Modern Standard High German, Bavarian and Swiss German *lang*, Swedish *lång*, and Danish and Norwegian *lang*.These forms are descended from ancestral forms which all had a nasal- velar stop combination at the end of the root; these languages simplified the nasal-velar stop combinations to the velar nasal [ŋ]. The reconstructed form is **langaz* (Orel, 2003). These were all coded as cognates.

LOUSE was represented by cognate forms in all but four of the languages. The root structure of the word for LOUSE was /lVs/, as in Old English and Old High German *lūs*, Old Norse *lús*, Modern English *louse*, Ferring Frisian *lüs*, West Frisian *lûs*, Saterland Frisian *Lúus*, Frasch Frisian *lüs*, Low German *Lus*, Standard Dutch, Flemish and Afrikaans *luis*, Modern Standard High German and Bavarian *Laus*, Swiss German *Luus*, Danish, Norwegian and Swedish *lus*, and Faroese and Icelandic *lús*. Orel's (2003) reconstructed word for Proto-Germanic is **lusz*. The vowel qualities in all of these forms generally suggest little borrowing; the diphthongs in both English *louse* [laʊs], and German and Bavarian *Laus* [laʊs], are not cognate, having arisen in parallel, with English having undergone the Great Vowel Shift (Lass, 1997; Barber, Beal & Shaw, 2012), and the German varieties having undergone a separate diphthongisation of the long vowels /iː/, /uː/ and /yː/ of Old High German from the 13th to 16th centuries (Salmons, 2012). The Netherlandic diphthongs arose in parallel as well. However, the words themselves are still cognates, with no lexical replacement having occurred. They were thus coded as cognates. The four languages which did not appear to have any words for LOUSE attested in their respective corpora were Old

Saxon, Gothic, Old Dutch and Old Frisian. The strategy of running three simulations under the conditions of Infodelete, Majority Wins and Infoclean was used here.

There were four attested forms for the concept MAN. The most widespread of these had the root structure /mV($C_{[+alv]}$)/, where the root-final consonant can be geminated or short, and the vowel can be long or short (although in this case it is generally short). This can be seen in Gothic *manna*, Old English *mann*, *mon*, Old Saxon, Old High German and Old Dutch *man*, Old Frisian *mon*, Old Norse *maðr*, Modern English *man*, Ferring Frisian *maan*, West Frisian *man*, Saterland Frisian *Mon*, Frasch Frisian *moon*, Low German *Mann*, Standard Dutch, Flemish and Afrikaans *man*, Modern Standard High German *Mann*, Bavarian *Moo*, Swiss German *Maa*, Danish *mand*, Faroese and Icelandic *maður*, Norwegian *mann*, and Swedish *man*. The [ð] of the Old Norse and Icelandic forms is descended from an original **n* (Faarlund, 1994) In the case of Faroese, the <ð> does not correspond with a particular phoneme in the modern language, as the sound [ð], which it inherited from Old Norse, was later lost from the language, resulting in many instances in vowel hiatus where these had formerly been broken by the fricative (Barnes & Weyhe, 1994). The Bavarian and Swiss German words, which lack a final nasal, appear to have come about by deletion of the original alveolar nasal, with compensatory lengthening of the vowel; in Bavarian the vowel appears to have also undergone raising. Originally, the alveolar consonant was a nasal, and the reconstructed form of the word given by Orel (2003) is **manan*. These were all assigned the same code. Another attested word for MAN has the structure /gVmV/; this form occurs in Old High German *gomo*, Old English *guma*, Gothic *guma* and Old Saxon *gome*. This form is older than the *man* forms, going back to Proto-Indo-European **g^{hw}omo-*; it is cognate with Latin *homo* (OED, 2015), showing the effects of Grimm's Law, where Proto-Indo-European **g^{h}* has become [g] rather than [h] in Germanic (Fortson, 2004). As these forms occur alongside the *man* forms, without any significant semantic differences between them, two semantically strict simulations had to be performed, with one substituting the *gomo* forms for the *man* forms in the languages in which they are attested. Alongside **manan*, Orel (2003) lists the word **gumon* as likely in Proto-Germanic. A third form, with the structure /$C_{[-vo,+obst]}$ Vrl/ can be seen in Old Norse and Icelandic *karl*. In Swedish the word *karl* exists, but it has prominent connotations of manliness which the more neutral *man* has not got, and is used less often than *man* (SAOB, 2015); due to this, it was thought prudent to only include the Swedish word under the semantically lax condition. These were also coded as cognates.

The concept MANY had three forms represented across the Germanic languages. The most widespread of this had the root form /mV$_{[+back]}$(*n*)-/, where the *n* in parentheses indicates a nasal sound which occurred in most of the languages (Old Norse and Icelandic were the exceptions), and the dash indicates that there was always something following this, although it varied dramatically between languages. For example, Gothic had *managa*, Old English *manig*, Old Saxon *manag*, Old Norse *margir*, Modern English *many*, Ferring Frisian *manigen*, Frasch Frisian *maning*, Danish *mange*, Faroese *mangir*, Icelandic *margir*, Norwegian *mange* and Swedish *många*. These descend from a common ancestor, Proto-Germanic **managaz* (Orel, 2003). In Old Norse Proto-Germanic **n* appears to have been rhotacised to [r]. This seems to have reverted to [n] in all of the North Germanic languages

but Icelandic. In Old English, the **g* of Proto-Germanic appears to have been lenited to a glide, before probably becoming silent in word-final position; however, in Gothic and Old Norse the **g* of Proto-Germanic, which was probably the sound [g], was preserved. These were all coded as cognates. The second widely attested word for MANY had the root form /fVl/, where V can be either long or short. This form can be seen in Old High German *filo*/*filu*, Old Dutch *filo*, Old Frisian *fele*/*felo*, Saterland Frisian *fúul*, Low German *vel*, Dutch and Flemish *veel*, Afrikaans *veel*, Modern Standard High German *viele*, and Swiss German *fili*. Bavarian has the form *fui*, which arose as part of a process by which syllable final trills and sonorants are regularly elided, often with lengthening or diphthongisation of the preceding vowel; compare Bavarian *foi*, versus Standrad High German *voll*, "full". Orel (2003) lists **felu* as a likely form in Proto-Germanic. In addition to the above forms, Old English has *fela* alongside *manig*, while Ferring Frisian has *fölen*; these were and are used regularly, without a noticeable semantic difference between them. This necessitated two semantically strict simulations, with *manig* and *manigen* being replaced by *fela* and *fölen* in one of them. In West Frisian the concept MANY is commonly conveyed using the construction *in soad*. This is quite obviously an innovation, and as such was assigned its own code.

There were four attested forms in the data for the concept FLESH. The most common forms, being attested in seventeen out of the twenty-three varieties of Germanic looked at, had the structure /flV$C_{[-vo,+sib]}$(k)/, where the /k/ in parentheses indicates that it is not present in all attested forms, but is common. This structure occurs in Old English *flæsc*, Old Saxon and Old Dutch *flēsk*, Old Frisian *flask*/*flēsk*, Old High German *fleisk*, Modern English *flesh*, Ferring Frisian *fleesk*, West Frisian *fleis*, Saterland Frisian *Flaask*, Frasch Frisian *flååsch*, Low German *Fleesch*, Standard Dutch *vlees*, Flemish *vleesch*, Afrikaans *vleis*/*vlees*, Modern Standard High German *Fleisch*, and Bavarian and Swiss German *Flaisch*. One of the two words for FLESH given by Orel (2003) is **flaiskaz*. These were all coded as cognates under the semantically strict condition. The second most common form in the data has the structure /$C_{[-vo,+obst]}V_{[+fro]}C_{[-vo,+stp]}$/. This form occurs exclusively in the modern North Germanic languages, with the exception of Icelandic, and can be seen in Danish *kød*, Faroese *kjøt*, Norwegian Nynorsk *kjøt*, Norwegian Bokmål *kjøtt*, and Swedish *kött*. The [ʃ] of Swedish *kött* was originally a [k], but was palatalised, and then became an alveolar fricative, as part of a process of palatalization of velar stops before front vowels (Lass, 1997). These were coded as cognates. The regular Old Norse term for FLESH is *hold*; this has been retained in Icelandic. Orel (2003) gives the word **huldan*, which would be ancestral to the Old Norse term, as a probable word in Proto-Germanic. These terms were thus coded as cognates. Finally, Gothic has the term *leik*; this form is not attested in the other Germanic languages with the meaning FLESH, or any related meaning. It was therefore assigned its own code. Another term which now has largely the same meaning as *flesh*, but is more widely used in Modern English, is *meat*. This is descended from Old English *mæt*, which had the more general meaning of "a type of food", and which could be used to denote a number of different foodstuffs, for example, *flæscmæt*, "meat" (OED, 2015). This use has been fossilised in Modern English *sweetmeat*, "something sweet and edible". A second semantically strict simulation was thus performed, with *meat* substituted for *flesh* in Modern English.

The concept MOON was represented by cognate forms across all but one of the Germanic languages; these forms all had the core structure /mVn/, as can be seen in Gothic *mēna*, Old English and Old Frisian *mōna*, Old High German *mānin*/*māno*, Old Saxon and Old Dutch *māno*, Old Norse *máni*, Modern English *moon*, Ferring Frisian *muun*, West Frisian *moanne*, Saterland Frisian *Moune*, Frasch Frisian *moune*, Low German *Moond*, Standard Dutch and Afrikaans *maan*, Flemish *maen*, Standard High German and Swiss German *Mond*, Bavarian *Moond*, Danish, Norwegian and Swedish *måne* and Faroese *máni*. These are all ultimately descended from a common ancestor, Proto-Germanic **menon* (Orel, 2003), and as such were coded as cognates. The only form which was different was that of Icelandic *tungl*, which originally meant "star" (Old Norse *tungl*; compare Old English *tungol*), but underwent a semantic shift to cover the meaning "moon", replacing the older *máni*, which has survived, however, in the Icelandic word for "Monday", *menidagur*. As *tungl* originally meant "star", it is not included under Old Norse or for that matter any of the other old Germanic languages under any conditions; it is assigned its own code under Icelandic.

There were five forms in the data which had the meaning MOUNTAIN. The most common of these had the form /bV(*r*)$C_{[+obst]}$/, where the vowel could be a monophthong or a diphthong and where the final consonant is generally the stop [g]. This form occurs in Old English as *beorg*, Old Saxon, Old High German and Old Dutch as *berg*, Old Frisian as *berch*, Ferring Frisian as *berig*, West Frisian as *berch*, Saterland Frisian *Bíerig*/*Bäierg*, Frasch Frisian *bärj*, Low German *Barch*, Dutch, Flemish and Afrikaans *berg*, Modern Standard High German and Swiss German *Berg*, Bavarian *Beag*, Danish *bjerg* and Swedish *berg*. The probable Proto-Germanic form given by Orel (2003) is **bergaz*~**bergon*. The second most widespread form in the data has the structure /f^{j}V(d)l/, where the segment [l] can be geminated or not, and the [d] in paretheses indicates the presence of a homorganic stop related to the liquid [l]; this stop is limited to Faroese and Icelandic, where geminated liquids and nasals were degeminated and had a homorganic stop inserted in between them and the preceding sound, resulting in clusters such as [dl] (Barnes & Weyhe, 1994; Thráinsson, 1994). This can be seen in Old Norse *fjall*, Faroese and Icelandic *fjall*, and Norwegian *fjell*. These were coded as cognates. English *fell* is a loanword from Old Norse, and is not used in every-day speech very often (OED, 2015); this was therefore not included.Gothic has the word *fairguni*; this does not appear to have a direct cognate in any of the other languages, and so was assigned its own code. Modern English *mountain* is a borrowing from Norman French, but has become the normal word used in most contexts (OED, 2015). Thus it was assigned its own code.

Item 56 on the Swadesh 100-word list is the concept MOUTH (of the human body). This was represented by cognate terms across most of the languages (twenty of the twenty-three); a particular alternation between some of the sounds in these words highlights an interesting potential problem with the lexical cognate approach. The core structure of these twenty words was /mV$C_{[+alv]}$($C_{[+alv,-nas]}$)/. This structure can be seen in Gothic *munþs*, Old English *mūþ*, Old Saxon and Old Frisian *mūth*, Old High German *mund*, Old Dutch *munt*, Old Norse *munnr*, Modern English *mouth*, Ferring Frisian *mös*, Frasch Frisian *müs*, Low German *Munt*, Standard Dutch, Flemish and Afrikaans *mond*, Modern Standard High German *Mund*, Danish *mund*, Icelandic *munnur*, Norwegian *munn* and Swedish *mun*. All of these are cognates,

having descended from a common ancestor, along the lines or Proto-Germanic **munþaz* (Orel, 2003); however, there is a widespread inter-language alternation in all of these related words, namely that in some languages the Proto-Germanic cluster **nþ* has been preserved, in others it has become [nd] or [nt], in some others it has been simplified to a fricative, with loss of the nasal and compensatory lengthening of the vowel, while in others it has been simplified to a nasal on its own. In addition to this, Bavarian has the word *Mai*, which may be a cognate of the above terms, having undergone a process similar to that of Dutch *Mui*, "depth between two coastal sandbanks through which a tidal stream runs" (OED, 2015), in which the nasal was lost, followed by the plosive; if this is the case, it would still be a cognate of the other terms, despite the drastic changes it has undergone. Faroese *muður*, [mɛavʊr], is pronounced without the interdental fricative (its orthography marks where it formerly occurred; the voiced labio-dental fricative is the product of a hiatus resolution strategy (Barnes, & Weyhe, 1994). In both cases the lexical cognate approach may be viewed as limited in that it does not necessarily mark out sound changes. This will be discussed further under "Discussion". All of these words were coded as cognates. The other set of words which meant MOUTH had the structure /muːl(V)/; these were West Frisian *mûle*, Saterland Frisian *Mule*, and Swiss German *Muul*. Two possible etymologies arise here: a possible connection between these words and Old Norse *mál*, "language, speech", or a connection with German *Maul*, "snout". Either is possible, for in both cases their semantic fields conceivably overlap with MOUTH; however, the first is harder to reconcile with Swiss German's geographic position, and would necessitate either positing Old Norse influence much further south than is generally thought, or treating the Frisian varieties as separate from Swiss German, which would be preferable. However, because either is possible here one simulation was performed in which these three were all treated as cognates, and one in which they were split, with Swiss German coded separately to the Frisian words.

The concept NAME (noun) was represented by cognates across all of the Germanic languages, all with the core root structure /nVC$_{[+nas]}$/. This can be seen in Gothic *nāmo*, Old English *nāma*, Old High German, Old Saxon and Old Dutch *nāmo*, Old Frisian *nōma*/*nāma*, Old Norse *nafn*, Modern English *name*, Ferring Frisian *nööm*, West Frisian *namme*, Saterland Frisian *Nome*, Frasch Frisian *noome*, Low German *Nåm*, Dutch *naam*, Flemish *naem*, Afrikaans *naam*, Modern Standard High German *Name*, Bavarian *Naama*, Swiss German *Naame*, Danish, Faroese and Norwegian Bokmål *navn*, Icelandic *nafn*, and Norwegian and Swedish *namn*. The Proto-Germanic form given by the Orel (2003) is **namon*~**namnan*. In the instances where a vowel follows the final nasal of the core root structure above, this is vowel is a reflex of the Proto-Germanic vowel. In many instances this vowel has been entirely elided, resulting in words ending in no vowel in a number of the modern languages, or has been reduced to a schwa (such as in German). The pre-nasal fricatives seen in Old Norse, Icelandic, Faroese, Norwegian Bokmål, and reflected in the spelling of the Danish word are the product of the fricativisation of a homorganic labial which developed in Pre-Old Norse before the bilabial nasal of the Proto-Germanic word as part of a dissimilation process (Snædal-Rosbergsson, personal correspondence). The final [n] in all of these instances is not related to the [m] seen in the West Germanic languages, or Norwegian Nynorsk or Swedish *namn*, being a continuation of the second **n* in the Proto-Germanic words (OED, 2015). This

can still be seen in words such as Old English *genemnan* and Modern Standard German *nennen*, "to name". These words are all cognates and were coded as such under all conditions.

Three separate forms were attested across the Germanic languages for the concept NECK (body part). The most widely attested term had the structure /hV(l)s/. This structure can be seen in Gothic, Old Saxon, Old High German, Old Dutch, Old Frisian, Old Norse, Ferring Frisian, Standard Dutch, Flemish, Afrikaans, Danish, Norwegian and Swedish *hals*, as well as Saterland Frisian *Hoals*/*Haals*, Frasch Frisian *håls*, Modern Standard High German, Low German and Swiss German *Hals*, Bavarian *Hois*, Faroese *hálsur* and Icelandic *háls*. These all appear to be cognates, with a number of them showing regular sound correspondences with each other, such as the presence of *l*-vocalism in Bavarian *Hois*, where the coda-position [l] of Old High German *hals* regularly becomes a vowel, leading to the formation of a diphthong with the original nucleus vowel. In this case the variant which was ancestral to *Hois* must have undergone vowel raising to result in the diphthong [ɔɪ]. In the case of Icelandic, the diphthong [au̯], spelt <á>, occurs where the continental North Germanic languages have the pure vowel [a] or [ɒ]; this is due to a vowel shift which affected Icelandic from the later Middle Ages onwards, and resulted in the dipthongisation of the long vowels of the old Norse spoken in Iceland; the vowels [aː] and [ɔː] merged to [au̯] (Thráinsson, 1994). In Faroese, Old Norse [aː] became [ɔ] (Barnes & Weyhe, 1994). These were all coded as cognates. The second most widely attested set of terms had the core structure /nVk/, as can be seen in Old English *hnecca*, Old High German *nak*, Modern English *neck*, West Frisian *nekke*, and Afrikaans *nek*. These terms all appear to descend from a West Germanic ancestor, and there is no evidence of borrowing from one into the others having occurred; they were therefore coded as cognates. Due to the fact that Afrikaans *nek* and Old High German *nak* co-occur with *hals*, two separate simulations were required under the semantically strict condition. Finally, in addition to *hnecca*, Old English had the term *sweora*/*suira* with the meaning "neck"; as this does not appear to be appreciably semantically different from *hnecca*, a third semantically strict simulation had to be performed. The Proto-Germanic term reconstructed by Orel (2003) is **halsaz*; this was coded as a cognate of the *hals* forms.

The concept NEW was represented by cognate forms across all of the languages examined. All had the core structure /nV($C_{[+gli]}$)/, as can be seen in Gothic *niujis*, Old English *niwe*, Old Saxon, Old High German and Old Dutch *niuwi*, Old Frisian *nī*, Old Norse *nýr*, Modern English *new*, Ferring Frisian *nei*, West Frisian *nij*, Saterland Frisian *näi*, Frasch Frisian *nai*, Low German *ni*, Standard Dutch and Flemish *nieuw*, Afrikaans *nuwe*/*nuut*, Modern Standard High German *neu*, Bavarian *nai*, Swiss German *nòi*, Danish, Norwegian and Swedish *ny*, Faroese *nýggjur* and Icelandic *nýr*. These all show regular sound correspondences and were thus taken to all be descended from a single ancestral form in Proto-Germanic; Orel (2003) does not supply a reconstruction of this form, but something along the lines of **niuwaz* would be expected.

Like NEW, NIGHT was represented by cognate terms across all of the languages concerned. The core structure for this set of terms was /nV($C_{[+fric]}$)($C_{[+plo]}$)/, where at least one of the consonants in parentheses always occurs. This structure can clearly be seen in Gothic *nahts*,

Old English *niht*, Old Saxon, Old High German and Old Dutch *naht*, Old Frisian *nacht*, Old Norse *nótt*, Modern English *night*, Ferring Frisian *naacht*, West Frisian *nacht*, Saterland Frisian *Noacht*/*Naacht*, Frasch Frisian *nåcht*, Low German *Nach*, Standard Dutch and Flemish *nacht*, Afrikaans *nag*, Standard High German, Bavarian and Swiss German *Nacht*, Danish *nat*, Faroese *nátt*, Icelandic *nótt*, and Norwegian and Swedish *natt*. The Proto-Germanic term given by Orel (2003) is **nahtz~*naxtz*; the OED (2015) gives **naχt-*. These were all coded as cognates, as there is no evidence of borrowing, and it seems more likely that the terms would descend from a common ancestor rather than have spread via borrowing from one variety through to all of them.

Words for the concept NOSE were attested in twenty of the twenty-three varieties. In these cases all of the words appeared to be cognates, with the core structure /nVC$_{[+fric]}$(V)/, where the vowel in parentheses indicates that this was a common, but not universal, element. This structure is obvious in Old English *nosu*, Old High German *nasa*, Old Frisian *nose*, Old Norse *nef*, Modern English *nose*, Ferring Frisian *nöös*, West Frisian *noas*, Saterland Frisian *Noze*, Frasch Frisian *noos*, Low German *Nääs*, Standard Dutch, Flemish and Afrikaans *neus*, Standard High German *Nase*, Bavarian *Naasn*, Swiss German *Naase*, Danish *næse*, Faroese *nøs*, Icelandic *nef*, Norwegian Nynorsk *nase*, Norwegian Bokmål *nese*, and Swedish *näsa*. The nasal at the end of the Bavarian form is a development of the vowel in the above given core structure, which took on a nasal quality, and then became a complete syllabic nasal. These forms are all thought to descend from a common ancestor, namely Proto-Germanic **naso* (Orel, 2003), and as such were all coded as cognates. The word for NOSE is not attested in Gothic, Old Saxon, or Old Dutch; because of this, the approach of running simulations under the conditions Majority Wins (missing information is assigned the most common code), Infodelete (missing information is treated as a deletion and coded as such) and Infoclean (missing items result in the characters under which they fall being left out of the simulation) was followed.

The concept NOT (used to negate verbs) was represented by several forms in the data; two of these highlight the effects semantic shift can have on an analysis such as this. In the old Germanic languages, the particles used to negate verbs generally had the form /nV/, as can be seen in Gothic *ni*, Old English *nē*, Old Saxon *nē*/*ni*, Old High German *ni*/*nē*, and Old Frisian *nē*/*ni*. These are all cognates, and were coded as such. In the modern West Germanic languages, the forms which are used to negate verbs are actually compounds based on the older verb negator and an element *wiht*, originally meaning "thing"; these came to be shortened, and their meaning grammaticalised to produce a new verb negator (OED, 2015), as can be seen in Modern English *not*, West Frisian *net*, Saterland Frisian *nit*, Low German *nich*, Standard Dutch and Flemish *niet*, Afrikaans *nie*, Standard High German *nicht*, Bavarian *ned*, and Swiss German *nit*. Due to the fact that these words contain segments descended from *nē*, they could be coded as cognates of the *nē* forms; however, this does not reflect the history of compounding in this cases. From a morphological point of view, it would be more accurate to treat Old English *newiht*, Old Saxon *neowiht*/*niowiht*, Old High German *niowiht*, Old Dutch *niewiht* and Old Frisian *nāwet*; however, these had the meaning "nothing" in these languages (OED, 2015). As a result of this, under the semantically strict condition, two

simulations were run, one with the modern forms coded as cognates of the old form, and the second with them coded differently. The Old Norse word for "not" is given as *eigi*; this is related to *ekki*, meaning "nothing" and derived via negation of the neuter form of the numeral *eitt*, "one" through the use of a negative suffix *-gi/-ki*. This gave rise to Danish and Norwegian Bokmål *ikke*, Faroese *ikki*, Icelandic *ekki*, and Norwegian Nynorsk *ikkje*. Swedish *inte* is related via *ingen*, from the Old Norse pronoun *engi*, "none, no one, no", also based on a similar construction, except using the feminine rather than the neuter numeral (SAOB, 2015); the form *inte* arose through assimilation of the velar plosive to the alveolar position of the nasal. These forms were all coded as cognates. Ferring Frisian and Saterland Frisian have the forms *ei* and *ai*, respectively; the etymology of these is uncertain. If one were to take their surface forms into consideration only, they might be taken to be related in some way to the North Germanic forms. However, this is not certain, and it is possible they are innovations in these varieties; as such, they were coded as cognates of each other, but not the Old Norse forms. In Old Norse there existed also the word *né* for "not" (Orel, 2003); this was included in a separate simulation. The Proto-Germanic forms supplied by Orel (2003) was **ne*.

The next concept is the numeral ONE. This was represented by cognates across all of the languages, and had the core structure /Vn/ where the V almost always represents either a diphthong or a long monophthong (the glide of Modern English *one*, [wʌn] could be analysed as part of a diphthong, as the glide [w] is phonetically not very different from the vowel [u]; however, it could equally be analysed as a prothetic consonant). This structure can be seen in Gothic *ains*, Old English *ān*, Old Saxon *ên*, Old High German *ein*, Old Dutch *ēn*, Old Frisian *ān*, Modern English *one*, Ferring Frisian *ian*, West Frisian *ien*, Saterland Frisian *een*:, Frasch Frisian *iinj* (displaying the widespread palatalization of final consonants seen in a number of Frasch Frisian words), Low German *een*, Standard Dutch and Flemish *een*, Afrikaans *een*, Modern Standard High German *eins*, Bavarian *oans*, Swiss German *ains*, Danish, Swedish and Norwegian Bokmål, *en*, Faroese *ein*, Icelandic *einn* and Norwegian Nynorsk *ein*. The Proto-Germanic word given by Orel (2003) is **ainaz*. These were coded as cognates under all conditions.

The concept PERSON was represented by five different forms, on of which had the core structure /mV$_{[+low]}$C$_{[+alv/+intdent]}$/. This form can be seen in Old English *man*, Old High German *man*, Old Norse *maðr* (this displays the fricativisation of an original [nː] conditioned by the masculine nominative marker *-r* (Faarlund, 1994)); compare the plural *menn*), Faroese *maður* and Icelandic *maður*. These were coded as cognates. A related set of words occurred in a number of the other languages, in which the stem of the word appears to be the same as the above form, but the words end in one of the following segments: /s/, /VʃV/, /ʃ/, /ʃV/, /sk/, /skV/, /VskV/. These endings can be seen in Standard Dutch and Afrikaans *mens*, Swedish *människa*, Norwegian *menneske*, Danish *menneske*, Icelandic *manneskja*, Old Dutch *mennisko*, Old Frisian *menneska*, Ferring Frisian *mensk*, West Frisian *minske*, Saterland Frisian *Moanske*, Frasch Frisian *mansche*, Modern Standard High German and Low German *Mensch*, Swiss German *Mansch*, and Bavarian *Månsch*. These terms are all originally derived from adjectival forms (OED, 2015; SAOB, 2015); this is not only indicated by the endings, but in many instances by the stem vowel, which is often a mid or high vowel, where the

unmodified stem in the *man* forms has a low vowel; this is the product of a derivational process which can still be seen to this day in Modern High German, in which stem vowels in derived words with a back or low vowel tend to be raised and forwarded (compare *Volk*, "people", [ɔ], versus *völkisch*, "popular, of the people", [ø]). These are remnants of the process known as *umlaut.* The adjectives have gradually come to be nouns in their own right, with parallel semantic shifts in which the adjective came to be used to refer more and more to the noun itself, eventually coming to be reanalysed as a noun. As these words appear to ultimately be based on the same stem as the *man* words, they are technically cognates; thus, they were coded as such in an initial simulation. However, this coding scheme does not indicate that there is a difference between these forms and the "man" forms, from which they are derived; thus, a second simulation was performed in which they were assigned different codes to the *man* forms. The usual word for PERSON in Modern English is *person*; this is a borrowing from Anglo-Norman; however, it has replaced the original Germanic terms for PERSON, and does not have any regularly used synonyms or near synonyms in Modern English. It was therefore assigned its own code. Gothic has the word *andwairþi* for PERSON; this contains the element *wair*, which is probably cognate with Old English *wer*, "man" (OED, 2015). This was assigned its own code. Orel (2003) gives the reconstructed word **manan* in Proto-Germanic.

The concept RAIN was represented by cognate forms across the Germanic languages, all with the structure /*r*V$_{[+fro]}$(C$_{[+vel]}$)(V)C$_{[+nas]}$/, where *r* indicates a rhotic segment, and the segments in parentheses are widespread but not universal. This structure can be seen in Gothic *rign*, Old English *rēn*, Old Saxon *regin*, Old High German and Old Dutch *regan*, Old Frisian *rein*, Old Norse *regn*, Modern English *rain*, Ferring and Frasch Frisian *rin*, West Frisian *rein*, Saterland Frisian *Rien*, Low German *Reng*, Standard Dutch *regen*, Flemish *regenen*, Afrikaans *reën*, Standard High German *Regen*, Bavarian *Reng*, Swiss German *Rääge*, Danish *regn*, Faroese *regn*, Icelandic *regn*, Norwegian *regn*, and Swedish *regn*. Orel (2003) has the reconstructed Proto-Germanic word **regnon*~**regnaz*. These were all coded as cognates under all conditions.

Like RAIN, the concept RED was represented by cognate forms across all of the languages studied. The core structure of all these forms was /*r*VC$_{[+obst,\ +alv/+intdent]}$/, as can be seen in Gothic *rauþs*, Old English *rēad*, Old Saxon *rōd*, Old High German *rōt*, Old Dutch *rōt*/*rōd*, Old Frisian *rād*, Old Norse *rauðr*, Modern English *red*, Ferring Frisian *ruad*, West Frisian *read*, Saterland Frisian *rood*, Frasch Frisian *rüüdj*, Low German *root*, Standard Dutch and Flemish *rood*, Afrikaans *rooi* (with elision of the stop), Standard High German *rot*, Bavarian *rood*, Swiss German *root*, Danish *rød*, Faroese *reyður*, Icelandic *rauður*, Norwegian Nynorsk *raud*, Norwegian Bokmål *rød* and Swedish *röd*. The form reconstructed for Proto-Germanic by Orel (2003) is **raudaz*. These were all assigned the same codes under all conditions.

The concept given as number sixty-seven on the Swadesh 100-word list is ROAD. This is a problematic item because it is not entirely clear if the meaning ROAD covers any form of thoroughfare, or if it has a more specific meaning, such as a large thoroughfare which can be used by large numbers of people, etc. It would seem that a better concept would be that of WAY, in the sense of a general physical path along which things can move or be conducted.

This more general concept may not be entirely unproblematic itself, as it may open up the range of potential words which could fall under it; despite this, if the Swadesh 100-word list was originally created with the intention of being used with any language or group of languages, a more general meaning would be preferable on the grounds that it is more likely to yield data, regardless of the degree of development of the society in which the language under consideration is found. Thus, for this analysis, ROAD was replaced by the more general WAY. This yielded cognates across all of the languages in the data set, which had the core structure /$C_{[+lab,]}V_{[-back]}(C_{[+vel,\ +obst]}$/, as can be seen in Gothic *wigs*, Old English, Old Saxon and Old High German *weg*, Old Fisian *wei*/*wi*, Old Norse *vegr*, Modern English *way*, Ferring Frisian *wai*, West Frisian *wei*, Saterland Frisian *wei*, Frasch Frisian *wai*, Low German *weg*, Standard Dutch, Flemish and Afrikaans *weg*, Standard High German, Bavarian and Swiss German *weg*, Danish *vej*, Faroese *vegur*, Icelandic *vegur*, Norwegian Nynorsk *veg*, Norwegian Bokmål *vei* and Swedish *väg*. These all descend from a common Germanic word reconstructed in the OED (2015) as the root **weg-*; Orel (2003) gives the whole word **wegaz* with a masculine ending. These were all coded as cognates.

The concept ROOT (of a plant, a tuber) was represented in the data by two sets of words with different core structures; the most widespread of these had the root structure /$C_{[+lab,\ -stp]}V(r)C_{[+alv/pal-alv]}$/, where the (*r*) indicates a rhotic segment which is very widespread, although not found in all attested forms. This structure occurs in Gothic *waurts*, Old English *wyrtruma*, Old Saxon *wurt*, Old High German *wurz*, Old Frisian *wirtel*/*wortel*, West Frisian *woartel*, Saterland Frisian *Wuttel*, Low German *Wuddel*, Standard Dutch, Flemish and Afrikaans *wortel*, Standard High German *Wurzel*, Bavarian *Wurzl*, and Swiss German *Wurzle*. These all descend from a common Germanic term, reconstructed in Orel (2003) as Proto-Germanic **wurtiz*. These were all coded as cognates. The second set of words had the core structure /$rV_{[-low]}C_{[+alv,+plo]}$/, as seen in Modern English *root*, Old Norse *rót*, Ferring Frisian *rut*, Frasch Frisian *rötj*, Danish *rod*, Faroese *rót*, Icelandic *rót*, and Norwegian and Swedish *rot*. The Modern English word is a borrowing from Old Norse (OED, 2015), but does not have a commonly used synonym, with Old English *wyrtruma* having been replaced by the Middle English period (OED, 2015). The Ferring and Frasch Frisian words are also borrowings from Old Norse, but have likewise replaced the original *wurt* forms; in the case of Frasch Frisian this appears to have occurred before the palatalization of final consonants became common, resulting in the original [t] of the loanword being palatalised later on to [t^j]. Under the semantically strict condition, these were all coded as cognates.The term *wort*, directly descended from Old English *wyrt* is now considered archaic in Modern English, although it does still exist in the names of a number of plants (for example, *liverwort*) (OED, 2015); as the word is now largely restricted in modern usage to certain plant names, it was decided not to include it in any simulations. In Old Norse the word *urt* occurs alongside *rót*, although it is less common; because of this, an additional simulation was run using *urt* (OED, 2015). *Urt* was coded as a cognate of *wortel* based on the regular loss of Old Norse word-initial glides. In Old Dutch, the word for ROOT is not attested, although Middle Dutch has *wortele*, suggesting that a similar word was probably found in Old Dutch; however, as it is not directly attested the approach to handling missing data, namely running simulations under the conditions Majority Wins, Infodelete and Infoclean was used.

The concept ROUND (adjective) was represented by four different word forms in the data. The most widespread of these had the structure /*r*Vn($C_{[+alv,+stp]}$)/ where *r* indicates a rhotic segment and the consonant in parentheses indicates it is very common, but not found in every single form. This structure can clearly be seen in Modern English *round*, West Fisian *rûn*,/*roun*, Saterland Frisian *rund*, Low German *runt*, Standard Dutch, Flemish and Afrikaans *rond*, Standard High German, Bavarian and Swiss German *rund*, Danish *rund*, Faroese *rundur*, and Norwegian and Swedish *rund*. These are all ultimately borrowings from Middle French *rond* (OED, 2015; SAOB, 2015); however, they have largely replaced older words meaning ROUND, and because of their origin in Middle French, they are all technically cognates of each other; they were thus coded as such. In Old English and Old High German the words *sinwēal* and *sinwel* are recorded, respectively; based on their almost identical forms and the fact that the Old English form shows characteristic breaking of the vowel, which in Old High German is a monophthong, it seems probable that they are cognates and not an instance of one borrowing from the other; they were therefore assigned the same codes. In Old Norse and Icelandic the general words for ROUND are *kringlóttr* and *kringlóttur*, respectively; the Icelandic word displays the epenthesis which developed as a way of breaking up consonant cluster which came to be prohibited by the phonotactic rules of later Old Norse (Faarlund, 1994); these were coded as cognates of each other. In Ferring and Frasch Frisian the words *trinj* and *trin* occur; these were coded as cognates. No words for ROUND were attested in Gothic, Old Saxon, Old Dutch and Old Frisian; these were handled by running simulations under the conditions Majority Wins, Infoclean and Infodelete.

Two forms for the concept SAND (noun) were attested in the data. The most common, which occurred in all but one language, had the structure /$C_{[+alv, +fric]}$Vn($C_{[+alv, +stp]}$)/, seen in Old English and Old Saxon *sand*, Old High German and Old Dutch *sant*, Old Frisian *sond*/*sand*, Old Norse *sandr*, Modern English *sand*, Ferring Frisian *sun*, West Frisian *sân*, Saterland Frisian *Sound*, Frasch Frisian *sönj*, Low German *Sant*, Standard Dutch and Flemish *zand*, Afrikaans *sand*, Standard High German, Bavarian and Swiss German *Sand*, Danish *sand*, Faroese *sandur*, Icelandic *sandur*, and Norwegian and Swedish *sand*. These are all descended from a common ancestor, the Proto-Germanic root **sando-* (OED, 2015); Orel (2003) reconstructs the word as **sandaz*. These words were all coded as cognates. Gothic was the exception, with the word *malma* (Forster, Pölzin & Röhl, 2006); this was assigned its own code.

The concept TO SAY was represented by two main forms in the data. One had the root structure /$k^{w}VC_{[+alv/dent, +obs]}$/, as seen in Gothic *qiþan*, Old English *cweþan*, Old Saxon and Old Dutch *kwethan*/*quethan*, Old High German *kwedan*/*quedan*, Old Norse *kveða*, and Saterland Frisian *kwede* (the vowel or vowel+nasal after the final consonant of the stems are infinitive markers). These forms appear to descend from a common Germanic ancestor, which has been reconstructed by Orel (2003) as the Proto-Germanic **kweþanan*. As there is no evidence of borrowing and they appear to be descended from a common ancestor, they were coded as cognates. The second, and more widely attested set of words for TO SAY has the core structure /$C_{[+alv,+fric]}$V(C)/, where the consonant in parentheses is generally a velar or palatal consonant, although it is also realised as a palato-alveolar affricate in Old English, a

voiced alveolar affricate in Old Frisian (although there also seems to have been a form with a palatal consonant), a voiced alveolar stop in Frasch Frisian, and a voiced alveolar fricative in West Frisian. The above core structure can be seen in Old English *secgan*, Old Saxon *seggian*, Old High German *sagen*, Old Dutch *sagon*, Old Frisian *sedza*, Old Norse *segja*, Modern English (*to*) *say*, Ferring Frisian *sai*, West Frisian *size*, Frasch Frisian *seede*, Low German *seng*, Standard Dutch and Flemish *zeggen*, Afrikaans *sê*, Standard High German *sagen*, Bavarian *sang*, Swiss German *saage*, Danish *sige*, Faroese *siga*, Icelandic *segja*, Norwegian Nynorsk *seia*, Norwegian Bokmål *si*, and Swedish *säga*. These all descend from a common ancestor, reconstructed by Orel (2003) as **sagjanan*; the OED (2015) gives **sagǣjan* and **sagjan* as possible ancestral forms in Proto-Germanic. The variations in the stem-final consonant are due to a number of developments which have affected the reflexes of Proto-Germanic **g* at various times in different languages histories; for example, the [x] or [ɣ] of Old and Modern Dutch, as well as Flemish, are the results of the lenition process of West Germanic **g* to a fricative in all positions in pre-Old Dutch, while the palatalization of the original velar to [ʤ] in Old English occurred before written records, possibly alongside the similar process seen in Old Frisian, in which the following glide **j* in the older form of the verb caused the preceding stop to take on the value [+palatal] (Fortson, 2004). In all instances where the final consonant is not present in a particular form, this is due to lenition of the final consonant or semi-vowel, resulting usually in a diphthong or a lengthened vowel. These were all coded as cognates. Due to the fact that a number of the languages in the data set had two words for TO SAY, two simulations were run under the semantically strict condition.

The next concept, the verb TO SEE, was represented by cognates across all of the languages under examination. The root structure of the verb was /(C)$C_{[+alv, +fric]}$V(C)/, where the consonants in parentheses represent consonants which are part of the root, but are not universal. This structure can be seen in Gothic (*ga*)*saihwan*, Old English *sēon*, Old Saxon (*gi*)*sehan*, Old High German *sehan*, Old Dutch *sian*, Old Frisian *sia*, Old Norse *séa*, Modern English (*to*) *see*, Ferring Frisian *sä*, West Frisian *sjen*, Saterland Frisian *sjo*, Frasch Frisian *siinj*, Low German *seen*, Standard Dutch and Flemish *zien*, Afrikaans *sien*, Standard High German *sehen*, Bavarian *seeng*, Swiss German *ksee*, Danish, Swedish and Norwegian Bokmål *se*, Faroese *siggja*, Icelandic *sjá* and Norwegian Nynorsk *sjå*. These are all descended from a common ancestor, the Proto-Germanic root of which is given in the OED (2015) as **sehw-*, and which is reconstructed in Orel (2003) as **sexwanan*. These words were all assigned the same codes and treated as cognates under all conditions.

For the concept SEED (noun) there were three forms attested in the data. The most widespread form had the structure /$C_{[+alv,+fric]}VC_{[+alv,+obst]}$/; this could be seen in Old English *sǣd*, Old Saxon *sâd*, Old High German *sât*, Old Norse *sáð*/*sǽði*, Modern English *seed*, Ferring Frisian *said*, West Frisian *sie*(*d*), Saterland Frisian *Säid*, Frasch Frisian *sädj*, Low German *Såt*, Standard Dutch *zaad*, Flemish *zaed*, Afrikaans *saad*, Standard High German *Saat*, Danish *sæd* and Faroese *sáð*. According to the OED (2015), these are all descended from Proto-Germanic **sǣđi-*/**sǣđo-*, a noun derived from the verb root **sǣ-*, "to sow". Orel (2003) however reconstructs the word as **sediz*. These were all coded as cognates, as there is

no evidence of borrowing having occurred between separate varieties. The second most widely attested form had the structure /fr(j)V/, as seen in Gothic *fraiw*, Old Norse *frjó*, Icelandic *fræ*, Norwegian *frø* and Swedish *frö*. The hypothetical Proto-Germanic antecedent of this word is **fraiwan* (Orel, 2003). These were coded as cognates. Lastly, several forms with the structure /$C_{[+alv,+fric]}$VmV/ occurred, as seen in Old Saxon *sāmo*, Old High German *sāmo*, Standard High German *Samen*, Bavarian *Saama* and Swiss German *Saame*; the form postulated by Orel (2003) as ancestral to these is **semon*. As a number of the languages had more than one word for the concept SEED, for instance Standard High German, Old Saxon and Old Norse, two simulations under the semantically strict condition where performed.

Two word forms for the concept TO SIT were found in the data. The most widespread form had the root structure /$C_{[+alv,+fric]}VC_{[+alv]}$(j)/, as in Gothic *sitan*, Old English *sittan*, Old Saxon *sittian*, Old High German *sizzen*, Old Dutch *sitten*, Old Frisian *sitta*, Old Norse *sitja*, Modern English *sit*, Ferring Frisian *sat*, West and Saterland Frisian *sitte*, Frasch Frisian *sate*, Low German *siddn*, Standard Dutch and Flemish *zitten*, Afrikaans *sit*, Standard High German *sitzen*, Bavarian *sitzn*, Swiss German *sitze*, Danish *sidde*, Faroese *sita*, Icelandic *sitja*, Norwegian Nynorsk *sittja*, Norwegian Bokmål *sitte*, and Swedish *sitta*. Orel (2003) reconstructs the Proto-Germanic words as **setjanan*, whereas the OED (2015) has the word **sitjan*. As there is no evidence which could convincingly indicate borrowing between these forms, it is almost certain they are descended from a common ancestor, and were thus treated as cognates. In Bavarian there also exists the verb *hockä* with the meaning TO SIT, alongside Swiss German *hocke*; these were substituted for the *sit* forms in an additional simulation and treated as cognates.

The next concept in the Swadesh 100-word list was that of SKIN (noun). This concept had four representative forms in the data; of these the most widespread form has the core structure/h$VC_{[+alv/intdent]}$/, as can be seen in Old English *hyd*, Old Saxon and Old Frisian *hūd*, Old High German and Old Dutch *hūt*, Old Norse *húð*, Ferring Frisian *hedj*, West Frisian *hûd*, Saterland Frisian *Häid*, Frasch Frisian *hüd*, Low German *Huut*, Standard Dutch and Flemish *huid*, Afrikaans *huid*, Standard High German and Bavarian *Haut*, Swiss German *Hut*, Danish *hud*, Faroese *húð* (the alveolar fricative is no longer pronounced, having been deleted in all positions (Barnes & Weyhe, 1994)), Icelandic *húð*, Norwegian Bokmål and Swedish *hud*. These are all cognates (OED, 2015) and are supposed to have descended from a Proto-Germanic form such as **hūðiz* (OED, 2015) or **hudiz* (Orel, 2003); as such they were assigned the same codes in the Network program. The second most attested word for SKIN had the core structure /fVl/; this form was found in Old English *fell*, Old Saxon, Old High German and Old Frisian *fel*, Gothic *fill*, Old Norse *fjall*/*fiall*,and Afrikaans *vel*. These words are all descended from a Proto-Germanic ancestor **fello* (OED, 2015) and, as such, were assigned the same codes in the semantically strict condition. As two forms with the same meaning within a single language could not be coded without creating a second line of codes for that language in Network's coding grid, two simulations had to be run under the semantically strict condition to accommodate the instances where there were synonyms in the data. Forms with the structure /(sk)(ʃ)Vn/, where the elements in parentheses in this case are in an either/or relationship, occurred four times in the data, as could be seen in Modern

English *skin*, Old Norse *skinn*, Norwegian Nynorsk and Swedish *skinn*. The latter three are cognates, with the Norwegian Nynorsk and Swedish forms descending directly from the Old Norse. Orel (2003) proposes the Proto-Germanic form **skenþō* as the ancestor of the Old Norse word. The Modern English word is a loanword from Old Norse, but it has displaced the original Old English words with the meaning "human skin" completely, and is now the normal term; due to this, it was assigned the same code as the other three forms under the semantically strict condition. Under the semantically lax condition, however, the Modern English forms which are descended from the Old English one, namely *hide* and *fell*, were included, along with Danish *skind*, which is frequently used to refer to animal skin and has slightly different uses and connotations to *hud*. The Frasch Frisian word *schan* also means *skin*; as it is most similar in structure to the *skin* forms, it was coded as a cognate of them. Due to the presence of two separate words with a meaning similar to SKIN in Modern English, two separate simulations under the semantically lax condition were run.

In the data two sets of forms occurred for the concept TO SLEEP. The most widespread had the root structure /$C_{[-vo,+alv/postalv]}$lVp/, as seen in Gothic *slepan*, Old English *slǣpan*, Old Saxon and Old Dutch *slāpan*, Old High German *slāfan*, Old Frisian *slēpa*, Modern English (*to*) *sleep*, Ferring Frisian *sliap*, West Frisian *sliepe*, Saterland and Frasch Frisian *släipe*, Low German *schlåpm*, Standard Dutch and Flemish *slapen*, Afrikaans *slaap*, Standard High German *schlafen*, Bavarian *schlaffa*, and Swiss German *schloofe*.The alternation between [s] and [ʃ] is due to a process which resulted in the shifting of the position of Old High German [s] from being purely alveolar to post-alveolar before another consonant (Salmons, 2012). All of these words are ultimately descended from the same Proto-Germanic term, given by Orel (2003) as **slēpjanan*. These forms were coded as cognates under all conditions. The second set of words with the meaning TO SLEEP was restricted to the north Germanic languages and had the structure /s$V_{[-low,+back]}$vV/, as seen in Old Norse *sofa*, Danish *sove*, Faroese *sova*, Icelandic *sofa*, Norwegian Nynorsk *sova*, Norwegian Bokmål *sove* and Swedish *sova*. These appear to be a North Germanic innovation, with the form *sofa* not being attested in any of the other Germanic varieties; as these forms are all descended from Old Norse *sofa*, they were assigned the same codes under all conditions.

Under the semantically strict condition, the concept SMALL had three separate forms attested in the data. The most common was a set of forms which had the root structure /lV($C_{[+alv,+obst]}$)($C_{[+alv,+app]}$)/, where at least one of the consonants in parentheses is always present. This root structure can be seen in Gothic *leitels*, Old English *lytel*, Old Saxon and Old Dutch *luttil*, Old High German *luzzil*, Old Norse *lítill*, Ferring Frisian *letj*, West Frisian *lyts*, Saterland Frisian *litje*/*litjet*, Frasch Frisian *latj*, Low German *lütt*, Danish *lille*/*liden*, Faroese *lítil*, Icelandic *lítill*, Norwegian *lille*/*liten*, and Swedish *lilla*/*liten*. The OED (2015) states that the relationship between the West Germanic forms and the East Germanic and North Germanic ones is uncertain, and suggests that the roots used (*-il*, *-l*, and *-en* were originally suffixes) were synonymous and phonetically similar, but nevertheless different from each other; this argument is based on the different root vowels. However, Orel (2003) reconstructs the Proto-Germanic ancestor word as having two possible forms, **litilaz* and **lutilaz*, presumably assuming some dialectal variation. The SAOB (2015) states that the

West and North Germanic forms developed in parallel. In all of the Old Germanic languages the root vowel is [+high], and because of this and the fact that in all instances the root of the word begins with the liquid [l] and ends with an alveolar consonant, it was decided to code them as cognates in one simulation and non-cognates in another; they were removed from a third simulation. The variation seen in the continental Swedish and Norwegian between forms with a root ending in a liquid and a root ending in a stop is governed by whether the adjective is used before a definite or indefinite noun; the stem with the final liquid is used before definite nouns, while that with the stop is used before indefinite nouns (Croghan & Holmqvist, 2010; Strandskogen & Strandskogen, 1995); in Danish this distribution is different, with the *l*-form the main adjectival form and the stop form mostly expressing a small quantity (Lundskær-Nielsen & Holmes, 2011).The second most common set of words had the root structure /$C_{[+obst]}$lv(n)/, as in Old High German *kleini*, Old Dutch *kleni*, Old Frisian *klene*, Standard Dutch, Flemish and Afrikaans *klein*, Standard High German *klein*, Bavarian *gloa*, and Swiss German *chlai*. These forms appear to be an innovation in West Germanic, although they have a patchy distribution in the old Germanic varieties, being attested in Old High German, Old Frisian and Old Dutch, but not Old Saxon or Old English; without clear evidence for borrowing having occurred, these were taken to be simply instances of a lack of attestation, and the forms were therefore coded as cognates. As a number of the languages in the data have more than one form attested without a major semantic difference between the terms, two simulations were run under the semantically strict condition. The last set of words for SMALL in the data had the core structure /smV(C)/, as in Old High German *smah*/*smal*, Old English and Old Frisian *smel*, Old Dutch and Old Saxon *smal*, and Modern English *small*. These were coded as cognates. Orel (2003) reconstructs the form **smalaz* for Proto-Germanic. Besides Modern English, all of the modern West Germanic languages have preserved these words with the semantic shift of SMALL to THIN, NARROW. The modern North Germanic plural adjectives, such as Swedish *små*, are thought to be unrelated, despite a superficial resemblance to Modern English *small* (OED, 2015); due to this, they were not included in any simulations, including those which were semantically lax.

The concept SMOKE (noun) had two representative forms sets in the data. The first and most widely attested has the form /*r*V$C_{[-vo,+obst, +vel/+pal]}$/, as in Old English *rēc*, Old Saxon and Old Dutch *rōk*, Old High German *rouh*, Old Frisian *rēk*, Old Norse *reykr*, Ferring and Frasch Frisian *riik*, West Frisian *reek*, Saterland Frisian *Rook*, Standard Dutch and Flemish *rook*, Afrikaans *rook*, Standard High German, Bavarian and Swiss German *Rauch*, Danish *røg*, Faroese *roykur*, Icelandic *reykur*, Norwegian Nynorsk *røyk*, Norwegian Bokmål *røyk*/*røk*, and Swedish *rök*. These forms descend from a common ancestral form, which is reconstructed as Proto-Germanic **roukiz* (Orel, 2013); they were thus coded as cognates under the semantically strict condition. The second attested form has the structure /smVk/, and is found in Old English *smōca* and Modern English *smoke*; these were coded as cognates. While *smoke* has become the most neutral and common term for SMOKE in Modern English, the word *reek* does still exist, although it has undergone a semantic shift from SMOKE to ODOUR; due to this, a second simulation was performed under the semantically lax condition. A term for SMOKE does not appear to be attested for Gothic, therefore this

character was one of those marked out for the three missing data treatments, Majority Wins, Infodelete and Infoclean.

The concept TO STAND was represented by cognate forms across all of the languages in the data, and all had the core structure /$C_{[-vo,+pal-alv,+fric]}$tV/. Often, there is an alveolar nasal after this structure, and frequently a homorganic stop following this. The variations in the forms can be seen in Swedish, Danish and Norwegian *stå*, Bavarian *schtee*, Swiss German *schtoo* versus Old High German and Old Dutch *stān* (the form *stantan* is also attested in Old High German), West Frisian *stean*, Ferring Frisian *stun*, Low German *schtån*, Standard Dutch *staan*, Flemish *staen*, Afrikaans *staan*, and Standard High German *stehen*, versus Gothic, Old English, and Old Saxon *standan*, Old Frisian *stonda*, Old Norse *standa*, Icelandic *standa* and Faroese *standa*. These are all descended from the same ancestral form, reconstructed by Orel (2003) as **standanan*; it can be suggested that this variation in the descendant forms is possibly dialectal in nature, going back to Proto-Germanic times, or as being due to assimilation to another series which alternated long and shortened verb stems in the verb "to go" (although there is no evidence for the direction of assimilation and the process may have gone the other way) (OED, 2015). Another possibility is that "weak" forms of the verb may have arisen and come to gradually replace the "strong" forms in certain tenses in some varieties. As these all have the same or similar core forms, and there is evidence to suggest that variations in these forms may be the result of sound loss or analogical extension of forms, they were assigned the same codes.

There were two different words for the concept STAR in the data. The most widespread had the core structure /$C_{[-vo,+pal-alv,+fric]}$tV(*r*)(n)/, as seen in Gothic *stairno*, Old English *steorra*, Old Saxon *sterro*, Old High German *sterno*, Old Dutch *sterno*/*sterro*, Old Frisian *stera*, Old Norse *stjarna*, Modern English *star*, Ferring Frisian *stäär*, West Frisian *stjer*, Saterland Frisian *Stiern*, Frasch Frisian *stäär*, Low German *Schtirn*, Standard Dutch *ster*, Flemish *ster*, Afrikaans *ster*, Standard High German *Stern* Bavarian *Schtean*, Swiss German *Schtern*, Danish *stjerne*, Faroese *stjørna*, Icelandic *stjarna*, Norwegian *stjerne* and Swedish *stjärna*. The ancestral form of this word reconstructed by Orel (2003) is **sternon*. However, the OED (2015) suggests that the nasal-less forms are a common West Germanic continuation of Proto-Germanic strong masculine **ster-*, while the forms with a nasal continue the weak feminine **sternon*, developed in parallel in Germanic. Based on this, the forms with a nasal were coded as not being cognate with those without one. It was uncertain whether to include the descendants of Proto-Germanic **tunglan* (Orel, 2003) in the simulation; this was eventually left out based on the fact that in many instances it appears to be used with a broader meaning than STAR.

STONE (the noun) was represented by cognates across all of the varieties; in all of the languages the structure of this word was /$C_{[-vo,+pal-alv,+fric]}$tVn/, as in Gothic *stains*, Old English *stān*, Old Saxon, Old Dutch and Old Frisian *stēn*, Old High German *stein*, Old Norse *stein*, Modern English *stone*, Ferring Frisian *stian*, West Frisian *stien*, Saterland Frisian *Steen*, Frasch Frisian *stiinj*, Low German *Schteen*, Standard Dutch and Flemish *steen*, Afrikaans *steen*, Standard High German *Stein*, Bavarian *Schtoa*, Swiss German *Schtei*, Danish *sten*, Faroese *steinur*, Icelandic *steinn*, Norwegian Nynorsk *stein*, Norwegian Bokmål *stein*/*sten*

and Swedish *sten*. These descend from Proto-Germanic **stainaz* (Orel, 2003) or **stainoz* (OED, 2015). These were all assigned the same codes.

The concept SUN (noun) had two sets of words for it represented in the data. The most common of these had the root structure /$C_{[+alv,+fric]}$Vn/, as in Gothic *sunno*, Old English *sunne*, Old Saxon *sunna*, Old High German *sunna*/*o*, Old Dutch *sunna*, Old Frisian *sunne*, Modern English *sun*, Ferring and Frasch Frisian *san*, West Frisian *sinne*, Saterland Frisian *Sunne*, Low German *Sunn*, Standard Dutch and Flemish *zon*, Afrikaans *son*, Standard High German *Sonne*, Bavarian *Sonn* and Swiss German *Sunne*. The second form was limited to the North Germanic languages, and had the structure /s$V_{[+hi,+ro]}$l/, as in Old Norse *sól*, Danish *sol*, Norwegian *sol*, Swedish *sol*, Faroese *sól* and Icelandic *sól*. These words are ultimately cognates of the *n*-forms, both going back ultimately to the Proto-Indo-European root **sau*-; however, the OED (2015) states that the forms are different because they make use of different formants (**n* and **l*, respectively) and shows that these formants are distributed differently across the Indo-European languages. Based on this, and the fact that these formants were not productive derivational segments in Proto-Germanic (they were, however, productive in Proto-Indo-European) it was thought probable that the two variant words existed in Proto-Germanic, and so the differences in forms are an example of already separate forms in Proto-Germanic coming to be selected as dominant in different branches of Germanic. Thus they were not treated as cognates. Despite this, Orel reconstructs only **sunnon* as a Proto-Germanic word. Two simulations were performed using different Proto-Germanic forms for this concept, Orel's (2003) and a hypothetical *l*-form.

The concept TO SWIM was represented by cognates across most of the languages, all having the root structure /$C_{[+pal/alv+fric]}$(v/w)V$C_{[+nas]}$/, where the elements in parentheses were common but not universal. This structure can be seen in Old English and Old High German *swimman*, Old Frisian *swimma*, Old Norse *svima*, Modern English (*to*) *swim*, Ferring Frisian *sweem*, West and Saterland Frisian *swimme*, Frasch Frisian *swume*, Low German *schwimm*, Standard Dutch and Flemish *zwemmen*, Afrikaans *swem*, Standard High German *schwimmen*, Bavarian *schwimma*, Swiss German *schwimme*, Danish *svømme*, Faroese *svimja*, Icelandic *synda*, Norwegian Nynorsk *symja*, Norwegian Bokmål *svømme* and Swedish *sima*. The reconstructed Proto-Germanic form is **swemmanan* (Orel, 2003); in some dialects of Germanic so-called "weak" forms of the noun arose which did not have the labio-velar glide or the labiodental fricative of the earlier forms (OED, 2015); however, these are nevertheless cognates, the loss of the second onset segment being a later change. These secondary forms came to replace the older forms in some of the North Germanic languages. The above words were all coded as cognates. The verb "to swim" is not attested in Gothic, Old Saxon and Old Dutch; the strategy of including these characters for special treatment under the Majority Wins, Infodelete and Infoclean conditions was followed.

TAIL had four groups of forms across the data, as well as two instances of missing data items. One of the forms had the root structure /stV(r)t/, as in Old English *steort*, Old High German *sterz*, Old Dutch *stert*, Old Frisian *stert*, Ferring Frisian *stöört*, West Frisian *sturt*, Saterland Frisian *Stäit*/*Stit*, Frasch Frisian *stjart*, Low German *Schteert*, Standard Dutch *start*, Flemish *steert*, and Afrikaans *stert*. These all descend from Proto-Germanic **stertoz* (OED,

2015) or **stertaz* (Orel, 2003); they were therefore taken to be cognates and were assigned the same codes. Another group of words had the structure /$C_{[-vo,+pal/alv,+fric]}$(w/v)V$nC_{[-vo,+sib]}$/, as in Standard High German, Bavarian and Swiss German *Schwanz* and Swedish *svans*. The Swedish word appears to be a borrowing from Middle Low German, originally with the meaning of the train of a dress (SAOB, 2015) (the word *swans* is attested in Middle Low German, but not Old Saxon); this word subsequently came to develop the more general meaning of TAIL and became the everyday word for TAIL. Under the semantically strict condition these words were listed as cognates. The rest of the North Germanic languages have words with the structure /h$V_{[+low]}$lV/ listed as their most usual, neutral words for TAIL, as can be seen in Old Norse *hali*, Danish *hale*, Faroese *hali*, Icelandic *hali*, and Norwegian *hale*. In the modern languages these are all instances of the continuation of the original Old Norse form. These were treated as being true cognates and were assigned the same codes. Lastly, under the semantically strict condition three words with the form /$C_{[-vo,+alv,+obst]}V_{[+low]}$(C)(V)l/ occurred: Old English *tægl*, Old High German *zagil* and Modern English *tail*. These are believed to be descended from a common Proto-Germanic word (OED, 2015) such as **taglan* (Orel, 2003), although the original sense of the word appears to have been "the hairy tail of an animal" (OED, 2015), with semantic broadening occurring later on. This would suggest that under the semantically strict condition Proto-Germanic **taglan* should not be used; because of this only **stertaz* was used. Gothic has the word *tagl*, "hair", but because this has preserved the older meaning, rather than being a semantically shifted word originally meaning TAIL, it was not included in a semantically lax simulation. However, Old Norse did have the word *stertr* for TAIL (OED, 2015); this was included in an additional semantically strict simulation.

The demonstrative pronoun THAT had forms which were technically cognate across almost all of the languages; in all of these instances the core structure of the word was /$C_{[+dent/alv,+obst]}$V$C_{[+dent/alv,+obst]}$/. This can be seen in Gothic *þata*, Old English *þæt*, Old Saxon and Old Dutch *that*, Old High German *daz*, Old Frisian *thet*, Old Norse *þat*, Modern English *that*, West Frisian *dat*, Saterland Frisian *dät*, Frasch Frisian *dåt*(*deer*), Low German, Standard Dutch and Flemish *dat*, Standard High German *das*, Bavarian *des*, Danish *det*, Icelandic *þetta*, and Norwegian and Swedish *det* (*där*) (older *detta*). Additionally, Faroese has the word *hatta*, which has a glottal fricative in place of an alveolar stop or interdental fricative; this is the result of a sound change which shifted the original [θ] of Old Norse to [h] (Barnes & Weyhe, 1994). These are all cognates, and ultimately go back to a postulated Proto-Germanic **þat* (Orel, 2003). Afrikaans has a construction which is somewhat like that of the Swedish construction *det där*, or the Frasch Frisian *dåt*(*deer*), where either the definite article or a demonstrative pronoun is used with a word meaning "there" to mean "that" in the demonstrative sense; Afrikaans has the word *daardie*, which is a compound of *daar*, "there", and the definite article *die*. As neither of these elements is descended from Dutch *dat*, it is not a cognate, and was thus assigned a different code to the *dat* words. Lastly, Swiss German has the unusual form *säb*; this word is not related to the *dat* words, being most probably descended from the pronoun *selb*, with *l*-vocalistation; the use of *selb* as a demonstrative is widespread in many souther German dialects (Rowley, personal correspondence). This was thus assigned a separate code.

The demonstrative pronoun THIS was likewise represented by cognates across most of the languages; in almost all instances the words had the core structure /$C_{[+intdent/alv,+obst]}VC_{[+intdent/alv,+obst]}$/, although in this instance the vowel is often higher than in the THAT words; this structure can be seen in Gothic *þatuh*, Old English *þis*, Old Saxon, Old Dutch and Old Frisian *thit*, Old High German *dese*, Old Norse *þetta*, Modern English *this*, West Frisian *dit*, Saterland Frisian *dut*, Low German, Standard Dutch and Flemish *dit/deze*, Standard High German *dies*, Bavarian *dees*, Swiss German *da*, Danish *dette*, Faroese *hetta*, Icelandic *þetta*, Norwegian *dette* and Swedish *detta*. These words are all related to the pronoun THAT, being ultimately derived from it by using it as the base and adding another particle to it (OED, 2015); in Northwest Germanic (i.e.: pre-Old Norse and Proto-West Germanic) the word appears to have been formed by adding the word **se~*si* (possibly meaning "see, behold" (OED, 2015)) to the original simple demonstrative and the definite article, with this compound later coming to be a blend (with different outcomes regarding the manner of the final alveolar obstruent and the vowel quality) and reanalysed as a single word (OED, 2015). In Gothic, the morpheme *-uh*, which strengthened the meaning of the word it was attached to, was suffixed to the simple demonstrative or definite article. This raises an interesting problem regarding the coding of lexical items for use with the network program: if two separate words exist in two different languages, but they are both descended from the same ancestral word, they are cognates; however, if different derivational strategies have been used by speakers of each language to come to these different forms, despite their being cognates ultimately, the two words will still be different and will have elements which are not cognate and cannot be ascribed to sound changes, such as different morphemes. Given that the software used in this only allows for coding items as either cognates or not, a strategy whereby both the cognate and non-cognate elements can be taken into account at the same time cannot easily be pursued; because of this, the items in question were coded as cognate and non-cognate in two different simulations to gauge the effects of doing this. A similar approach was taken with the Ferring and Frasch Frisian words, which make use of an article and the adverb *heer* as a compound. These two were assigned different codes. Despite Afrikaans using a similar strategy, it was not coded as a cognate of the Frisian forms as it is descended from Middle Dutch, which did not use this strategy, and underwent this development in parallel. The Afrikaans form was assigned its own code. As all the words for THIS arose after the Proto-Germanic period, and most are based on Proto-Germanic **þat*, there is no distinct reconstructed form for THIS, and so **þat* was used again.

The second person pronoun YOU was represented by a large number of cognates in the Germanic languages; the most widespread form-set had the structure /$C_{[+intdent/alv,+obst]}V$/, as in Gothic *þu*, Old English *þū*, Old Saxon, Old Dutch and Old Frisian *thū*, Old High German *dū*, Old Norse *þú*, Ferring and Frasch Frisian *dü*, West Frisian *do*/*dû*, Saterland Frisian *du*, Low German and Standard High German *du*, Bavarian and Swiss German *duu*, Danish *du*, Norwegian and Swedish *du*, Faroese *tú* and Icelandic *þú*. These all descend from Proto-Germanic **þu* (Orel, 2003). The variation between the initial consonant being a stop or a fricative is due to later sound changes in various languages; Old High German's interdental fricatives became alveolar stops in all positions at an early date (Fortson, 2004; van der Wal & Quak, 1994), while in Faroese the fricative [θ] became [t] at a later date (Barnes & Weyhe,

1994). All of these forms were coded as cognates. Modern English *you* is descended from Old English *ēow*, which was the accusative/dative second person plural pronoun (OED, 2015); this development with the plural accusative as the normal word for YOU in Modern English came about through the use in the late Middle English period up to the Early Modern period of the plural second person pronoun as a polite way of addressing an individual, with reanalysis of the accusative form as a new nominative, with *you* coming to displace *ye* (Barber, Beal & Shaw, 2012). As the modern form is not derived from the usual Old English singular form, it was decided to assign it a different code. Standard Dutch *jij*, Flemish *jij*/*gij* and Afrikaans *jy*, arose independently (GTB, 2015), and thus were not coded as cognates of any of the other forms (Modern English *you* is a cognate of the Dutch polite form *u* (OED, 2015)). The archaic English *thou* is a cognate of the *þū* forms, as is the obsolete Dutch form *du* (OED, 2015); these were included in a separate simulation.

All of the Germanic varieties examined had cognate words for TONGUE, with the root structure /$C_{[-vo,+pal/alv,\ +obst]}$Vŋ(g)/, as in Gothic *tuggo*, Old English and Old Frisian *tunge*, Old Saxon, Old Dutch and Old Norse *tunga*, Old High German *zunga*, Modern English *tongue*, Ferring Frisian *tong*, West Frisian *tonge*, Saterland Frisian *Tunge*, Frasch Frisian *tung*, Low German *Tung*, Standard Dutch, Flemish and Afrikaans *tong*, Standard High German *Zunge*, Bavarian *Zunga*, Swiss German *Zunge*, Danish *tunge*, Faroese *tunga*, Icelandic *tunga*, Norwegian *tunge*, and Swedish *tunga*. Orel (2003) reconstructs **tungon* for Proto-Germanic. These were all coded as cognates.

The concept TOOTH was represented overwhelmingly by words which all formed one cognate group. These words all had the structure /$C_{[+alv,+stp]}$ V($C_{[+vo,+alv,+nas]}$)($C_{[+intdent/alv]}$)/, as seen in Gothic *tunþus*, Old English *tōþ*, Old Saxon *tand*, Old High German *zan(d*), Old Dutch *tant*, Old Frisian *tōth*, Old Norse *tonn*, Modern English *tooth*, Ferring Frisian *tus*, Frasch Frisian *täis*, Low German *Tään*, Standard Dutch, Flemish and Afrikaans *tand*, Standard High German *Zahn*, Bavarian *Zaan*, Swiss German *Zaa*, Danish *tand*, Faroese *tonn*, Icelandic *tönn*, Norwegian *tan/tonn*, and Swedish *tand*. The reconstructed Proto-Germanic form based on these words is **tanþz* (Orel, 2003) or **tanþ-~*tunþ-* (OED, 2015). In Old English and Old Frisian, the lack of a nasal in the attested forms is due to a process whereby the nasal was deleted, leading to compensatory lengthening of the preceding vowel (Barber, Beal & Shaw, 2012; OED, 2015). The Ferring Frisian *tus* and the Frasch Frisian *täis* have a word-final voiceless alveolar fricative due to a process which affected the North Frisian varieties where the place of stricture in the vocal tract of the interdental fricative [θ] of Old Frisian was shifted backwards slightly, leading to the production of the alveolar fricative [s] instead (Hoekstra, personal correspondence). The front vowel in the nucleus of the Frasch Frisian word is due to the reanalysis of the original plural, with umlaut, to a new singular (Hoekstra, personal correspondence). The Danish and Swedish forms are borrowings from Middle Low German; however, they have completely replaced the original Old Norse reflexes (SAOB, 2015; ODS, 2015). The affricate in the Old High German form (as well as in its descendants) is a product of the Old High German consonant shift, which lead to a number of stops becoming affricates (Salmons, 2010; Fortson, 2004). These were all coded as cognates. In West and Saterland Frisian the words for TOOTH are *tosk* and *Tusk*, respectively; these are

cognate with Modern English *tusk*, and originally had the same meaning, although a semantic shift led to them replacing Old Frisian *tōth* as the normal word for TOOTH (OED, 2015). These were assigned the same code as each other, but were not coded as cognates of the *tooth* forms.

The concept TREE was likewise represented by two groups of words; however, particularly as concerns the old Germanic languages, the semantic status of these words presented some challenges when it came to coding. One set of words had the form /bVm/, as in Gothic *bagms*, Old English *bēam*, Old Saxon and Old Dutch *bōm*, Old High German *boum*, Old Frisian *bām*, Ferring and Frasch Frisian *buum*, West Frisian *beam*, Saterland Frisian and Low German *Boom*, Standard Dutch and Flemish *boom*, Afrikaans *boom*, Standard High German and Swiss German *Baum*, and Bavarian *Baam*. Additionally Old Norse had the word *baðmr*, although this is less common than *tré*. These words are all cognates, descending from an ancestral form which Orel (2003) has reconstructed as **boumaz* in Proto-Germanic; the OED (2015) reconstructs **baumoz* for Proto-West Germanic, and states that the Gothic and Old Norse forms, with their apparent velar stop and interdental fricative before the nasal, respectively, present phonetic complications which make the sounds of a Proto-Germanic form difficult to reconstruct. Under a semantically strict simulation these were all coded as cognates. The second set of forms has the structure /t*r*V/, where *r* represents a rhotic segment. These can be seen in Old English *trēow*, Old Frisian *trê*, Old Saxon *trio*, Gothic *triu*, Old Norse *tré*, Modern English *tree*, Danish *træ*, Faroese *træ*, Icelandic *tré*, Norwegian *tre* and Swedish *träd*. These forms are also descended from a term in Proto-Germanic, **trewan* (Orel, 2003) or **trewo-* (OED, 2015), and were thus coded as cognates. A problem which can be seen here is that there is no evidence that one term is older than the other, or, in the case of the older Germanic languages, had a particularly great semantic difference, and it is clear that much of the difference in the modern languages has been due to one form being preferentially selected; this naturally leads to the question of which forms should be selected. Based on Forster, Pölzin and Röhl (2006) it was decided to use the terms they used for an initial semantically strict simulation, as these were taken to be the most common terms for the concept in each language. An additional semantically strict simulation was performed with the less common words being substituted. Lastly, a semantically lax simulation was performed with English *beam* instead of *tree.*

The numeral TWO was represented by cognate forms across all of the Germanic languages. In a number of the languages concerned there is more than one form of the numeral, each generally being used with a noun of a different gender; this is particularly the case in the older Germanic languages, as in Old English *twegen* (masculine), *twā* (feminine and neuter), and *tū* (neuter); Gothic *twai* (masculine), *twos* (feminine) and *twa* (neuter); Old Saxon *twene* (masculine), *twâ*/*twô* (feminine), *twê* (neuter); Old High German *zwene* (masculine), *zwâ*/*zwô* (feminine), *zwei* (neuter); Old Norse *tveir* (masculine), *tvær* (feminine) and *tvau*/*tvö* (neuter). With the reduction of the complexity of the gender system in most of the daughter languages there has been a corresponding reduction in the forms of the numeral TWO, although these do not always reflect the status of noun gender in the modern languages (Modern Standard High German has three noun genders, but only one form of the numeral TWO is used); in

most of the modern languages there has also been a reduction in the extent to which adjectives have to agree in case, gender and number with their corresponding nouns, of which this reduction in the forms of the numeral TWO is an example. The modern forms are: Modern English *two* (although *twain* exists, it is an archaism (OED, 2015)), Ferring Frisian *taav*/*tâw* (with metathesis of the labiodental fricative and the vowel), West Frisian *twa*, Saterland Frisian *twô*, Frasch Frisian *tou*, Low German *twee*, Standard Dutch and Flemish *twee*, Afrikaans *twee*, Standard High German *zwei*, Bavarian *zwoa*, Swiss German *zwai*, Danish *to*, Faroese *tvey*, Icelandic *tveir*, Norwegian *to* and Swedish *två*. The different forms in the older Germanic languages are paradigmatic, and are due to the adding of inflectional endings for noun gender to the root of the numeral; as can be seen from all of the older forms, as well as their descendants, the root had the structure /$C_{[-vo,+pal\text{-}alv,+obst]}(C_{[+lab,-stp]})V$/, where the element in parentheses is missing in some languages due to its elision at various times (OED, 2015). The reconstructed Proto-Germanic form is **two*(*u*). These were all assigned the same code.

The concept TO WALK was problematic in some ways, as it had four different forms which varied semantically from TO WALK (implying motion using the feet and legs) to GO (expressing movement to a location without implying the particular manner of going). This led to a situation in which some of the words may not actually have been semantically equivalent. When verbs implying movement involving the legs or feet were absent or semantically difficult to judge themselves (such as instances where a verb could mean "run", "walk" and "jump" depending on context) the most neutral motion verb was used. This resulted in two different simulations being done, with different concepts used; TO WALK and TO GO. One of the attested forms had the root structure /$C_{[+vel,+obst]}V(C_{[+nas]})$/, where the element in parentheses is widespread but not universal; this structure can be seen in Gothic *gaggan*, Old English *gān*/*gangan*, Old Saxon *gangan*/*gān*, Old Norse *ganga*, Ferring Frisian *gung*, Low German *gån*, Danish *gå*, Faroese *ganga*, Icelandic *ganga*, and Swedish and Norwegian *gå*. These words generally convey the meaning TO WALK in their respective languages. These are descended ultimately from Proto-Germanic **gēnan* (Orel, 2003). These were coded as cognates. A second form, limited to only some of the older West Germanic languages, but common to all of the modern day ones, had the root structure /l$VC_{[+lab]}$/, as in Old High German *loufan*, Old Dutch *lōpan*, Old Frisian *hlōpa*/*hlāpa*, Saterland Frisian *lope*, Frasch Frisian *luupe*, Standard Dutch *lopen*, Flemish *loopen*, Afrikaans *loop*, Standard High German *laufen*, Bavarian *laafa*, and Swiss German *laufe*. These all descend from Proto-Germanic **hlaupen* (OED, 2015), possibly with the meaning "to jump, spring, run". In this instance these words do imply use of the legs and feet for locomotion, but as it is uncertain whether or not the Proto-Germanic verb implied this, or referred to some other bodily motion, resulted in the choice to not include the Proto-Germanic item (from which Modern English *leap* has descended; *leap* was not included under any conditions due to this uncertainty as well); as Modern English *lope* is a borrowing from Old Norse and is not used with the meaning TO WALK or TO GO, it was not included at all. These *l*-forms were assigned the same code. Modern English *(to) walk* is the most neutral word with the meaning TO WALK (implying motion with the feet and legs); this is the result of the fusion of the Old English strong verb *wealcan* and the weak verb *wealcian*, both meaning "to knead" or "to full cloth";

the OED (2015) suggests that this semantic shift refers possibly to an old method of working cloth with the feet. The Modern English verb was assigned its own code. West Frisian has the word *rinne*, cognate with Modern English *run*, Standard High German and Dutch *rennen*, "to run"; this was assigned its own code. Another simulation was performed replacing the concept TO WALK with the more general TO GO; in this instance the *g*-forms above were retained, but cognate forms in the other languages were used as well, such as Old Frisian and Old Dutch *gān*, Old High German *gēn*, Standard High German *gehen*, Standard Dutch, Flemish and Afrikaans *gaan*, West Frisian *gean*, etc. These were coded as cognates[17]

For WARM (adjective) there were three forms in the data. The most common of these had the root structure /$C_{[+lab,-plo]}$Vrm/, where *r* represents a rhotic segment; this can be seen in Old English *wearm*, Old Saxon, Old High German, Old Dutch and Old Frisian *warm*, Old Norse *varmr*, Modern English *warm*, Ferring Frisian *warem* (with a prothetic vowel between the trill and the nasal), West Frisian *warm*/*waerm*, Saterland Frisian *woorm*, Frasch Frisian *wurm*, Low German, Standard Dutch and Flemish *warm*, Afrikaans *warm*, Standard High German *warm*, Bavarian *waam*, Swiss German *warm*, Danish *varm*, and Norwegian and Swedish *varm*. These are all ultimately descended from Proto-Germanic **warmaz* (Orel, 2003) or **warmo*- (OED, 2015). These are thus all cognates and were assigned the same codes. Faroese had the word *heitur*, and Icelandic had *hlýr*; these were assigned their own individual codes. In Gothic the adjective WARM is not attested, although the root *warm*- is found in the transitive verb *warmjan*, "to warm" (OED, 2015); however, to reduce the complexity of the simulations involved, it was decided not to include words which were of different word classes but were related; thus, this was not included in any simulations. The missing data item meant that its data slot was handled according to the Majority Wins, Infodelete and Infoclean strategies.

The concept WATER (noun) had cognate forms in all of the Germanic languages, with the root structure /$C_{[+lab,-stp]}$V$C_{[+alv,+obst]}$/, as in Gothic *wato*, Old English *wæter*, Old Saxon *water*, Old High German *wazzar*, Old Dutch *water*, Old Frisian *weter*, Old Norse *vatn*, Modern English *water*, Ferring Frisian *weeder*, West Frisian *wetter*, Saterland Frisian *Woater*/*Water*, Frasch Frisian *wååder*, Low German *Wåter*, Standard Dutch and Flemish *water*, Afrikaans *water*, Standard High German *Wasser*, Bavarian *Wassa*, Swiss German *Wasser*, Danish *vand*, Faroese *vatn*, Icelandic *vatn*, Norwegian Nynorsk *vatn*, Norwegian Bokmål *vann*, Swedish *vatten*. These are all cognates, and descend from Proto-Germanic **watnan*~**watnar* (Orel, 2003); the differences in the ends of these words, with the West Germanic words almost all ending in -*Vr* (in Received Pronunciation in Modern English, as well as many other varieties in the UK, Australia, New Zealand and South Africa, syllable final rhotics are no longer pronounced; in Bavarian the final rhotic has been lost), Gothic -*o* and Old Norse and its descendants -*n* is due to the selection of different formative endings (OED, 2015). The Old Norse and Gothic forms are in fact the same formative, revealed by the Gothic genitive

[17] It is thought that doublets such as *gangan* and *gān*, both meaning "go", may in fact indicate the presence of what were originally two separate verbs whose meanings were similar, and who underwent a partial merger, although the precise nature of this merger and these two words is uncertain (OED, 2015).

watins; a later sound change led to the Gothic nominative *wato* (OED, 2015). As the roots of all these words were cognates, they were assigned the same code.

The concept of WE (the first person plural pronoun) had two representative forms in the data. The most common of these had the root structure /$C_{[+lab]}V_{[-low]}$/ as in Gothic *weis*, Old English *wē*, Old Saxon, Old Dutch and Old Frisian *wī*, Old High German *wir*, Old Norse *vér*, Modern English *we*, Ferring Frisian *wi*, West Frisian *wy*, Saterland Frisian *wie*, Frasch Frisian *we*, Low German *wi*, Standard Dutch *wij*, Flemish *wij*/*wy*, Standard High German *wir*, Bavarian *mia*, Swiss German *miir*, Danish, Norwegian and Swedish *vi*, Faroese *vit* and Icelandic *við*. The reconstructed Proto-Germanic form given by Orel (2003) is **wēz~*wīz*. In the cases of Bavarian and Swiss German, where a bilabial nasal is found where other varieties have either a labiovelar glide or a labiodental fricative, a sound change took place in the Old High German period which led to the original labiovelar glide losing its velar quality and becoming a bilabial nasal instead (Rowley, personal correspondence; Salmons, personal correspondence); additionally, in Bavarian *r*-vocalisation led to the word-final rhotic becoming a vowel (Rowley, personal correspondence). The presence of a rhotic at the end of the Old Norse form *vér*, and an alveolar fricative at the end of Gothic *weis*, attests to the presence of Proto-Germanic **s~*z* in that position; this fricative was preserved in Gothic (Fortson, 2004) but came to be rhotacised in the ancestor of both North and West Germanic, where it was retained as a rhotic in Old Norse but lost in all of West Germanic except for Old High German. The Faroese and Icelandic forms have an alveolar stop and a voiced alveolar fricative syllable finally, respectively; this is due to the fact that both were originally the dual article, which came to be used later on as the default first person singular plural (OED, 2015), although the honorific *vér* still exists in Icelandic. As the base of both of these words is still Old Norse *vér*, it was decided to code them as cognates. The theoretical problems which could arise as a result of this will be examined in the discussion. Afrikaans was the odd one out with the pronoun *ons*, which is etymologically related to the first person plural accusative English *us*, Standard High German *uns*, Dutch *ons* (its direct ancestor), Swedish *oss*, etc. As this word is not derived from Dutch *wij*, and is an example of a semantic shift which generalised the accusative pronoun to all instances of the first person plural pronoun, it was assigned a separate code from the other data characters under WE.

The concept WHAT (interrogative) was represented by cognates across the entire data set. All of the words in question had the core structure /$C_{[+lab,-stp]}V(C_{[+intdent/alv,\ +obst]})$/, as in Gothic *hwa*, Old English *hwæt*, Old Saxon *hwat*, Old High German *waz*, Old Dutch *wat*, Old Frisian *hwet*, Old Norse *hvat*, Modern English *what*, Ferring, West and Frasch Frisian *wat*, Saterland Frisian *wät*, Low German *watt*, Standard Dutch, Flemish and Afrikaans *wat*, Standard High German and Bavarian *was*, Swiss German *waas*, Danish *hvad*, Norwegian Nynorsk *kva*, Norwegian Bokmål *hva*, Swedish *vad*, Faroese *hvat* and Icelandic *hvað*. The reconstructed Proto-Germanic word is **hwat* (Orel, 2003) or **χwat* (OED, 2015). These were all coded as cognates.

The colour WHITE likewise was represented by cognates throughout the data, all of the words having the form /$C_{[+lab]}VC_{[-vo,+stp]}$/ with regular sound correspondences, as in Gothic *hweits*, Old English and Old Frisian *hwīt*, Old Saxon *hwît*, Old High German *wīz*, Old Dutch

wīt, Old Norse *hvítr*, Modern English *white*, Ferring Frisian *witj*, West Frisian *wyt*, Saterland Frisian *wiet*, Frasch Frisian *wit*, Low German *witt*, Standard Dutch, Flemish and Afrikaans *wit*, Standard High German *weiß*, Bavarian *waiss*, Swiss German *wiss*, Danish *hvid*, Faroese *hvítur*, Icelandic *hvítur*, Norwegian Nynorsk *kvit*, Norwegian Bokmål *hvid*, and Swedish *vit*. The reconstructed Proto-Germanic word is **hwitaz* (Orel, 2003). These were all assigned the same code.

The interrogative pronoun WHO had four representatives across the data, the most common of which had the structure /(C)$C_{[+dist]}$V(C)/, as seen in Gothic *hwas*, Old English *hwā*, old Frisian *hwā*/*hwē*, Old Saxon *hwē*, Old High German *wer*, Old Dutch *wie*, Old Norse *hverr*, Modern English *who*, West Frisian *wa*, Low German *weer*, Standard Dutch, Flemish and Afrikaans *wie*, Standard High German *wer*, Bavarian *wea*, Swiss German *wèèr*, Danish *hvem*, Faroese *hvør*, Icelandic *hver*, Norwegian Nynorsk *kvem*, Norwegian Bokmål *hvem* and Swedish *vem*. These are all descended from a common Proto-Germanic word, reconstructed as **hwaz*~**hwez* (Orel, 2003) or **χwaz*~**χwez* (OED, 2015). These were assigned the same code. The postvocalic consonant was originally an alveolar fricative, as reflected by the Gothic form and used in the reconstructed one; in the North and West Germanic languages this fricative was rhoticised (Fortson, 2004); it was subsequently lost in all of the West Germanic languages except for Old High German (Fortson, 2004). The forms in Danish, Norwegian and Swedish, which have the bilabial nasal [m] word finally, are the product of apprehending the dative form of the interrogative as the normal form and subsequently using it without distinction as to the syntactic status of its referent (Hoekstra, personal correspondence). Ferring Frisian has the word *hoker*, derived from Old Low Franconian *hok*, from *hwelik*, "which' (Hoekstra, personal correspondence); this was assigned its own code. The Frasch Frisian word is *huum*; this is descended from the Old Frisian dative pronoun *hwam*, which has come to no longer be analysed as a dative pronoun, in much the same way as the continental North Germanic pronouns above (Hoekstra, personal correspondence). This was thus taken to more probably be a separate word, and was assigned its own code. The Saterland Frisian word *wäl* is a loanword from Low German *welk*, "which", with loss of the final velar stop (Hoekstra, personal correspondence). In addition to *weer*, Low German also has the word *wokeen* for WHO; this appears to be etymologically derived from *wolk*/*welk een*, "which one', but has come to mean WHO. This was thus used in an additional semantically strict simulation with its own code. It was decided to perform one simulation with these as cognates and another with them coded separately to reflect the different origins (loans from Low German and Old Low Franconian) or strategies underlying these forms.

The concept WOMAN had a number of word-forms attached to it, particularly in the older Germanic languages; many of these do not appear to reflect major semantic distinctions between them. It was therefore decided to focus on some of the more common forms attested in these languages. A widespread form had the root structure /$C_{[+lab,-stp]}$V$C_{[+lab]}$/, as in Old English, Old Dutch, and Old Saxon *wīf*, Old Frisian *wîf*, Old High German *wīb*/*wīp*, Ferring Frisian *wüv*, Frasch Frisian *wüset* (from Old Frisian **wüfshood*, with *hood* cognate with English *head* and having the same distributive function as in "head of cattle' (Hoekstra, personal correspondence)), Saterland Frisian *Wieuw*(*moanske*) and Bavarian *Weiberz*.

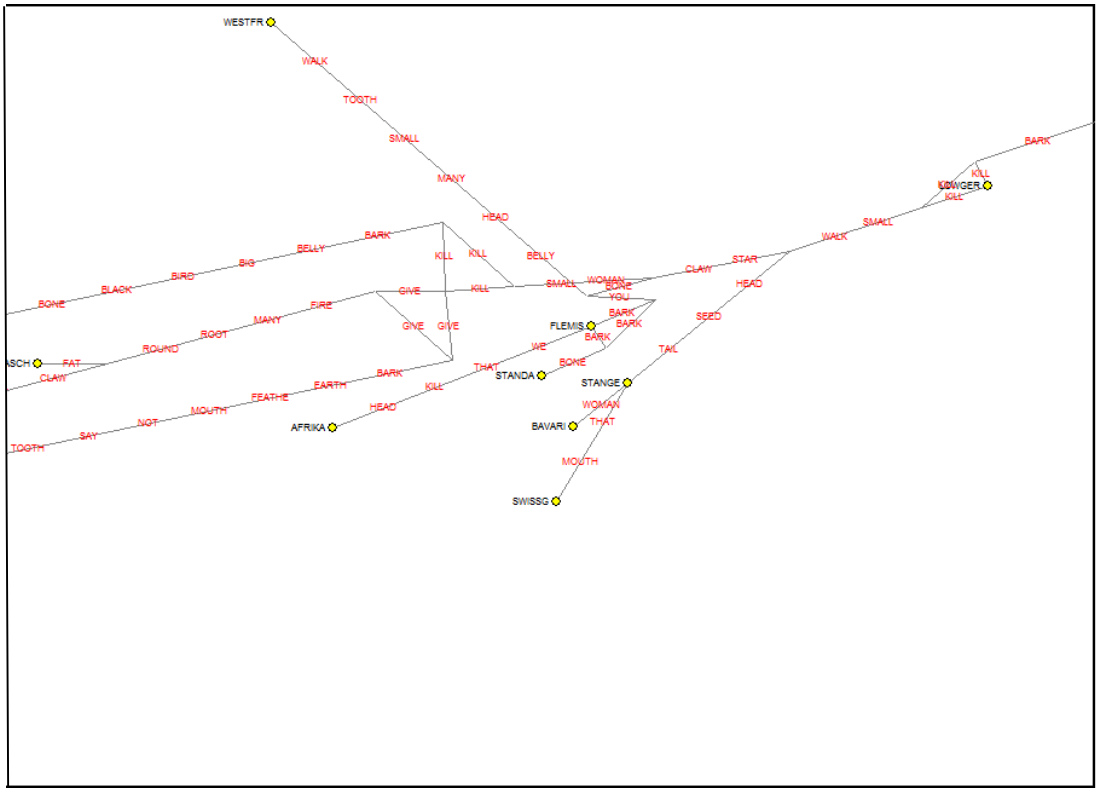
Additionally, Modern English has *woman* from Old English *wifmann*; as the modern word is the descendant of a compound with *wif* as one element, the word is technically a cognate, and was assigned the same code as the other *wif* forms. Orel (2003) reconstructs the ancestor of these forms as Proto-Germanic **wiþan*. These were all assigned the same code. A second widespread form had the root structure /k($C_{[+lab]}$)Vn/, as in Gothic *qino*, Old Norse *kona*, Danish *kvinde*, Faroese *kvinna*, Icelandic *kona*, Norwegian Nynorsk *kvinnfolk*, Norwegian Bokmål *kvinne* and Swedish *kvinna*. The instances in the North Germanic languages in which the cluster [kv] is found, despite there only being [k] in the Old Norse form, is due to selection in the daughter languages of a form in the genitive (Old Norse nominative singular *kona* but genitive plural *kvinna*) as the default form. These were all coded as cognates. Old English *cwēne* and Modern English *queen* were not included at all as the use of *cwēne* in manuscripts suggests it had already taken on a meaning more specialised then WOMAN; this has translated into the modern English meaning, and so these were treated as having had a different meaning altogether. Another form had the structure /f*r*V/, as seen in West Frisian *frou*, Low German *Fru*, Standard Dutch and Flemish *vrouw*, Afrikaans *vrou*, Standard High German *Frau*, and Swiss German *Frau*. These words are all cognates, and all ultimately stem from a word used in the continental West Germanic languages originally with the meaning "a noble woman" (GTB, 2015; Duden, 2015); this later had its meaning generalised to WOMAN. These words were assigned the same code. Bavarian additionally has the word *Frau*; this was used in an additional semantically strict simulation. Standard High German has the word *Weib*, although it has taken on a pejorative connotation and is no longer used as the neutral term for WOMAN (Duden, 2015); this was thus included in a semantically lax simulation.

The concept YELLOW (colour) was represented by cognates across all of the languages except Gothic, in which the word is unattested. The structure of these words was /$C_{[+vo,+vel/pal]}V_{[+mid]}$(l)(V)($C_{[+vo,+lab]}$)/, where the elements in parentheses are widespread but not universal in the data. This can be seen in Old English *geolu* (with palatalization of the initial **g* to [j] before the front vowel), Old Saxon *gelu*, Old High German and Old Dutch *gelo*, Old Frisian *gēl*, Old Norse *gulr*, Modern English *yellow*, Ferring Frisian *güül*, West Frisian *giel*, Saterland Frisian *jeel*, Frasch Frisian *göösl*, Low German *chääl*, Standard Dutch and Flemish *geel*, Afrikaans *geel*, Standard High German *gelb*, Bavarian *geib*, Swiss German *gääl*, Danish *gul*, Faroese *gulur*, Icelandic *gulur*, and Norwegian and Swedish *gul*. These are all ultimately continuations of Proto-Germanic **gelwaz* (Orel, 2003). These were all assigned the same code. The missing Gothic item was handled using the Majority Wins, Infodelete and Infoclean strategies.

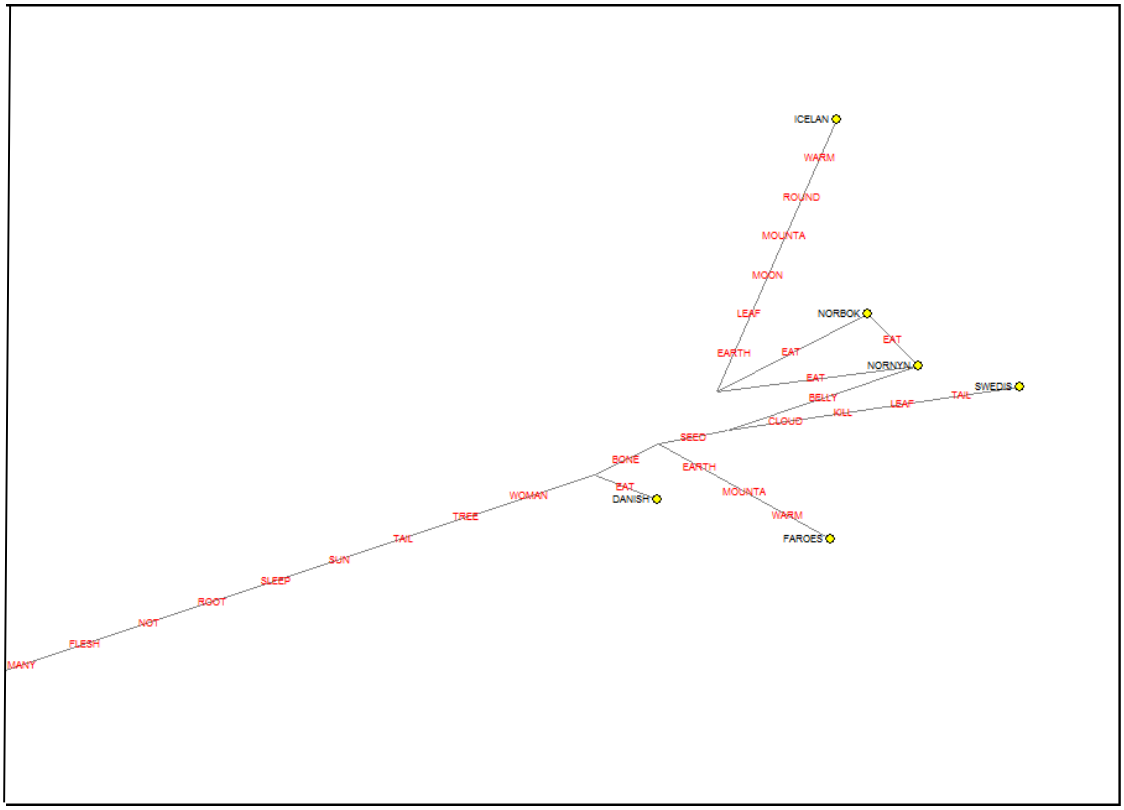
Appendix B

Close-Ups of Parts of Select Diagrams

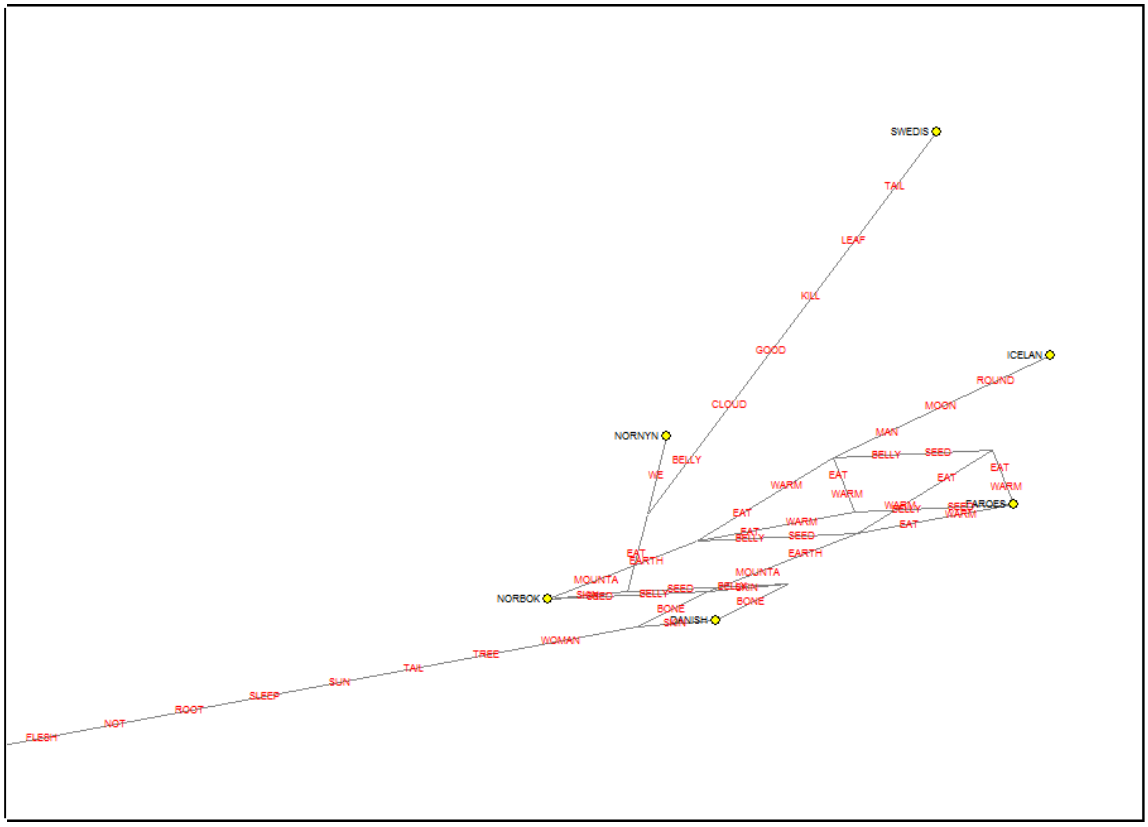
Subgroup 2:



Modern Germanic Languages, Semantically Strict Simulation, ε=0 Language n=18, Concept n=100

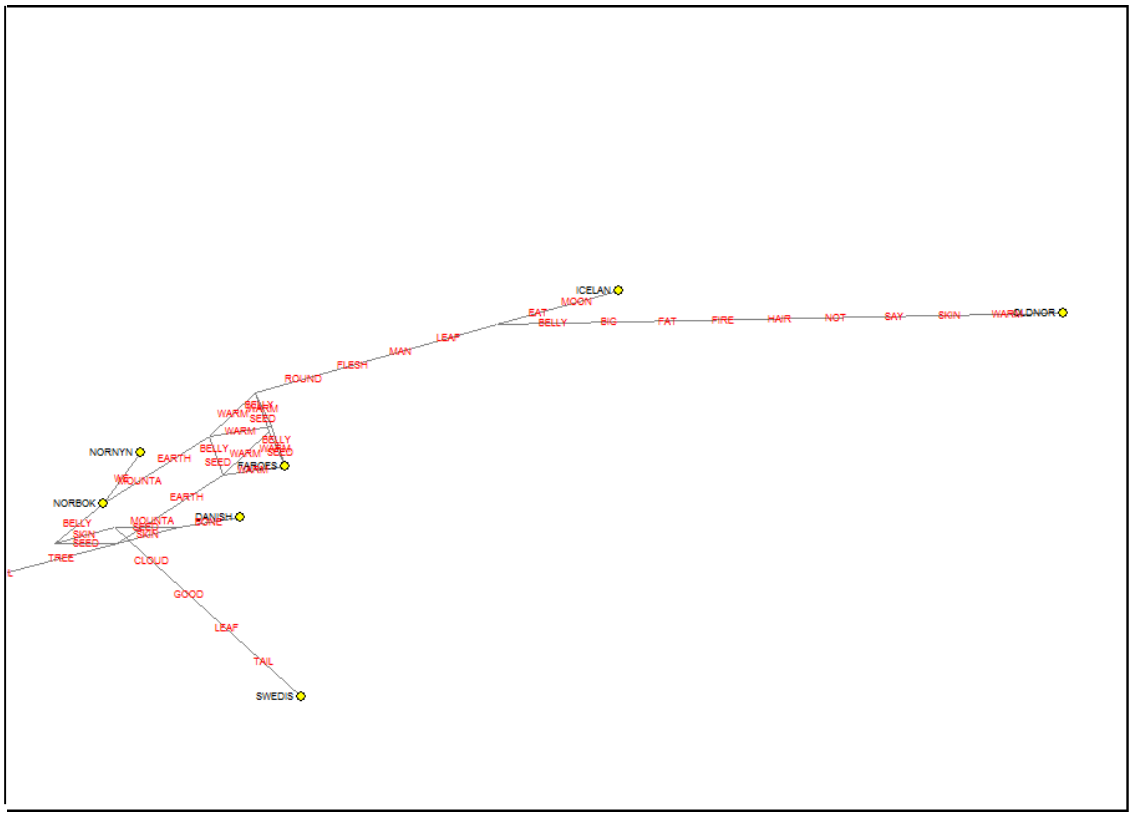


Modern Germanic Languages Semantically Strict Simulation, ε=0 Language n=18, Concept n=100



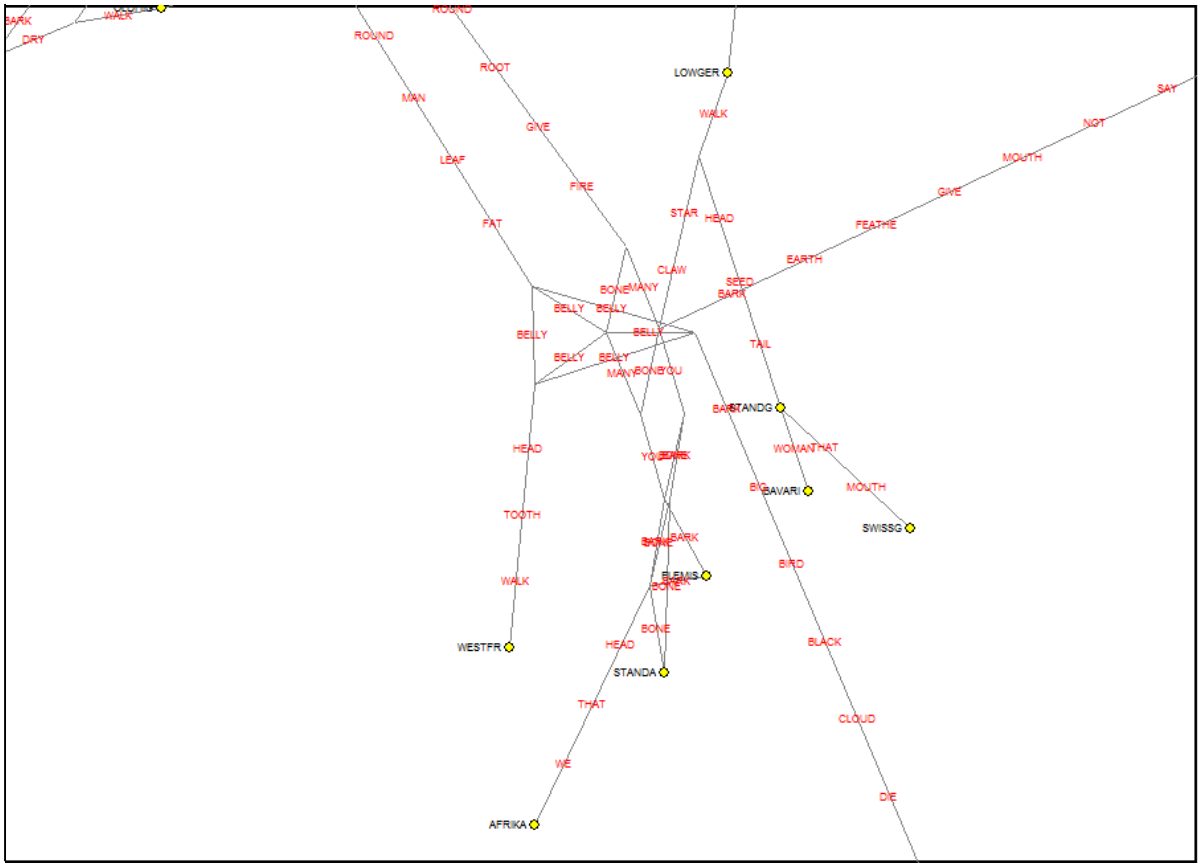
Modern Germanic Languages Semantically Strict, Other Words Simulation, ε=0
Language n=18, Concept n=100

Subgroup 3:

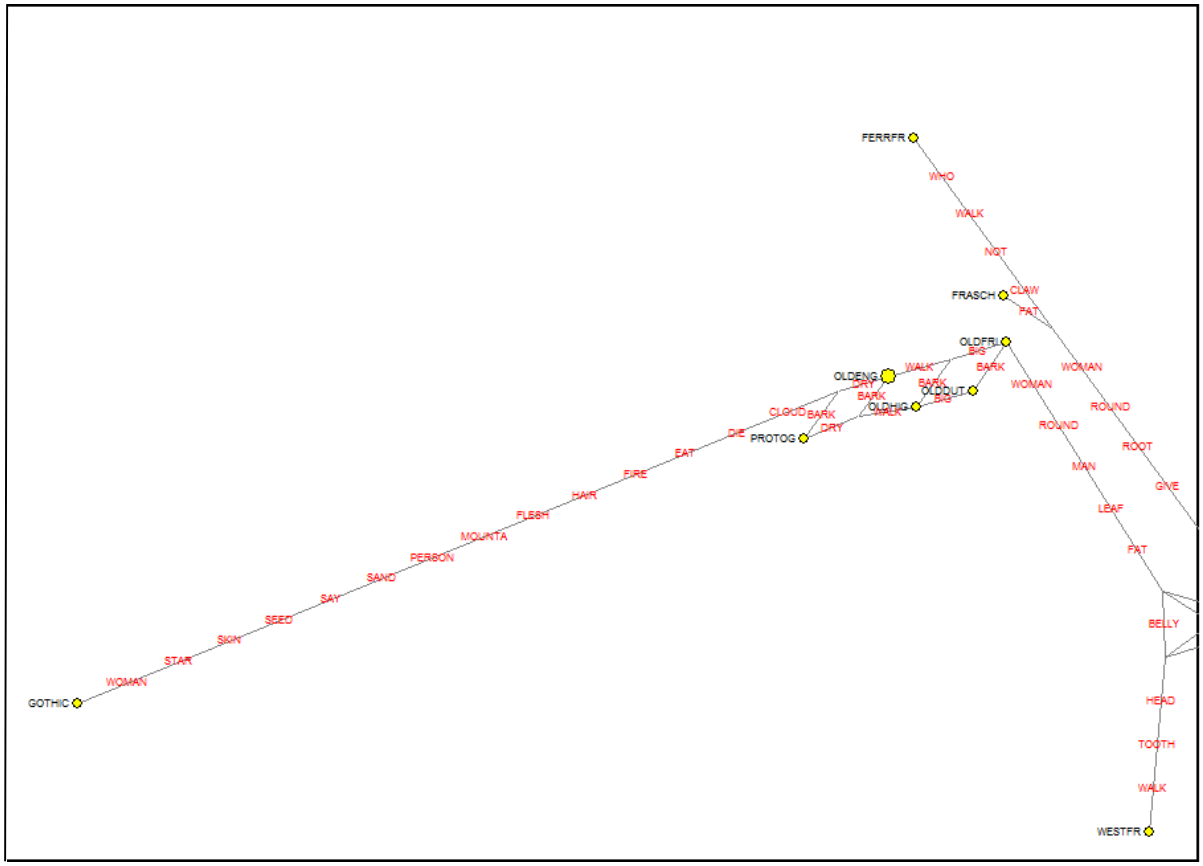


Old and Modern Germanic Languages, Majority Wins, Additional Words, Semantically Strict Simulation, ε=0 Language n=25, Concept n=99 (Note, different angle from Results diagram)

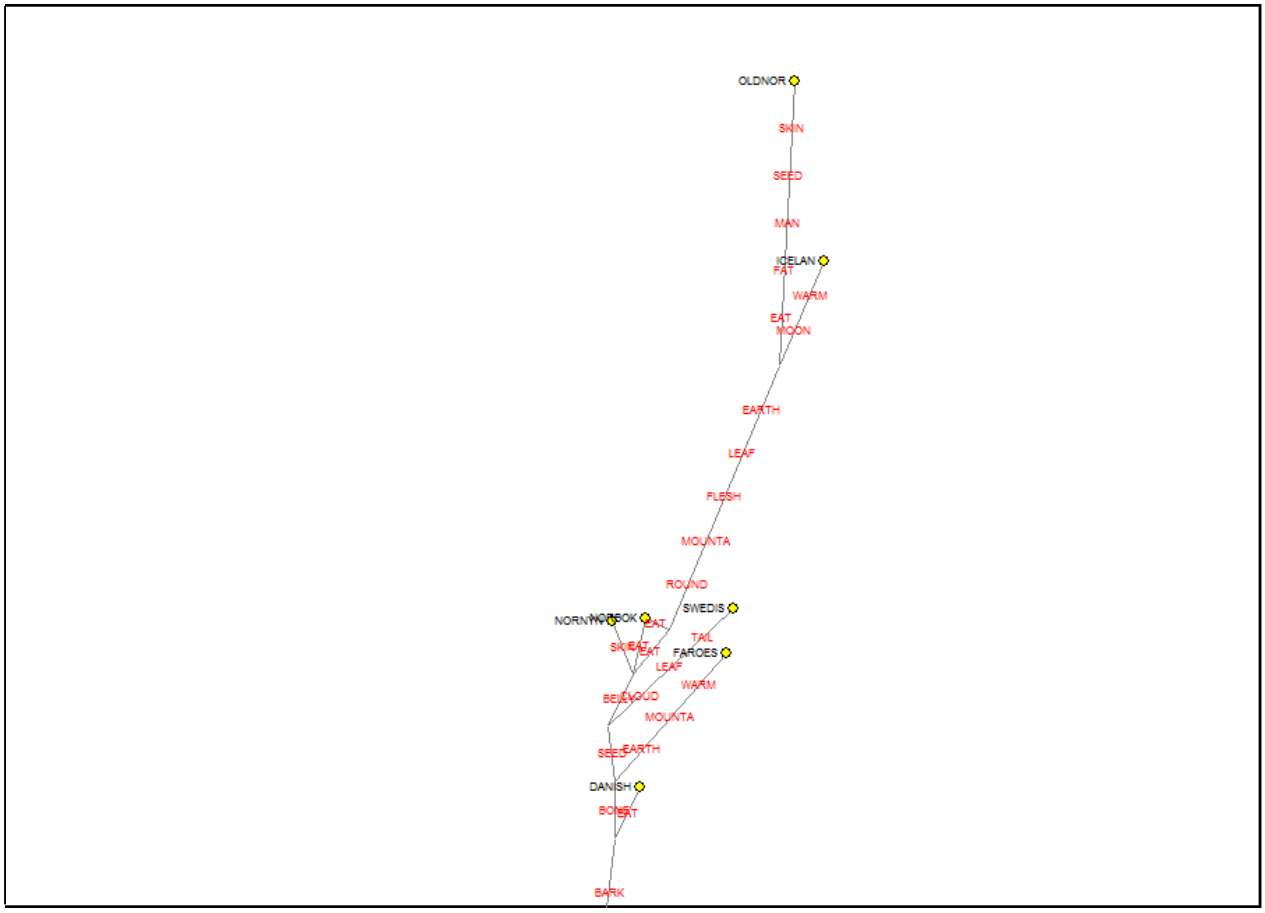
Subgroup 4:



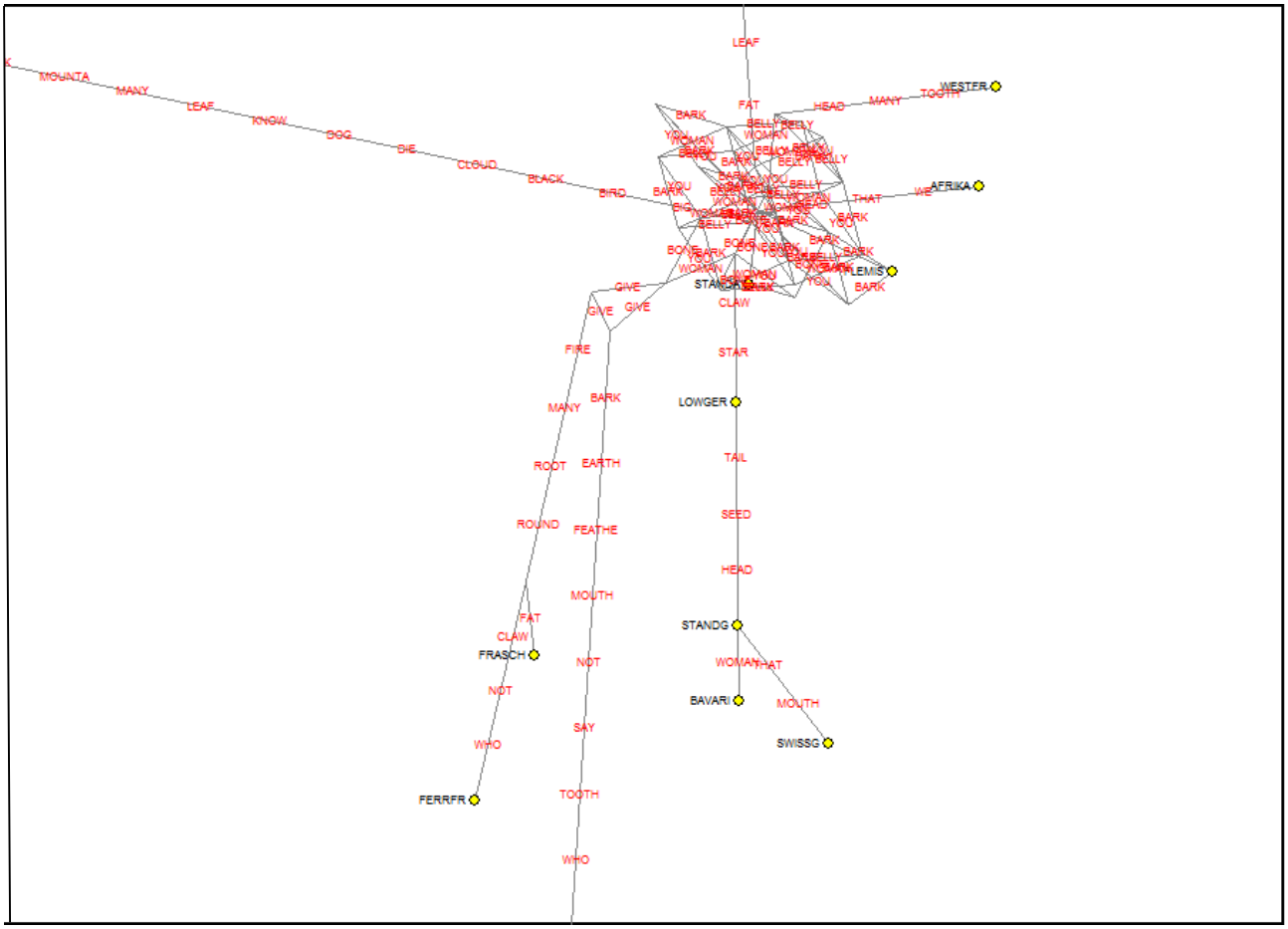
Old and Modern Germanic Languages, with Proto-Germanic, Word Choice One, Majority Wins, Semantically Strict Simulation, ε=0, Language n=26, Concept n=99



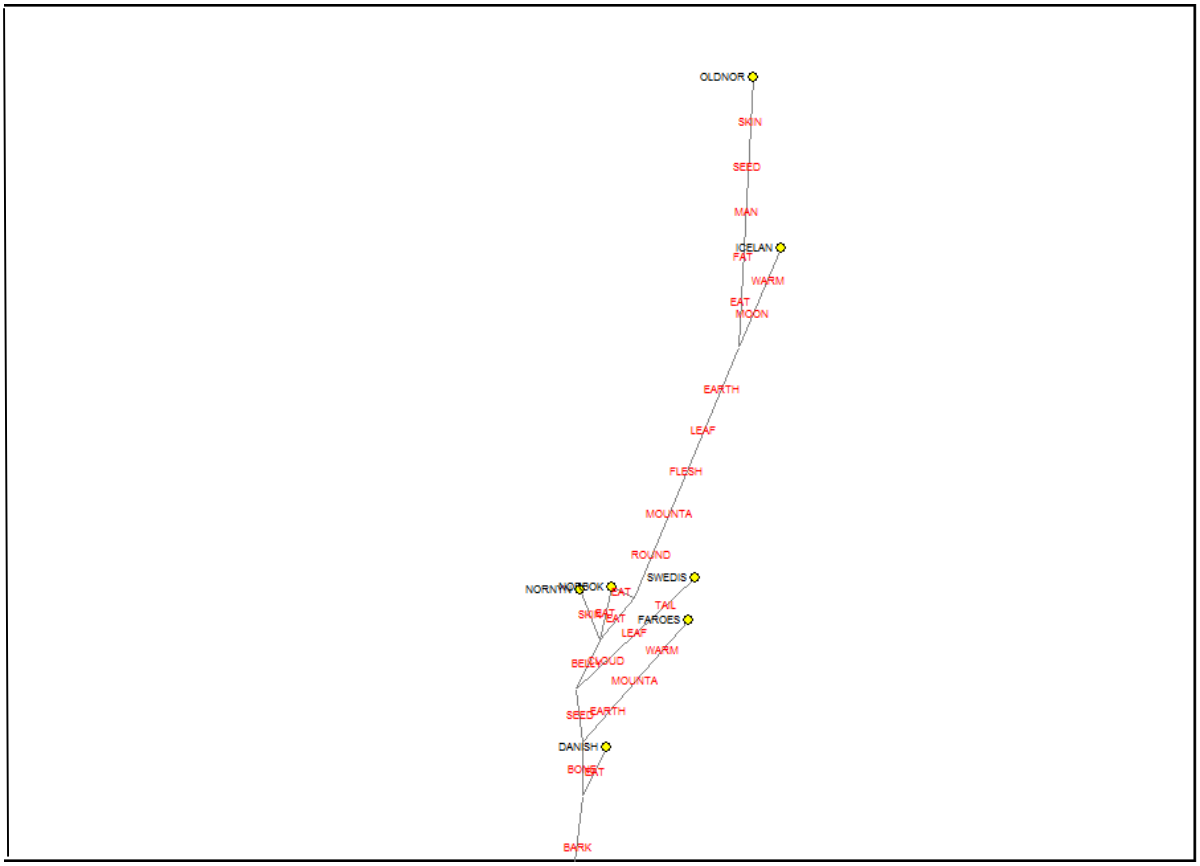
Old and Modern Germanic Languages, with Proto-Germanic, Word Choice One, Majority Wins, Semantically Strict Simulation, ε=0, Language n=26, Concept n=99



Old and Modern Germanic Languages, with Proto-Germanic, Word Choice One, Majority Wins, Semantically Strict Simulation, ε=0, Language n=26, Concept n=99



Old and Modern Germanic Languages, with Proto-Germanic, Word Choice Two, Majority Wins, Semantically Strict Simulation, ε=0, Language n=26, Concept n=99



Old and Modern Germanic Languages, with Proto-Germanic, Word Choice Two, Majority Wins, Semantically Strict Simulation, ε=0, Language n=26, Concept n=99