Motifs and Folktales: A New Statistical Approach.

Julien d'Huy

Motifs and Folktales: A New Statistical Approach

Julien d'Huy, Institute of the African World (IMAF), Paris I Sorbonne

On his way home from Troy, Odysseus and his twelve ships are captured by the cyclops Polyphemus while visiting his island. The monster moves a massive stone to cover the door of the cave so that the men cannot escape, and then he eats many of them. The hero brings the cyclops a barrel of wine and says his name is 'Nobody'. When the cyclops falls into a drunken sleep, Odysseus and his men blind the monster with a wooden stake. The monster calls for his brothers, who come, but they leave when they hear that 'Nobody' has caused the harm. Later, Odysseus ties himself and his men to the bellies of sheep and they escape, despite the blind Polyphemus feeling the backs of animals to ensure that the men are not getting out with his herd.

This famous story of Homer has been recorded in modern times among the folklore of many widely separated European groups (Hackmann 1904). In some variants, the giant tries to recapture the man using a magic ring that raises alarm and reveals where the fugitive is. The man needs to cut off his finger to escape.

Stith Thompson (1961) numbered five traditional elements or motifs in this tale-type:

> G100. Giant ogre, Polyphemus
> K1011. Eye-remedy. Under pretence of curing eyesight the trickster blinds the dupe. (Often with a glowing mass thrust into the eye.)
> K521.1. Escape by dressing in animal (bird, human) skin
> K602. "Noman"
> K603. Escape under ram's belly

Uther (2004) adds five additional motifs:

> F512. Person unusual as to his eyes.
> F531. Giant.
> K1010. Deception through false doctoring.
> K521. Escape by disguise.
> D1612.1. Magic objects betray fugitive. Give alarm when fugitive escapes.

The term *motif* has commonly been used by folklorists to refer to distinguishable and consistently repeated story elements used in the traditional plot structures of many stories and folktales. Stith Thompson developed a *Motif-lndex of Folk-Literature* (1932–1936; revised and enlarged second edition appearing in 1955–1958). However, Thompson's (1955: 7) criteria for identifying and delineating motifs were unsophisticated: "It makes no difference exactly what they are like; if they are actually useful in the construction of tales, they are considered to be motifs." The present analysis raises a number of questions related to the identification of motifs and the assessment of their uniformity and coherence as socially and historically circulating narrative elements. Here, I will explore potential tools and methods that may enable researchers to control these assessments in a statistical and more objective way. Developing a systematic means of doing this would be new and potentially very useful.

Applying different software to textual corpora in order to identify narrative elements like motifs presents a number of methodological issues. Once the sample corpus used here has been introduced, I will thus discuss the application of two different software programs. First, the corpus was treated using Treecloud. Treecloud was applied with the hypothesis that the software had the potential to present evidence of narrative motifs when applied to a textual corpus of variants as raw data. Rather than a positive outcome, this pilot study instead produced information that illustrates a number of problems that arise when using software of this type. The program Iramuteq 0.6 alpha 3 was then applied to the same corpus with an ability to account for additional parameters in the data-set. The ways to identify motifs with Iramuteq are also tested on both the raw text of the corpus and also with tagging of essential elements. With regard to identifying motifs in the classic sense of Thompson, the use of these programs proved better suited to identifying certain motifs rather than others, and the pilot studies show that these and similar programs are not well-suited to identifying motifs in a text corpus. These pilot

studies therefore have significance for revealing methodological problems in the use of software for narrative analysis that may help point the way to future innovations. More significantly, the analyses had an unexpected outcome of providing a new model for approaching tales in terms of semantic networks of elements. Rather than revealing 'motifs', the pilot studies present new ways of looking at tale-types.

The Test Corpus

In order to test the hypothesis that software could identify the main topics of a tale-type on the basis of the lexical surface of a text, we chose to analyze the tale-type of Polyphemus (AT 1137) on the basis of the test-corpus of the 36 versions of the tale published *in extenso* in English in a chapter of James Frazer's *Apollodorus: The Library II* (1921: 404–455). The text of each narrative was embedded in Frazer's critical introduction and conclusion, which have been excluded from the present analysis. The fact that these tales were translated (when necessary) by the same scholar allows a relative uniformity in the text's lexical field. It is assumed that the analysis will work best when using material translated by the same individual. The impact of lexical variation according to translator and its implications of the tree produced from the data would be worth exploring.

It must be acknowledged that the examples collected in *Apollodorus* are drawn from diverse published sources that Frazer had available. These texts were not selected according to modern source-critical standards. Some of these source texts have potentially been subject to significant editing for the earlier publication, or they may reflect summaries and paraphrases of, for example, early 19th century scholars. In addition, Frazer's translations are, in a number of cases, based on, for example, earlier German translations of the narrative from another language – although Lévi-Strauss claims that a mythical message is preserved even through the worst translation (Levi-Strauss 1958: 232). Additionally, it is not clear that Frazer was interested in critically reflecting the lexical field of these sources rather than the narrative content, and especially that narrative content which he considered relevant for comparative discussion. The lexical field of his translations is nevertheless anticipated to be more uniform than texts by multiple translators would be, and this thus increases the probability that the pilot study will yield positive results. Consequently, in terms of the international Polyphemus tradition, the findings of this pilot study necessarily remain conditional on the quality of the data to which the software is applied. Methodological issues surrounding the source-critical quality of source-texts and translations in a data-set remains distinct from the focus here, which is on the potential of Treecloud and Iramuteq 0.6 Alpha 3 as methodological tools in the motif analysis of a body of texts.

The Treecloud Pilot Study

The software Treecloud (Gambette & Veronis 2009) allows the most frequent words of a text to be arranged on a tree that reflect their 'semantic proximity', i.e. the co-occurrence of distinct semantic elements according to the text. The size and the color of each word reflects its individual frequency. The length of the path between two words in the tree represents the distance between them on the basis of their linear word proximity (i.e. analyzing the text as a linear sequence in which each semantically tagged word equates to one unit of distance). Such a tool may help to identify the main topics of a tale-type on the basis of recurring concentrations of words associated with plot patterns. This software analyzes the lexical surface of texts, and therefore the analysis of multiple texts is subject to a degree of language dependence. In addition, variation in the lexical surface (e.g. owing to synonymy, phraseology or alternation between common noun and an agent's proper name) are not accounted for by the software. Yet, it may offer a general path for motif analysis.

For the purposes of this initial pilot study, no attempt was made to tag texts' lexica according to number or categories of semantic equivalence. This avoided the possibility that the researcher-interpretation might conflate narrative elements which otherwise maintained patterns of use associated with certain motifs and not others (e.g. 'ram' and

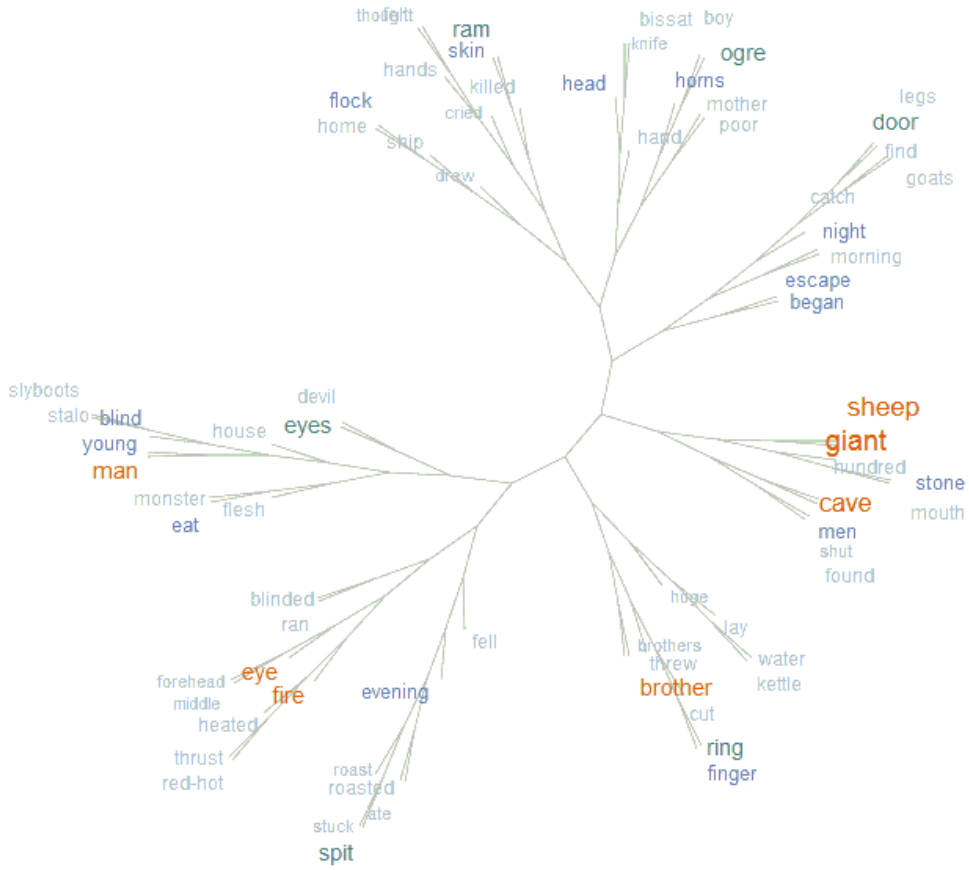

Figure 1. Treecloud analysis of our corpus.

'sheep', or 'ram', 'sheep', 'goats' and 'flock'). Of course, the critical standards of the source texts and their stadial translation have already problematized the validity of such distinctions, but they will remain distinguished here on a methodological principle for the present test. Similarly, nominal designations for agent roles were not tagged to standardize within a text (e.g. 'monster' and 'giant'), nor across different texts (e.g. 'giant', 'devil' and untranslated vernacular beings such as Sámi *stalo*). The decision 'to tag or not to tag' agentive roles across texts from different cultures presents methodological issues that shape the outcome of the analysis in either case. On the one hand, not doing so avoids the problems already mentioned, and the effect of treating only agentive roles in this way could have unpredictable consequences for the data. On the other hand, not doing so may make motifs associated with redactions linked to certain terms more observable, but also affect the reflection of the common agentive role identifiable with Homer's Polyphemus in connection with clusters of other semantic elements within the data-set as a whole. As an initial test, it was here preferred to analyze the test corpus with as little impact from the present researcher as possible and as a more automated outcome of applying the tool to the raw data. Further work will be able to identify and measure the fluctuation due to the standardization – if any – of the vocabulary.

Treecloud was first used to explore our data with the following parameters: english stoplist, NJ tree, number of words: 75; width of the sliding window: 20; distance: Jaccard's co-occurrence. 'Neighbor joining' is a bottom-up clustering method for the creation of phylogenetic trees (NJ tree); the branch lengths as well as the topology of a parsimonious tree can quickly be obtained by using this method. I retain the 75 most frequent words from which the tree is formed (number of words: 75). This portion studied corresponds to a sliding window, with a width of 20 words and 1 as the size of the sliding steps between two consecutive windows (width of the sliding window: 20). The

statistical tool of the Jaccard index is used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. The Jaccard distance measures dissimilarity between sample sets and is obtained by subtracting the Jaccard coefficient from 1. Results were found to be robust to changes in co-occurrence distance formula and number of words; all distance formulas perform approximately equally well and the number of words including in the sliding window did not affect our results.

The program reflects the relative frequency of words by the font size and color in which these appear in the tree, where larger words can be inferred to be more significant. The result is shown in Figure 1, which reveals five primary clusters. These five groups match to varying degrees with motifs identified by Thompson and Uther.

Cluster (1) presents the prominent words 'skin', 'killed', 'ram' and 'flock', as well as 'ship' and 'hands' in a lighter colour and smaller size. This cluster appears to correlate quite strikingly to the hero's escape under the skin or the belly of a ram and can be interpreted as referring to the relevant narrative sequence (motifs K521, K521.1. and K603).

In cluster (2), 'cave', 'giant' (motif F531 and the largest word of the cluster) and 'sheep' are prominent, with 'stone' and 'men' somewhat lighter and not far from 'escape', 'begin' and 'door'. The words 'begin' and 'sheep' appear ambiguous, but this cluster otherwise appears to correlate strongly with the narrative element of the cave of the giant as a locked place from which human beings want to escape. Between cluster (1) and cluster (2), it is also interesting to observe that the words 'ram' and 'flock' group separately from the word 'sheep', which may be associated with this cluster as the word most frequently co-occurring with the giant's activity of housing or caring for the animals, or opening the cave to let out the animals.

In cluster (3), the words 'ring', 'finger', 'cut' and 'kettle' all correlate with the episode concerning the magic ring. The word 'brother' appears peculiar here. This could suggest a correlation between the ring episode, described as motif D1612.1 Magic Objects Betray Fugitive, and the presence of brothers at the beginning of the story. This correlation appears centrally in Frazer's sixth and seventh examples (Frazer 1921: 415–418) and a 'brother' is mentioned in connection with a magic ring possessed by the giant in a different motif (i.e. not motif D1612.1) in a medieval example from *The Book of Dede Korkut* (Frazer 1921: 453, 455). Frazer's sixth example refers to characters as 'brother' rather than 'man' etc. or a personal name, while the seventh treats 'Little Brother' as the name of the protagonist. Together, these two versions account for the majority of occurrences of the word 'brother' in the data set and have led to its disproportionate association with the motif with the ring.

Cluster (4) centers on the words 'fire' and 'eye' with a number of minor words. The subgroup 'middle', 'forehead' and 'eye' points to the unique eye of the monster found in several variants. The subgroup 'fire', 'heated', 'roast', 'roasted', 'red-hot', 'spit', 'stuck', and 'thrust' correlate with the weapon used by the protagonist (a spit or hot water), while the words 'eye' and 'blinded' correlate with the weapon's target and the resulting blindness of the giant. This cluster presents a striking correlation with narrative elements of the tale-type; the words 'eye' (used in the singular), 'forehead', 'middle' involve F512, i.e. 'person unusual as to his eye'. 'Heated', 'red-hot', 'roast', 'roasted', and 'thrust' point toward the 'glowing mass thrust into the eye' of K1011; 'spit' is included in both K1010 (deception) and K1011 (blinds the dupe); finally, 'blinded' involves K1011.

Cluster (5) centers on 'monster', 'eat', 'flesh', 'devil' and 'man'. This cluster can be correlated with the fact that the monster eats human flesh. Again it is noteworthy that 'giant' appears associated with cluster (2), separate from 'monster' and 'devil' here as well as from 'ogre' in still another group, although all of these fill the same role or function in a set of motifs in this tale-type or might be identified with motif F531 Giant. Note that a bias can occur because Treecloud will not, for example, identify a giant at the

level of narrative content if it is identified through description (e.g. as a man many times the size of other men) without using a word that is tagged as indicating 'giant'.

'Giant' is linked to the sealed cave, with the word 'sheep' potentially linked to the motif of releasing the animals from the sealed cave while blind. Similarly, the word 'eyes' appears here at the root of this cluster while the singular 'eye' appears in cluster (4) linked to blindness and the motif of blinding in the form 'blinded', while here the form 'blind' appears. The plural 'eyes' may cluster with 'devil' as one of the only monsters that possesses two eyes. A second analysis could reveal that this cluster groups more particular data that only occur in some versions of the story (such as '*stal*o', 'Sly-Boots', 'devil').

The overall impression of the result is that the lexical surface of the examples analyzed in this way does produce some evidence of motifs. However, this statement must be nuanced: motif words associable with G100 Giant Ogre, Polyphemus, are divided between the first and the fifth clusters. Moreover, 'ogre' appears as an important word between the first and the second clusters, noting that this term, however, is only used in Frazer's examples twenty-one, twenty-four and twenty-five and then once in the translation of the example from *The Book of Dede Korkut*. Motif K602 "Noman" does not appear, nor is this motif mentioned in Uther's revised classification. However, the software would only reveal the presence of this sort of name-disguise if *a*) multiple texts used the same 'No Man' / 'Noman' / 'No One' / 'Nobody' as a word, and *b*) the name would be recurrent within a text rather than only used once.

Viewed uncritically from the perspective of broad motifs mentioned by Thompson and Uther in their descriptions of the narrative, statistics provide a largely positive correlation between motifs (although sometimes as groupings of motifs) and the clusters of lexical items (85% of the whole motif; 75% when we take into account and delete the duplicated motifs F512, F531, K1010, K21). Lexicometic tools could potentially open up new areas for research and may be able to reconstruct large numbers of motifs automatically.

This pilot study also reveals certain problematic aspects of the use of Treecloud. First, exceptional features of certain narratives may significantly impact the lexical surface of individual examples, producing a concentration of a particular word. This is the case with 'brother' in cluster (3), where two examples account for significantly more than half of the examples of the word. This concentration appears directly connected to the appearance of 'brother' as a high-frequency word in the overall corpus and also offsets the relative frequency with which 'brother' co-occurred with other narrative elements by linking it especially to those elements prominent in the two particular examples. In this case, the word 'brother' was seen as linked to the motif of the magic ring (or D1612.1), which was prominent in those two examples but was not a motif found throughout the corpus. This type of problem can be moderated in the future by increasing the number of examples of the tale studied. The number and the way to treat multiple cultures and periods remains to be investigated. Nevertheless, it also highlights that information generated by applications of the software cannot be taken at face value and it is the responsibility of the researcher to consider the information in dialogue with the material being analysed.

Another problematic aspect of this use of the lexicometric software is that it reveals only the highest degree of co-occurrence of each word singly throughout the whole diagram. When words are equated with semantic elements and the clusters of elements are identified with motifs, this means that any single word can only appear in one cluster, and thus any single element will only be correlated with one motif in the whole of the narrative. The relevance of words to multiple clusters is highlighted by the distribution across clusters of words that can be considered synonyms or potentially equivalent variations in different versions, such as 'blind'–'blinded', 'eye'–'eyes', 'man'–'men'–'boy'–'brother', 'giant'–'ogre'–'monster' and so forth. When compared with the example of 'brother' above, the clusters in which some of these terms appear may potentially also be influenced by unusual uses

in the lexical surface of a few particular texts. This dispersal would be eliminated if all of the terms for the monster were tagged 'giant', but this would also consolidate that role as appearing linked to only a single cluster. This difficulty can also be linked to the issue of motifs as narrative elements. For example, cluster (2) appears associated with the men trapped in the cave by the giant. The word 'mouth' appears here owing to the recurrent expression 'mouth of the cave'. This appears equivalent to 'door' in the adjacent cluster, where 'legs' is found, linked to the door by the motif of the giant letting his sheep and goats out of the cave between his legs. Yet 'giant' appears with 'sheep' and 'cave' while 'legs' appears with 'door' and 'goats'. These two clusters could be interpreted as reflecting narrative elements of the trapping of the men and their escape, respectively. However, it becomes questionable how accurately individual clusters may represent motifs if some of their key elements do not appear linked to them because their relative frequency is slightly higher in connection with a different cluster. Treecloud effectively reduces the whole lexical surface of the corpus into exclusive clusters of elements. What it does not do is reveal the concentrated open clusters of co-occurring elements recurrent through the corpus which would enable, for example, acknowledging multiple clusters in which 'giant' was a key element.

The Iramuteq Pilot Study

Iramuteq 0.6 alpha 3 (Ratinaud 2009; 2012; see additional material in Schonhardt-Bailey 2013) allows for statistical analysis of the corpus text (width of the sliding window: 40; for a good synthesis). The classification done by this software is based on lexical proximity and the idea that words used in similar contexts are associated with the same lexical and mental worlds. Iramuteq assumes that as the speaker speaks, he is investing in a succession of different worlds, which each successively impose their properties and a specific vocabulary. The software could also be very useful for the reconstruction of the successive 'lexical worlds' that a folktale teller successively inhabits. By classifying together the co-occurring words, we may understand what semantic territories were behind the construction of the observed folktales.

Each text of the corpus (all the different texts collected) was individualized during the lexical analysis (vs. other software like Treecloud, which treats all the texts together). This individualization accounts for an additional variable in the analysis. This method also eliminates the largest bias of over-represented words that may exhibit a remarkably high frequency in only a few texts and thereby off-set the data, such as 'brother' (59 occurrences; see Figure 2).

Iramuteq software constructs a dictionary of 'lexical forms' which are lemmatized, i.e. Iramuteq automatically reduces words to their root forms and grammatical classification to eliminate function words. This includes the conversion of verbs to their infinitive, plurals to singular, and so forth. The lemmatization deletes the impact of synonyms terms such as 'blind'–'blinded', 'eye'–'eyes', 'man'–'men' and to some degree makes the lexical field more uniform. This is already an advantage of Iramuteq over other pieces of software such as Treecloud.

Table 1. Number of the most widespread occurrences (only nouns and verbs) in the untagged and tagged corpus. In the data set "Tagged texts 1", 'devil', 'ogre', 'stalo', 'monster', 'cyclops', 'Basa-Jaun', 'Tartaro' and 'Depe Ghoz' have been tagged as 'giant'; 'ram', 'flock' and 'goat' have been tagged as 'sheep'; 'hatchet' as 'ring'; and 'myself' as 'nobody'. The data set "Tagged text 2" differs from Tagged text 1 by not tagging 'cyclops' as 'giant' and tagging 'one-eyed' as 'cyclops' (which appears reflected in the number of occurrences of 'eye' in the present table).

Lexical unit	**Instances in Corpus**		
	Untagged	**Tagged text 1**	**Tagged text 2**
giant	184	312	305
eye	110	172	172
sheep	109	102	102
man	84	84	84
cave	73	73	73
fire	60	60	60
brother	59	59	59
ogre	53	= 'giant'	= 'giant'
day	48	48	48
skin	47	47	47
eat	46	46	46
find	46	46	46
ring	45	45	45

Each text is cut into segments. The segmentation is automatically obtained as sentences or parts of sentences cut by natural punctuation and sometimes as somewhat larger units made by the concatenation of several succeeding sentences. Within each segment, the software maps the distribution of the forms selected by the researcher for analysis (nouns, verbs, etc.). The results are then collated and brought together to be analysed. The software aims to cluster forms according to similarity and differences in the distribution of the vocabulary. The analysis is based on a series of bi-partitions calculated from the binary table (presence / absence) crossing lexical forms and segments. The set of partitions that maximizes the inter-classes inertia leads to the first set of partitions. Then the software tests whether each unit is exchangeable from one class to another to control the robustness of the result. After all the text segments have been partitioned into two classes, the algorithm repeats the operation at every step for the larger of the remaining classes until the required number of iterations have been done.

When applying Iramuteq to the corpus, I first applied the software to the raw, untagged text, and then to the corpus with lexica tagged according to number or categories of semantic equivalence. It should be noted that the Iramuteq software's distinction of each text as

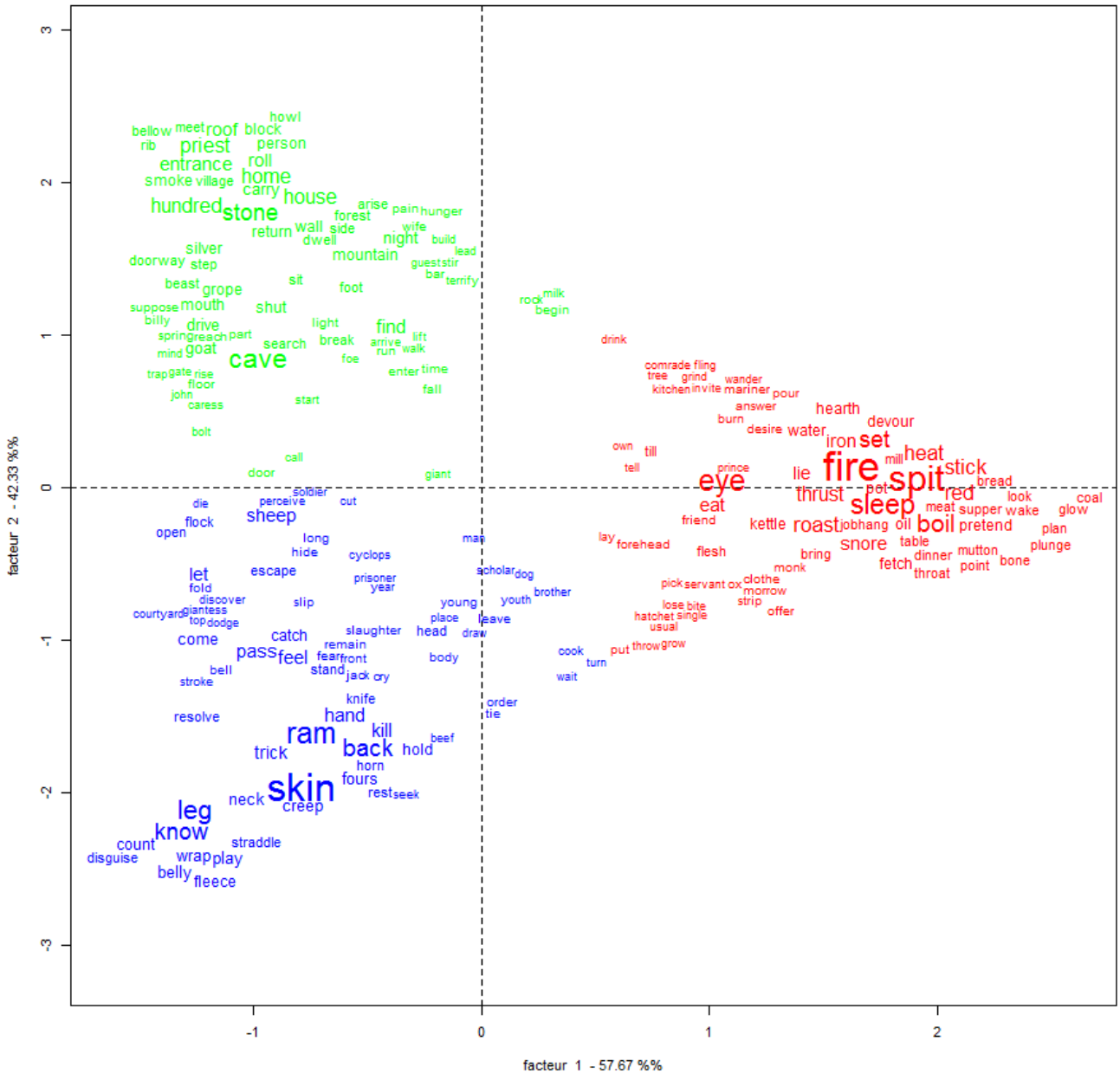

Figure 2. Principal Component Analysis of the untagged corpus.

a factor in analysis has implications, especially for analyzing agent roles. When the term for the adversary is consistent within each text but varies between texts, that agent will not appear to the software as consistently co-occurring with other elements of a motif. In other words, the variation between texts could produce interference in the data so that the agent would not appear as an element of the motif. To test for this problem, I therefore analyzed the sample corpus with both untagged and tagged texts in order to identify and measure the fluctuation – if any – resulting from this standardization of the vocabulary.

Iramuteq Correspondance Factorial Analysis

The most common nouns and verbs have been classified with a correspondence factorial analysis (method GNEPA, formerly called ALCESTE; factor 1: 57.67%; factor 2: 42.33%). This factorial analysis is based on calculations of inertia (or of variance) – i.e. of differences between the classes. It specifically reveals the contrasted use of vocabulary in the different lexical groups and the proximity of lexical items inside each of them.

The default options values of the program were maintained (size of rst 1 = 12; size of the rst 2 = 14; number of terminal classes during the first phase 1 = 10; minimum number of segments of text per class = automatic; minimum frequency of an analysed form: 2; maximum number of analysed form 3000; method of the singular value decomposition: irlba). The classification obtained is based on lexical proximity; it is not a matter of counting occurrences, but of relations among words, consequently 'giant' does not appear prominently in the principal component analysis shown in Figure 2 (untagged text), although it is the noun most frequently appearing in the corpus (184 instances including all morphological inflections), nor does 'giant' appear in Figures 3 or 5, which show the result with tagged text. Instead, the word appears floating in the center of the multiple groupings. It should be noted immediately that, as in the Treecloud analysis, each element occurs only once in a diagram, which means that semantic clustering of elements in any one group necessarily requires their exclusion from other groups.

Iramuteq's Principal Component Analysis found three classes, covering 38,4% (red), 32,3% (green) and 29,3% (blue) of segments in the untagged text, as seen in Figure 2. The first lexical group includes essentially the house of the giant and its lexical field ('home', 'house', 'stone', 'wall', 'mountain', 'arrive', 'enter', etc.) and the lexical field of the village ('house', 'home', 'entrance', 'smoke', etc.) Neither Thompson nor Uther address this as a motif. The second lexical group can be associated with the moment when the giant is blinded ('fire', 'spit', 'sleep', 'eye', 'roast', 'boil', 'forehead', etc. – K1010 and K1011). The third lexical group can be interpreted as reflecting the flight of the hero under the skin or the belly of a ram ('skin', 'ram', 'back', 'trick', 'disguise', etc. – K521, K521.1. and K603). The magic ring episode does not appear and might not be an essential motif, which is to be defined not in terms of the number of occurrences of the motif in the corpus, but rather by belonging to a core lexical group that appears constitutive of the tale. In order to prevent a circular representation of 'motif' (i.e. circularity as the method of text analysis circularly defining the phenomenon that is its object of study according to the parameters through which it is identified), the data obtained should be carefully re-analysed with other algorithms. Furthermore, our result should be reproduced with a larger database.

With the first tagged text (cyclops = giant), represented in Figures 3 and 4, the scores of the factors are far worse than those obtained with the untagged data (Factor 1: 31.02%; factor 2: 27.39%, for a total of 58.41% versus a total of 100% for the untagged data); this Principal Component (Figure 3) explains fewer things from a statistical point of view and so appears less reliable. Five categories were found. Whereas the untagged data presented three groupings on a more or less evenly distributed grid, the tagged data presented two groupings as outliers on the grid, while three are interpenetrating to varying degrees. Group (1), which appears here in red, covers 22.72% of the segments in the text, and presents a

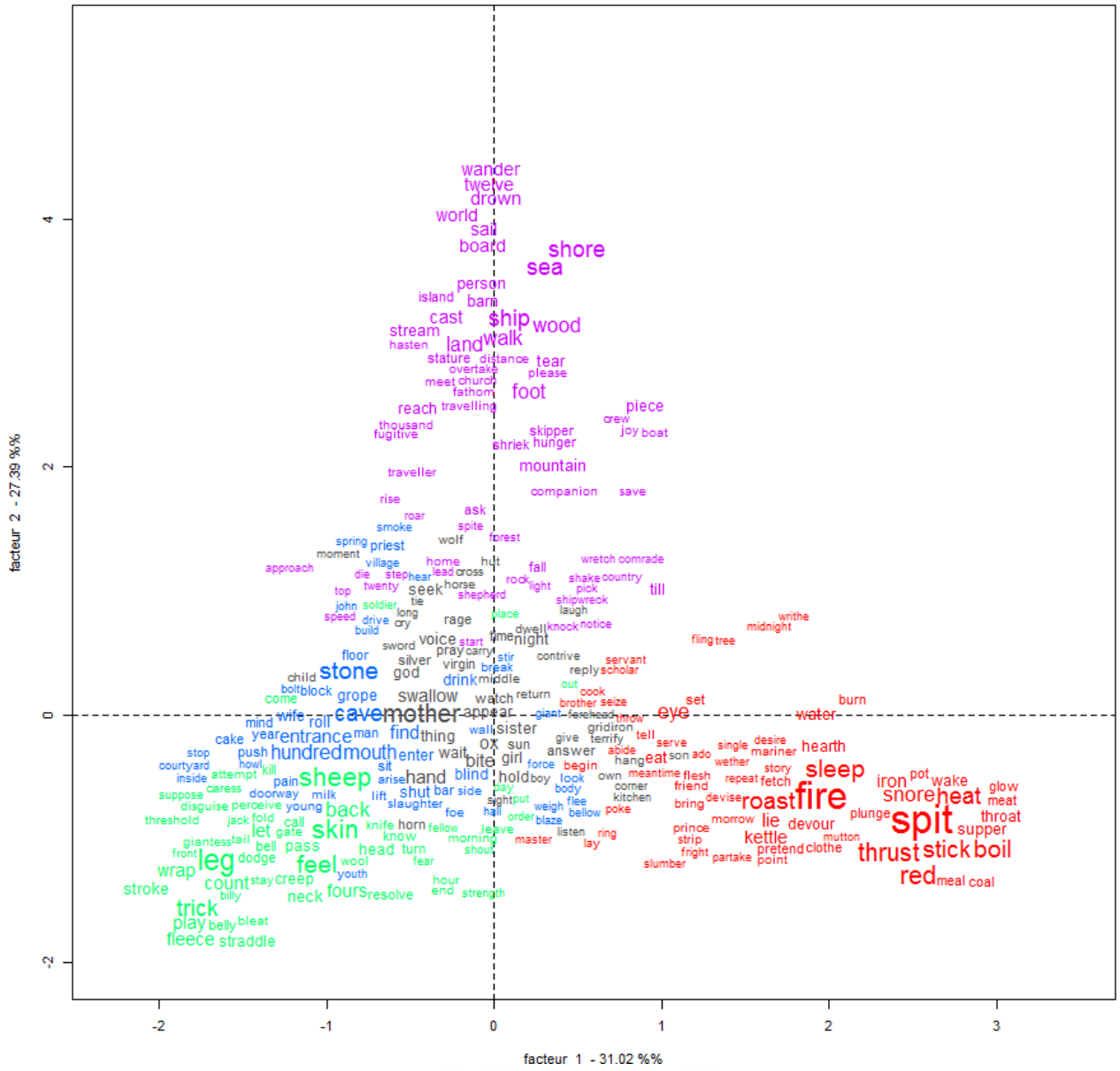

Figure 3. Principal Component Analysis of the Tagged texts 1 corpus (all terms for monster = 'giant').

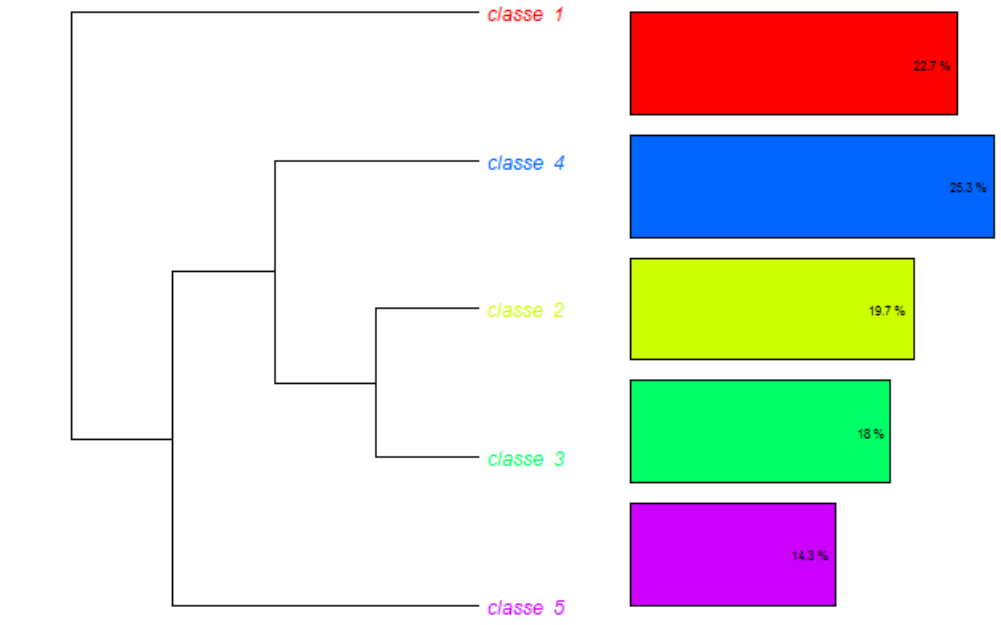

Figure 4. Dendrogram of the Principal Component Analysis in Figure 3. Percentages represent the percentage of segments of the texts.

fairly cohesive grouping that centers on the words ‘fire’, ‘spit’, ‘eye’, ‘thrust’, ‘boil’, ‘stick’, ‘red’, ‘roast’, ‘heat’, ‘snore’, ‘sleep’, with a number of minor words. This group was also one of the three groups in the analysis of the untagged corpus, although here with a slightly different concentration. The appearance of ‘single’ near ‘eye’ may here point to the unique eye of the monster found in several variants and the way used by the protagonist to blind him (‘thrust’ points toward the ‘glowing mass thrust into the eye’ of K1011). Group (2) only appears distinguished as a group for the Tagged texts 1 corpus. It covers 19.67% of the segments and is presented here in black at the center of the chart, interpenetrating all other groups while lacking any words more prominent than ‘mother’. This scattered group appears connected to human relations (‘mother’, ‘voice’, ‘sister’, ‘girl’, ‘son’ ‘man’, ‘virgin’, ‘boy’, ‘child’) and interactions (‘voice’, ‘laugh’, ‘answer’, ‘pray’, ‘reply’). Group (3) in green and covering 18,03% of the segments, presents the words ‘sheep’, ‘skin’, ‘leg’, ‘back’, ‘trick’, ‘count’, ‘belly’, ‘kill’, and so forth. ‘Sheep’ and ‘skin’ appear as quite pronounced elements, with ‘sheep’ the more prominent element here, in contrast to ‘ram’ in Figure 2, where the corresponding group is set apart. This cluster appears to correlate quite strikingly with the hero’s escape under the skin or the belly of a ram and can be interpreted as referring to the relevant narrative sequence (motifs K521, K521.1. and K603). Group 4 centers on the words ‘stone’, ‘cave’, ‘mouth’, ‘entrance’, ‘roll’, ‘enter’, ‘block’, ‘house’ and so forth, which are represented in a somewhat larger size. The group lacks particularly centralized elements although it was a clearly distinct group in the analysis of the untagged data, where it also exhibited more prominent words. This group points towards the home of the giant. In group (5), covering 14.29% of the segments here shown in pink, the larger words ‘ship’, ‘sea’, ‘shore’, ‘walk’, ‘wood’, ‘walk’, ‘land’, ‘board’ belong to the lexical field for travel.

Although the extensive interpenetration of Groups (3) and (4) may be because of a more regular co-occurrence of their constituents overall, this does not explain why these would have appeared as clearly distinguished groupings in Figure 2. The tagging of the agent adversary seems to have led to a much more distinctive clustering of elements that appear associated with the blinding. This has produced a shift in the grid and distribution in it. At the same time, tagging elements associated with the giant’s livestock has also affected the outcome: in Figure 2, ‘ram’ is associated with ‘skin’, ‘sheep’ is at the periphery of the grouping close to the cluster linked to the cave, and ‘goat’ is grouped with the cave cluster (with implications for the identification of the lexical item ‘goat’ with the giant’s livestock but not with the escape of the hero). Thus the tagging of the data has increased the potential representation of two motifs, one of which was not reflected in the untagged data, while situating the distinct groupings associable with the escape and location together.

The second tagged corpus is identical to the first except that ‘cyclops’ (= ‘one-eyed’) is separated from other terms for monster. The result (factor 1: 51,37%; factor 2: 48,63%) is similar to the analysis of the untagged corpus in the sense that it produces three groups on an evenly distributed grid. The first (31,07%) includes the words ‘sheep’, ‘skin’, ‘leg’, ‘pass’, ‘hand’, ‘back’, ‘horn’, ‘belly’ and points towards the hero’s escape. In contrast to both the untagged data and the Tagged texts 1 corpus, the elements associated with the escape do not exhibit a coherent grouping: for example, ‘cave’ (and ‘village’) group with ‘sheep’ while ‘home’ appears with the second cluster. In the second cluster, ‘wood’, ‘land’, ‘ship’, ‘walk’, ‘sea’, ‘shore’, ‘foot’, ‘drown’, ‘board’, and ‘home’ seem linked with the journey of the hero, which was not distinguished in the untagged data which also appeared in a slightly different configuration in the Tagged texts 1 corpus. Class 3 shows the words ‘fire’, ‘spit’, ‘eye’, ‘boil’, ‘stick’, ‘heat’, ‘thrust’, ‘red’, and so forth, pointing towards the blinding of the monster.

Across the three tests, groups associable with the blinding of the adversary and with the escape of the hero can be observed in all three cases. Tagging the terms for the livestock of the adversary appears to have

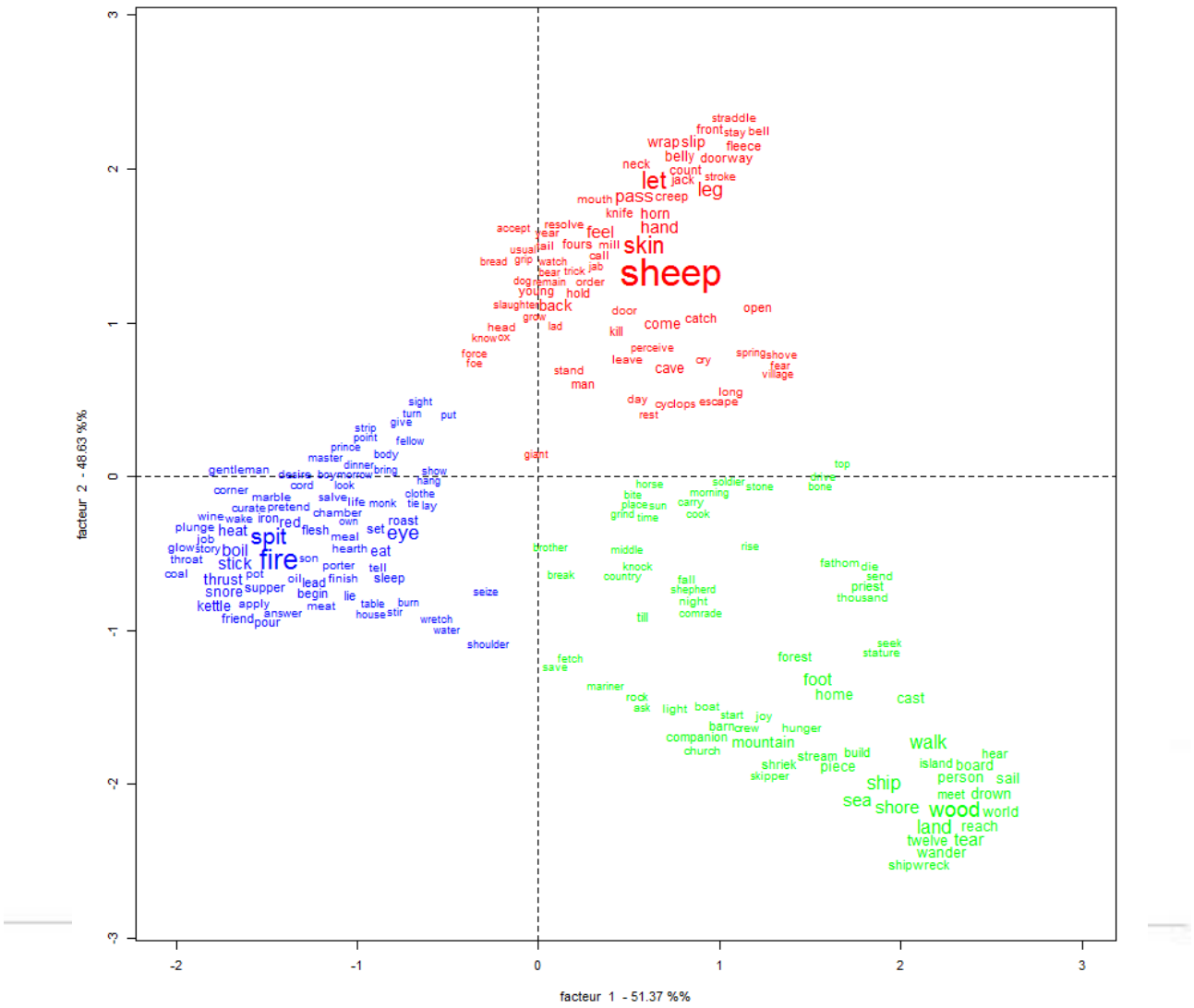

Figure 5. Principal Component Analysis of Tagged texts 2 corpus ('cyclops' = 'one-eyed'; all other monsters = 'giant').

Table 2. Chart of results comparing motifs identified with ATU 1137 by Thompson and Uther against those which appear identified by the Iramuteq Correspondance Factorial Analysis. (G100: Giant ogre, Polyphemus; F531: Giant; F512: Person unusual as to his eyes; K1011: Eye-remedy. Under pretence of curing eyesight, the trickster blinds the dupe. (Often with a glowing mass thrust into the eye.); K1010: Deception through false doctoring; K602: 'Noman'; K521.1: Escape by dressing in animal (bird, human) skin; K603: Escape under ram's belly; K521: Escape by disguise; D1612.1: Magic objects betray fugitive. Give alarm when fugitive escapes; PROPOSED MOTIF 1: Hero's habitat and relationship; PROPOSED MOTIF 2: Monster habitat; PROPOSED MOTIF 3: The journey; PROPOSED MOTIF 4: The monster owns sheep.)

Motif	**Untagged text**	**Tagged text 1**	**Tagged text 2**
G100/ F531	*Not found*	*Not found*	*Not found*
F512	Group 2 (forehead?)	Group 2 (forehead?)	*Not found*
K1011/K1010	Group 2	Group 1	Group 3
K602	*Not found*	*Not found*	*Not found*
K521.1/K603/K521	Group 3	Group 4	Group 1
D1612.1	*Not found*	*Not found*	*Not found*
PROPOSED MOTIF 1	Group 1	Group 2	*Not found*
PROPOSED MOTIF 2	Group 1	Group 3	Group 2 ?
PROPOSED MOTIF 3	Group 1 ?	*Not found*	Group 2
PROPOSED MOTIF 4	Group 3 ?	Group 4 ?	Group 1 ?

been significant in the intermingling of groups associated with the hero's escape and the place of habitation. The dissolution of the group associated with the place, when 'cyclops' (grouping closer to 'sheep') is distinguished from 'giant' while 'ogre' and other terms are not, is rather surprising.

To sum up, Table 2 offers an overview of the motifs for which potential evidence could be identified in the data using Iramuteq's Principle Component Analysis.

These analyses of tagged texts confirms the three categories found in the untagged data, but only two of these consistently, and they added three additional ones. The results seem robust and the variation in the lexical surface texture of texts may affect much less than what was initially expected in the outcome of analysis, although significant variations in some areas were clearly evident. However, it is noteworthy that, for example, G100: Giant ogre, Polyphemus / F531: Giant does not appear as a prominent element, but is rather represented in small font near the center of the three distributed groupings in Figures 2 and 5, suggesting a more or less equal association with each of these groups.

Iramuteq Similarities Analysis

A similarities analysis has also been done (index: co-occurrence; layout: fuchterman reingol; maximum tree; size of text: 10). This approach is based on properties of the connectivity of the corpus. The result is the graphic tree shown in Figure 6 (untagged text) and in Figure 7 (tagged text), where nodes are

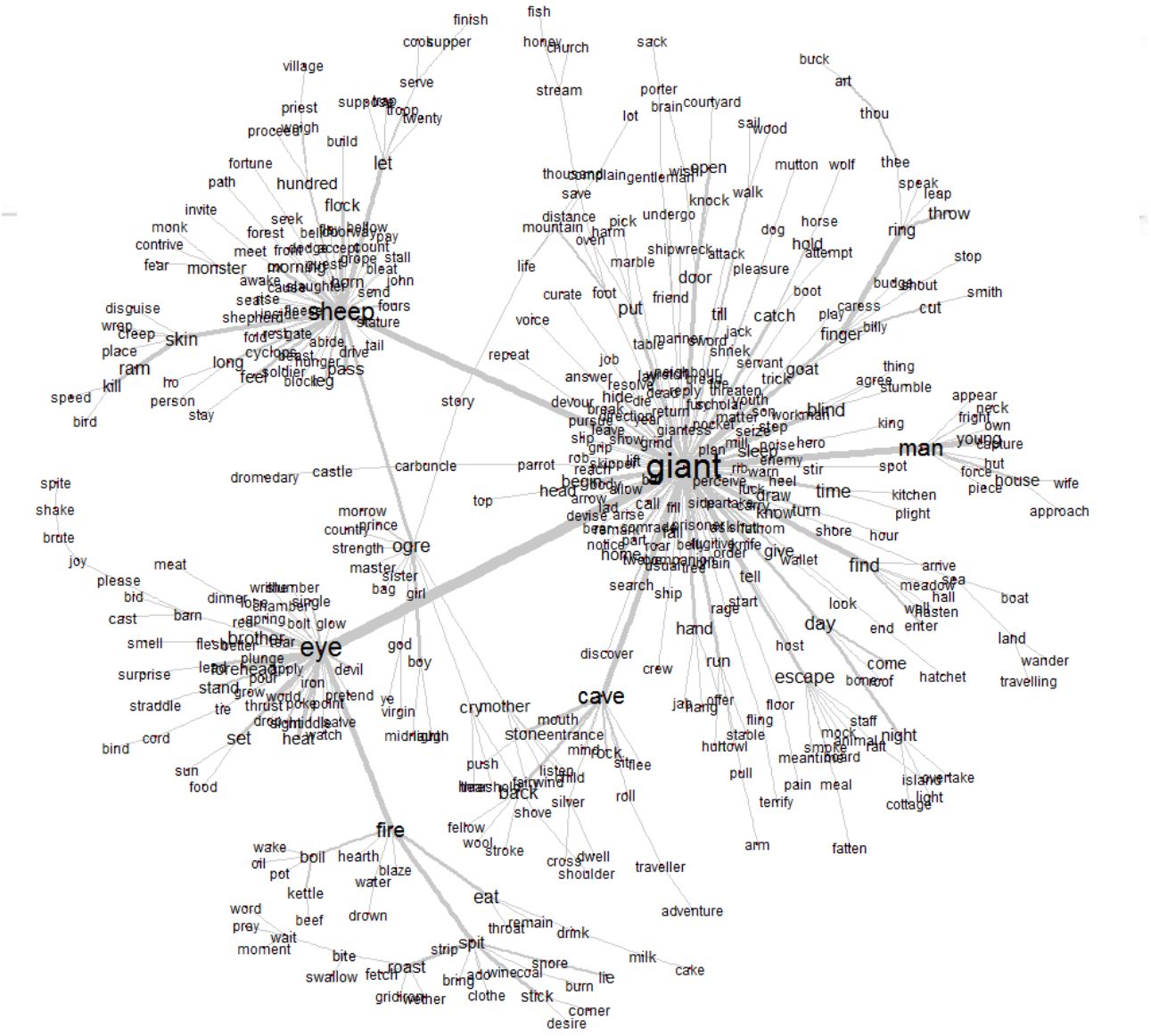

Figure 6. Similarities Analysis of the untagged corpus.

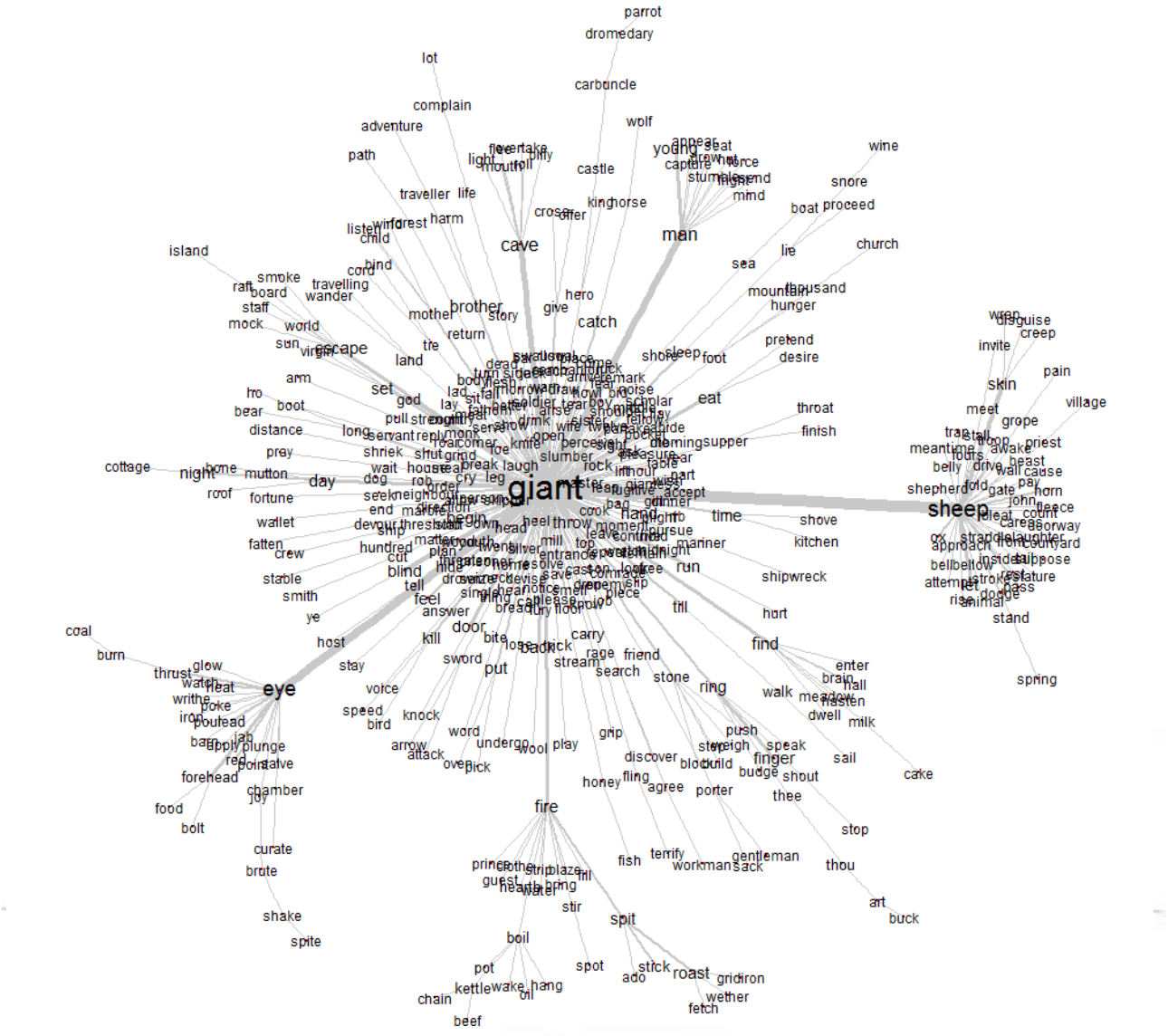

Figure 7. Similarities Analysis of the Tagged text 1 corpus (all terms for monster = 'giant').

lexically based elements revealed in the form of lexical communities.

This algorithm shows the proximity between the elements (co-occurrence). With the untagged data, the 'giant' (F.531) is the central figure, according to the fact that he should be associated with all essential motifs of the tale-type. Tree 6 allows the word 'giant' to be connected to many lexical groups, and it is linked to many important groups, organized around the words 'eye' (singular; F512), 'sheep', 'man' and 'cave'. These words, bigger than the others, could be the most important categories of being, around which less important beings, actions or objects could be organized. Only two groups seem to be very important: the cluster surrounding 'eye' is linked to two small groups: 'fire' and 'spit' (K1010 and K1011); the cluster surrounding 'sheep' is linked with the smaller groups 'ogre' (G100) and 'skin' (K521, K521.1. and K603). These results generally correspond to the group 2 and the group 3 found with the principal component analysis shown in Figure 2. It is noteworthy that 'ogre' appears in a position centered among the smaller clusters but connected to 'sheep' rather than to 'giant'. On the one hand, 'ogre' and 'giant' appear to function as mutually exclusive terms in the corpus, thus 'ogre' would not be linked as co-occurring with 'giant'. On the other hand, 'ogre' is associated with the same motifs but only links to one of these. The Iramuteq Similarities Analysis only allows each element to appear once in the tree and only allows semantic relations to branch outward, thus 'ogre' cannot be linked to elements in other branches from 'giant' although its position in the tree seems otherwise to reflect its relationship to them. A small group around the words 'finger' and 'ring' can be associated with the

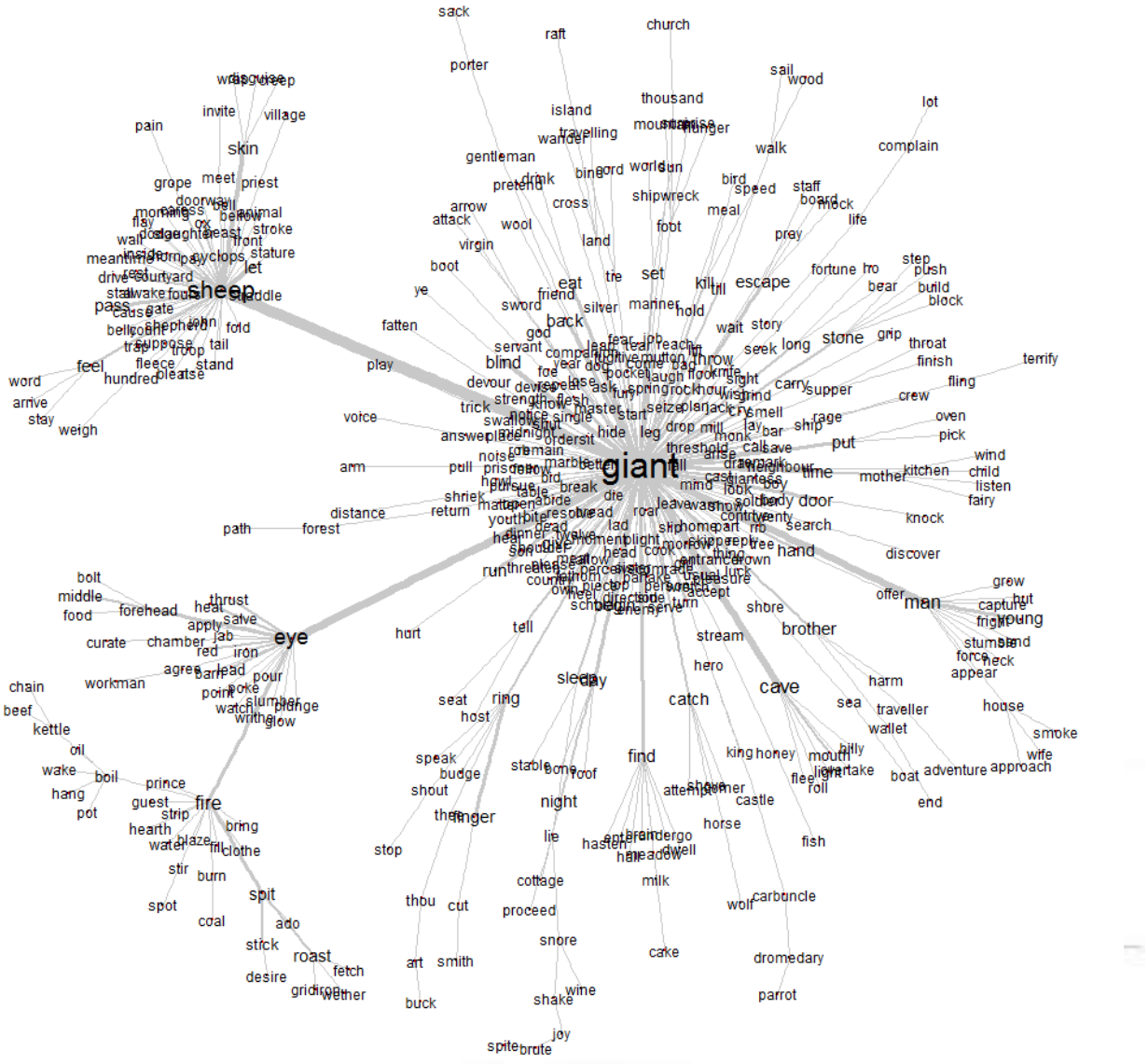

Figure 8. Similarities Analysis of the Tagged text 2 corpus ('cyclops' = 'one-eyed'; all other monsters = 'giant').

motif B1612.1. Another small group points toward the ogre's habitat (around 'cave').

The degree to which these can be seen as unambiguously linked to motifs in narration is problematized by words that appear fundamental to the motif but are dispersed elsewhere in the diagram. A striking example is precisely the linkage of 'eye', 'fire' and 'spit' that we are likely to associate with the motif of blinding the adversary (especially K1011) on the basis of our previous knowledge of the tale. However, the word 'blind' is distantly removed from these elements on another side of the cluster around 'giant'. Put another way, the key semantic element of the motif is absent from the prominent lexical cluster with which it seems most readily identifiable.

The program was relaunched with the same tagged texts as in Figures 3 and 5; significantly. The results are shown in Figures 7 and 8, which show less detailed yet similar clusters.

An overview of comparisons is surveyed in Table 3.

The untagged corpus produced a tree with wider dispersal and many more smaller branches in Figure 6 than the tagged corpora in Figures 7 and 8, as was expected. Tagging the term for the adversary and his livestock significantly tightened the groupings around each lexical-semantic center. This did not, however, significantly impact the centers for 'giant', 'sheep', 'eye' or 'fire', although the smaller centers 'cave' and 'man' were reduced in relative prominence while the center of 'ogre' was eliminated entirely.

Our results show that statistical tools can be placed productively in dialogue with motifs already claimed to be present by Thompson and Uther (e.g. K1011 / K1010, K521.1. / K521 / K603). More significantly,

Table 3. Chart of results comparing motifs identified with ATU 1137 by Thompson and Uther against those which appear identified by Iramuteq Similarities Analysis. (G100: Giant ogre, Polyphemus; F531: Giant; F512: Person unusual as to his eyes; K1011: Eye-remedy. Under pretence of curing eyesight the trickster blinds the dupe. (Often with a glowing mass thrust into the eye.); K1010: Deception through false doctoring; K602: 'Noman'; K521.1: Escape by dressing in animal (bird, human) skin; K603: Escape under ram's belly; K521: Escape by disguise; D1612.1: Magic objects betray fugitive. Give alarm when fugitive escapes; PROPOSED MOTIF 1: Hero's habitat and relationship; PROPOSED MOTIF 2: The Capture of the hero; PROPOSED MOTIF 3: Monster habitat PROPOSED MOTIF 4: The journey; PROPOSED MOTIF 5: The monster owns sheep.)

Motif	**Untagged text**	**Tagged text 1**	**Tagged text 2**
G100/F531	Around 'giant'	Around 'giant'	Around 'giant'
F512	Around 'eye' ('forehead', 'middle')	Around 'eye' ('forehead')	Around 'eye' ('forehead', 'middle')
K1011/K1010	Around 'fire' & 'spit'	Around 'fire', 'spit' & 'boil'	Around 'fire'
K602	*Not found*	*Not found*	*Not found*
K521.1/K603/K521	Around 'sheep' & 'skin'	Around 'sheep' & 'skin'	Around 'sheep' & 'skin'
D1612.1	Around 'ring' (small cluster)	Around 'ring' (small cluster)	Around 'ring' (small cluster)
PROPOSED MOTIF 1	Around 'man' (small cluster)	Not found	*Not found*
PROPOSED MOTIF 2	*Not found*	Around 'man' (small cluster)	Around 'man' (small cluster)
PROPOSED MOTIF 3	Around 'cave'	Around 'cave' (?)	Around 'cave' (?)
PROPOSED MOTIF 4	*Not found*	*Not found*	*Not found*
PROPOSED MOTIF 5	Around 'sheep'	Around 'sheep'	Around 'sheep'

these tools also makes it possible to consider new motifs, such as 'the home of the giant is a cave' (around the word 'cave'), 'a giant owns sheep' (around 'sheep'), 'a young man is captured' (around 'man').

Conclusion

When comparing the Iramuteq analysis to the the classical motifs identified with this tale by Thompson and Uther, it was possible to propose good correspondences with many of them, as shown in Table 4.

As one can see, the software remains far from fully satisfactory. It only found K1010/K1011 (the blinding event) and K521.1./ K.603 / K.521 (escape under the skin). In more than 50% of cases, one can accept the detection of F512 (Person unusual

Table 4. Chart comparing results from Tables 2 and 3.

Motif	**Correspondence factorial analysis (% of the results)**	**Similarities analysis (% of the results)**
G100/ F531	*Not found*	100 %
F512	66%	100 %
K1011/K1010	100 %	100 %
K602	*Not found*	*Not found*
K521.1/K603/K521	100 %	100 %
D1612.1	*Not found*	100 %
PROPOSED MOTIF 1	66 %	33 %
PROPOSED MOTIF 2	*Not found*	66 %
PROPOSED MOTIF 3	100 %	100 % (?)
PROPOSED MOTIF 4	33 %	*Not found*
PROPOSED MOTIF 5	100 %	100 %

as to his eyes), even if it remains questionable (to find the lexical cluster linking to each motif, see above). D1612.1 (the ring episode) and G100 were found in only 50% of the results. To explain this difference, we must remember first that tests on the corpus were problematic. Certain classic motifs such as the 'Noman' false name may reduce to a single lexical item according to this approach. Similarly, Iramuteq automatically reduces words to their root forms, and thus cannot distinguish between 'eye' and 'eyes', which might be relevant for the cyclops having one eye as opposed to two (F512). These elements have highlighted problems of identifying all the elements purely on the basis of the lexical surface of the text because the single term may vary from text to text and also vary with other lexica such as personal names and pronouns, as well as being rendered through description as opposed to a keyword.

Our method may have detected two new motifs: 'the monster's habitat is a cave' and 'the monster has sheep'. However, when considering comparison with Thompson's motif index, it is necessary to observe that Thompson was concerned with motifs that could be found across tale-types, and consequently motifs as quite abstract or general elements. In contrast, the present study analyzes only a single tale-type in order to identify recurrent elements characteristic of that type at the lexical surface of texts in translation. Of course, the findings using this corpus necessarily remain conditional on the degree to which this corpus is representative of the tradition addressed, and the quality of information produced is dependent on the quality of the sources. However, if we imagine for the sake of experiment that these texts ideally render English lexical equivalents of the sources of the tradition, it is not clear whether these studies reveal 'motifs' as conventional units of this particular tale-type or 'motifs' at the more abstract level of Thompson's types. Additional research in this field is certainly needed. Similar studies across narratives of different types are required to confirm the new motifs preliminarily identified here and to find others. Moreover, the use of multiple types of software and algorithms is highly recommended to compare the results. Additionally, if, following the present study, we define a motif as a semantic attractor, a central point which underpins a set of related words, it is also necessary to observe that a constellation of lexica such as 'skin', 'ram', 'back', 'trick' and 'disguise' cannot necessarily be reconstituted as a single, coherent motif. Similarly, the constellation 'fire', 'spit' and 'eye' might be interpretable as a blinding motif, but this constellation begins to appear chaotic when it is accompanied by 'sleep', 'roast' and 'boil'. At the present state of research, it is interesting to apply these tools in research on motifs, but it is not possible to reconstitute motifs from the information produced without presupposing narrative elements (as already described motifs) and placing these as well as the information produced by analysis in dialogue with the source data.

This pilot study initially set out to use software to demonstrate 'motifs', which proved highly problematic in a number of respects. However, the outcome did produce a new model for approaching tales in terms of semantic networks of elements. The graphic representations in Figures 2–3 and 5–8 are not representations of motifs *per se*, but of whole tales. Given a particular tale, forthcoming software programs may determine if this story belongs to a particular tale-type (previously determined as a certain cloud of words) and if it could be brought closer to other tales belonging to the same group on the sole basis of the shared semantic elements.

Acknowledgements: The author wishes to thank Sándor Darányi, Jean-Loïc Le Quellec and Jamshid J. Tehrani. My special thanks and acknowledgement go to Mr. Frog, who has been very helpful and his comments have improved a lot of passages in this paper. This text owes much to him.

Works Cited

Gambette, Philippe, & Jean Veronis. 2009. "Visualising a Text with a Tree Cloud". *IFCS'09*, software freely available from www.treecloud.org.

Frazer, James George. 1921. *Apollodorus: The Library* II. Londres: William Heinemann / New York: G.P. Putnam's sons.

Hackman, Oskar. 1904. *Die Polyphemsage in der Volksüberlieferung.* Helsinki.

Lévi-Strauss, Claude. 1958. *Anthropologie structurale*. Paris: Plon.

Ratinaud, Pierre. 2009. "IRaMuTeQ : Interface de R pour les Analyses multidimensionnelles de Textes et de Questionnaires". Available at: http://www.iramuteq.org

Ratinaud, Pierre. 2012. "Analyse automatique de textes". Unpublished manuscript accessed: http://repere.no-ip.org/Members/.

Schonhardt-Bailey, Cheryl. 2013. "Book Appendix I: Methodology". *Deliberating American Ponetary Policy: A Textual Analysis*. London: MIT Press. Available at: http://mitpress.mit.edu/sites/default/files/BOOK%20APPENDIX%20I.pdf

Thompson, Stith. 1932–1936. *Motif-Index of Folk-Literature: A Classification of Narrative Elements in Folk-Tales, Ballads, Myths, Fables, Mediæval Romances, Exempla, Fabliaux, Jest-Books and Local Legends* I–VI. FF Communications 106–109, 116–117. Helsinki: Academia Scientiarum Fennica.

Thompson, Stith. 1955. *Narrative Motif-Analysis as a Folklore Method*. FF Communications 161. Helsinki: Academia Scientiarum Fennica.

Thompson, Stith. 1955–1958. *Motif-Index of Folk-Literature: A Classification of Narrative Elements in Folk-Tales, Ballads, Myths, Fables, Mediæval Romances, Exempla, Fabliaux, Jest-Books and Local Legends* I–VI. 2nd rev. edn. Bloomington: Indiana University Press.

Thompson, Stith. 1961. *The Types of the Folktale: A Classification and Bibliography: Antti Aarne's Verzeichnis der Märchentypen (FFC No. 3) Translated and Enlarged*. 2nd rev. edn. Helsinki: Academia Scientiarum Fennica.

Uther, Hans-Jörg. 2004. *The Types of International Folktales: a Classification and Bibliography, Based on the System of Antti Aarne and Stith Thompson. Tales of the Stupid Ogre, Anecdotes and Jokes, and Formula Tables.* Helsinki: Suomalainen Tiedeakatemia.

The U Version of *Snorra Edda*

Daniel Sävborg, University of Tartu

Snorra Edda has been preserved in four independent manuscripts. Codex Regius, Codex Wormianus and Codex Trajectinus are close to each other and can – in spite of certain differences – be said to represent one version, RTW. The text of Codex Upsaliensis is at several points very different from the other manuscripts and is usually seen as the sole representative of another version, U. What distinguishes the two versions is mainly the length and style of the narrative sections. The U version is, as a whole, remarkably shorter than RTW. Its style and narrative technique is terse and panoramic, mentioning only details necessary for the plot or the purpose of the story, while the style and narrative technique of RTW is broad, scenic, and full of rhetorically effective but factually irrelevant details (a fuller analysis is given in Sävborg 2012: 13–16).

Scholars have long argued about which version is closest to the original. Scholars such as Finnur Jónsson (1898: 306–355) and D.O. Zetterholm (1949: 46–54) argued for the priority of RTW, while e.g. Eugen Mogk (1879: 510–537) and Friedrich Müller (1941: 146) argued that U best represents Snorri's original version. Recently, Heimir Pálsson has revived the arguments in favor of U's priority in the introduction to his edition of U (Heimir Pálsson 2012: cxvii). The main scholarly overviews describe the matter as unsolved (e.g. Lindow 1988: 352; Faulkes 1992: 601).

So far, scholars have used criteria such as the degree of quality, accuracy and logic to determine the priority. Just a few examples will be mentioned. Eugen Mogk points to details where U, according to him, has the better text ("Dass dieser lesart die von A [= U] […] vorzuziehen ist, unterliegt wol keinem zweifel" ['That this reading in A (= U) [...] is preferable, there can indeed be no doubt'], etc.; Mogk 1879: 528), while RTW, in contrast, has elements – absent in U – that are "störend" ['disturbing'] (1879: 508). He also mentions alleged contradictions, inconsequences and illogical features in RTW, which in the corresponding parts of U are consequent and logical (1879: 511–514). For him, these are strong arguments for the priority of U. Finnur Jónsson, on the other hand, comes to a conclusion opposite to Mogk by arguing in exactly the same way. He points to cases where "det eneste logiske" ['the only thing that is logical'] is found in RTW but not in U (Finnur Jónsson 1898: 335). Friedrich Müller turned the discussion upside-down in 1941. He argued in favor of Mogk's conclusion that U represents the original version, and that RTW is a reworking of it, but his arguments were exactly opposite to Mogk's. For Müller, U can be established