Inferring Systemic Nets with Applications to Islamist Forums

David B. Skillicorn and N. Alsadhan

Abstract Open-source intelligence often requires extracting content from documents, for example intent and timing. However, more interesting and subtle properties can be extracted by directing attention to the thought patterns and framing that is implicitly present in the writings of groups and individuals. Bag-of-words representation of documents are useful for information retrieval, but they are weak from the perspective of intelligence analysis. We suggest that systemic functional linguistics, with its focus on the purpose an author intends for a document, and its abstraction in terms of choices, is a better foundation for intelligence analysis. It has been limited in practice because of the difficulty of constructing the systemic nets that are its representation of these choices. We show that systemic nets can be constructed inductively from corpora using non-negative matrix factorisation, and then apply this to infer systemic nets for language use in islamist magazines published by three different groups: Al Qaeda, Daish (ISIS), and the Taliban. We show that the structures captured are also present in posts in two large online forums: Turn to Islam and Islamic Awakening, suggesting a widely held mindset in the Islamic world.

1 Motivation

Applying intelligence collection and analysis strategies to open source data is an obvious strategy because of the availability of vast numbers of documents online: web pages, but also social network status updates, tweets, forum posts, blogs, and podcasts. Leveraging this data requires solving two problems: (1) finding documents relevant to a subject of interest, and (2) extracting the intelligence content implicit in those documents.

D. B. Skillicorn (✉) · N. Alsadhan
School of Computing, Queen's University, Kingston, ON, Canada
e-mail: skill@cs.queensu.ca

M. A. Tayebi et al. (eds.), *Open Source Intelligence and Cyber Crime*, Lecture Notes
in Social Networks, https://doi.org/10.1007/978-3-030-41251-7_5

The first problem, information retrieval, has been solved using the bag-of-words approach to representing the content of each document, followed by large-scale index search and careful ranking of the document set. Web search businesses depend on this as a crucial technology on which they build monetized services such as focused advertisements.

The bag-of-words representation of text has proven extremely successful, even for languages such as English where word order is crucial to meaning. However, it is less useful for extracting intelligence content: sentences such as “the criminal shot the officer” and “the officer shot the criminal” are equally plausible responses to queries about criminals and officers, but much less equivalent from the perspective of law enforcement and the media. Solving the second problem, intelligence extraction, depends on understanding what a document is ‘about’ in a semantic sense, as well as aspects of each document’s meta-properties: who wrote it, what their intent was in doing so, how it might be understood by an audience, whether it was intentionally deceptive, what attitudes and emotional tone it conveys, and a long list of other possibilities. In other words, many useful properties require understanding what might be called the social, or even sociological, properties of documents.

Determining such properties is key to domains such as e-discovery (finding significant emails in a corporate archive), intelligence (finding meaningful threats in a set of forum posts), measuring the effectiveness of a marketing campaign (in online social media posts), or predicting an uprising (using Twitter feed data).

Although bag-of-words approaches have been moderately successful for such problems, they tend to hit a performance wall (80% prediction accuracy is typical) because the representation fails to capture sufficient subtleties [30]. There have been attempts to increase the quality of representations, for example by extracting parse trees (that is, context-free grammar representations) but this focuses entirely on (somewhat artificial) language structure, and not at all on mental processes [14]. Other approaches leverage syntactically expressed semantic information, for example by counting word bigrams, by using Wordnet [26], or using deep learning [29]. Recent developments in deep learning, particularly LSTMs and biLSTMs have increased prediction accuracy; but these predictors are black boxes, so they bring little understanding of why and how a property is present in a document.

One approach that shows considerable promise is *systemic functional linguistics* [9, 12, 19], a model of language generation with sociological origins and an explicit focus on the effect of the creator’s mental state and social setting on a created document. In this model, the process of generating an utterance (a sentence, a paragraph, or an entire document) is conceived of as traversing a *systemic net*, a set of structured choices. The totality of these choices defines the created document. At some nodes, the choice is disjunctive: continue by choosing *this* option or by choosing *that* one. At others, the choice is conjunctive: choose a subset of these options and continue in parallel down several paths.

Figure 1 shows a simple example of a systemic net. The decision to communicate requires a parallel (independent) choice of the level of formality to be used and the communication channel to be used. The level of formality could be formal or

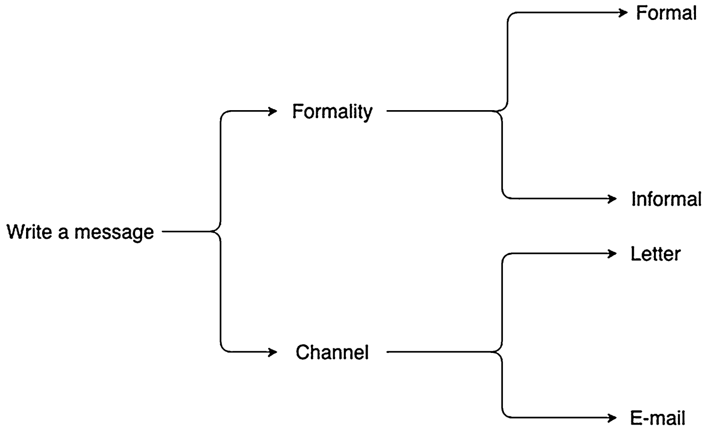


Fig. 1 A simple example of a systemic net

informal; and the channel could be via physical letter or email. These choices at the second level are disjunctive—it has to be one or the other. Further choices exist below these ones, so the systemic net notionally continues to the right until it results in concrete language.

Production using a context-free grammar also requires a structured set of choices, but the choices are top-down (so that the first choice is to instantiate, for example, a declarative sentence as a subject, an object, and a verb). In contrast, the order of the choices in a systemic net has no necessary relationship to the concreteness of the implications of those choices. For example, the choice to use formal or informal style is an early choice with broad consequences that limit the possibilities for subsequent choices. The choice to write a letter or an email is also an early choice but its immediate consequence is narrow and low level: typically whether the first word of the resulting document will be "Dear" (for a letter) or not (for an email).

Another example of a well-used systemic net, called the Appraisal Net [1], is shown in Fig. 2. It describes the way in which choices of adjectives are made when evaluating some object. The choice process is not arbitrary; rather an individual chooses simultaneously from up to three parallel paths: appreciation, affect, and judgement. Within two of these choices, there are then subsequent parallel choices that lead to particular adjectives—one example adjective is shown at each leaf. These choices are associated with different aspects of the situation: composition-complexity captures aspects of the object being appraised, while reaction-quality captures aspects of the person doing the appraising.

The power of systemic nets comes because these choices are made, not simply with the goal of constructing a syntactically valid sentence, but because of the limitations and exigencies of *social purpose* (certain things cannot be said in certain circumstances although syntactically valid); *mental state* (because language generation is a largely subconscious process), and the properties of the language in use. In other words, the choice of adjective in an appraisal certainly says something

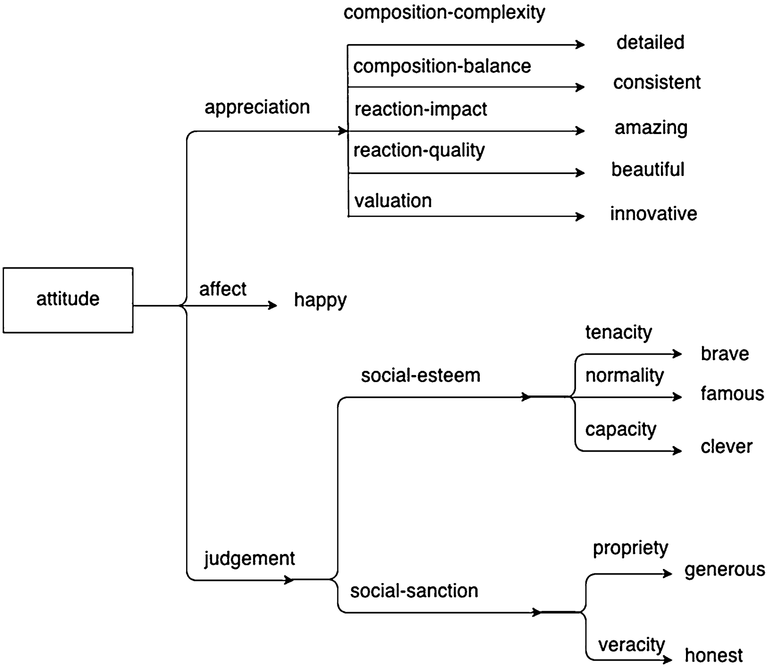


Fig. 2 The Appraisal systemic net, appropriate for representing judgements or reviews

about the object being assessed, but also reveals something about the person doing the assessing; and the structure of the choices would be different in English from, say, French or Japanese.

A systemic net is explanatory at three different levels. First, the existence of a net organizes constructions into categories and so explains some aspects of how the pieces in a text fit together.

Second, the choices made by individuals traversing a net are not typically unique; rather, they cluster into common choice patterns that reflect particular kinds of textual targets. This is because there are social rules that govern acceptable end-products. Each individual can write with an individual style, but we can also say that some set of documents by different authors are written in a down-to-earth style, and another set in a flowery style. This idea of a consistent set of choices in a net, leading to detectable consistencies in the resulting documents is called a *register*. Thus the set of registers associated with a net are also explanatory.

Third, for any particular document we can list the choices made in its construction, and this becomes a record that describes that document at a higher level of abstraction than as a bag of words. This level of explanation is most directly useful for analytics—such choices can be used as attributes for clustering or for prediction.

The advantages of a systemic functional approach to textual analytics are:

- The choices within the net are a smaller, more abstract, and more structured set than the choice of individual words, and therefore provide a stronger foundation for knowledge discovery—a kind of structured attribute selection; and
- These choices reflect, and make accessible, the mental state of the author or speaker and his/her perception of the social situation for which the text was constructed. This enables a kind of reverse engineering of how the text came to be, that is analytics about authors and settings.

The reason why systemic net approaches have not been more widely used in text analytics is because they have, so far, been constructed by computational linguists, often requiring several person-years to build, even when of modest size. Some substantial systemic nets have been built, but usually within the context of projects where they have been kept confidential; those that are public, like the Appraisal Net above, are usually small.

The contributions of this chapter are:

- We show that it is possible to infer systemic nets from corpora using Non-Negative Matrix Factorization (NNMF), and that these nets are plausible. Thus we are able to construct systemic nets for any corpus, and for any set of relevant words. This creates a new path to representing corpora at a deeper level, but without the need (and cost) for substantial human input.
- We show that the resulting systemic nets organize corpora more strongly than the corresponding bags of words, and that this organization improves both clustering and prediction tasks, using authorship prediction as a demonstration task.
- We apply systemic functional nets to a real-world intelligence problem, learning systemic nets from a set of Islamist magazines, and applying the resulting structure to two large Islamist forums. We show that the top-level distinctions derived from the magazines can also be clearly seen in the forum posts, suggesting a widespread mindset shared by the audience for these ideas.

2 Related Work

There have been several applications of predefined systemic nets to textual prediction problems. For example, Whitelaw et al. [32] show improvement in sentiment analysis using the Appraisal Net mentioned above. Argamon et al. show how to predict personality type from authored text, again using systemic functional ideas [31]. Herke-Couchman and Patrick derive interpersonal distance from systemic network attributes [13].

The most successful application of systemic functional techniques is the Scamseek project. The goal of this project was to predict, with high reliability, web pages that represented financial scams and those that represented legitimate financial products. This is a challenging problem—the differences between the two classes are small and subtle, and even humans perform poorly at the margins. The fraction of documents representing scams was less than 2% of the whole. This project's predictive model was successfully deployed on behalf of the Australian Securities and Investments Commission [21]. However, the effort to construct the registers corresponding to normal and (many varieties of) scam documents was substantial.

Kappagoda [15] shows that word-function tags can be added to words using conditional random fields, in the same kind of general way that parsers add part-of-speech tags to words. These word-function tags provide hints of the systemic-functional role that words carry. This is limited because there is no hierarchy. Nevertheless, he is able to show that the process of labelling can be partially automated and that the resulting tags aid in understanding documents.

Especially since the World Trade Center attacks of 2001, there has been a great deal of academic work on open-source intelligence [4, 17, 24]. This includes leveraging text [16, 25, 33] and graph data, including social networks [5, 6, 10, 22]. A large number of commercial platforms have also been developed and are in widespread use, for example i2 Analysts' Notebook and Palantir.

3 Inductive Discovery of Systemic Nets

The set of choices in a systemic net lead eventually, at the leaves, to choices of particular (sets of) words. One way to conceptualize a systemic net, therefore, is as a hierarchical clustering of words, with each choice representing selection of a subset.[1] We use this intuition as a way to inductively construct a systemic net: words that are used together in the same document (or smaller unit such as a sentence or paragraph) are there because of a particular sequence of choices. An inductive, hierarchical clustering can approximate a hierarchical set of choices.

Our overall strategy, then, is to build document-word matrices (where the document may be as small as a single sentence), and then cluster the columns (that is, the words) of such matrices using the similarity of the documents in which they appear. The question then is: which clustering algorithm(s) to use.

In this domain, similarity between a pair of documents depends much more strongly on the *presence* of words than on their *absence*. Conventional clustering algorithms, for example agglomerative hierarchical clustering and other algorithms that use distance as a surrogate for similarity, are therefore not appropriate, since mutual absence of a word in two different documents is uninformative, but still increases their apparent similarity.

Singular value decomposition is reasonably effective (J.L. Creasor, unpublished work) but there are major issues raised by the need to normalize the document-word matrix so that the cloud of points it represents is centered around the

[1] Complete systemic nets also include a downstream phase that defines the process for assembling the parts of a constructed document into its actual linear sequence. We ignore this aspect. In declarative writing, assembly is usually straightforward, although this is not the case in, for example, poetry.

origin. Typical normalizations, such as z-scoring, conflate median frequencies with zero frequencies and so introduce artifacts that are difficult to compensate for in subsequent analysis.

We therefore choose to use Non-Negative Matrix Factorization, since a document-word matrix naturally has non-negative entries. An NNMF decomposes a document-word matrix, A, as the product of two other matrices:

$$A = WH$$

If A is $n \times m$, then W is $n \times r$ for some chosen r usually much smaller than either m or n, and H is $r \times m$. All of the entries of W and H are non-negative, and there is a natural interpretation of the rows of H as 'parts' that are 'mixed' together by each row of W to give the observed rows of A [18].

Algorithms for computing an NNMF are iterative in nature, and the results may vary from execution to execution because of the random initialization of the values of W and H. In general, the results reported here are obtained by computing the NNMF 10 times and taking the majority configuration. We use a conjugate gradient version of NNMF, using Matlab code written by Pauca and Plemmons.

There are two alternative ways to use an NNMF, either directly from the given data matrix, or starting from its transpose. If we compute the NNMF of the transpose of A, we obtain:

$$A' = \bar{W}\bar{H}$$

and, in general, it is not the case that $\bar{H} = W'$ and $\bar{W} = H'$. Experiments showed that results were consistently better if we applied the NNMF to A', that is to the word-document matrix. The textual unit we use is the paragraph. A single sentence might, in some contexts, be too small; a whole document is too large since it reflects thousands of choices.

We extracted paragraph-word matrices in two ways. A parts-of-speech-aware tagger made it possible to extract the frequencies of, for example, all pronouns or all determiners [7]. For larger word classes, such as adjectives, it was also possible to provide the tagger with a given list and have it extract only frequencies of the provided words. Frequency entries in each matrix were normalized by the total number of words occurring in each paragraph, turning word counts into word rates. This compensates for the different lengths of different paragraphs.

Superior results were obtained by choosing only $r = 2$ components. In the first step, the $\bar{W}$ matrix has dimensionality *number of words* $\times$ 2, with non-negative entries. Each word was allocated to the cluster with the largest entry in the corresponding row of $\bar{W}$, and the process repeated with the two submatrices obtained by splitting the rows of A' based on this cluster allocation. This process continued until the resulting clusters could not be cleanly separated further. These clusters therefore form a binary tree where each internal node contains the union of the words of its two children.

Each NNMF was repeated 10 times to account for the heuristic property of the algorithm. We were able to leverage this to estimate the confidence of each clustering. For example, there were occasionally particular words whose membership oscillated between two otherwise stable clusters, and this provided a signal that they didn't fit well with either. We were also able to use this to detect when to stop the recursive clustering: either clusters shrank until they contained only a single word (usually a high-frequency one), or their subclusters began to show no consistency between runs, which we interpreted to mean that the cluster was being over-decomposed.

The result of applying this recursive NNMF algorithm to a word-paragraph matrix is a hierarchical binary tree whose internal nodes are interpreted as choice points, and whose leaves represent the 'outputs' that result from making the choices that result in reaching that leaf. A leaf consists of a set of words that are considered to be, in a sense, equivalent or interchangeable from the point of view of the total set of words being considered. However, this view of leaves contains a subtle point. Suppose that a leaf contains the words 'red' and 'green'. These are clearly not equivalent in an obvious sense, and in any given paragraph it is likely that an author will select only one of them. In what sense, then, are they equivalent? The answer is that, from the author's point of view, the choice between them is a trivial one: either could serve in the context of the document (fragment) being created. Thus a leaf in the systemic net contains a set of words from which sometimes a single word is chosen and sometimes a number of words are chosen—but in both cases the choice is unconstrained by the setting (or at least undetectably unconstrained in the available example data).

We have remarked that choices at internal nodes in a systemic net can be disjunctive or conjunctive. However, in our construction method each word in a particular document is allocated to exactly one cluster or the other. We estimate the extent to which a choice point is conjunctive or disjunctive by counting how often the choice goes either way across the entire set of documents, that is we treat conjunction/disjunction as a global, rather than a local, property. (It would be possible to allocate a word to both clusters if the entries in the corresponding row of $\bar{W}$ had similar magnitude, and therefore detect conjunctive choices directly. However, deciding what constitutes a similar magnitude is problematic because of the variation between runs deriving from the heuristic nature of the algorithm.)

4 Inferred Systemic Nets

The data used for proof of concept of this approach is a set of 17 novels downloaded from gutenberg.org and lightly edited to remove site-specific content. These novels covered a period of about a century from the 1830s to the 1920s and represent well-written, substantial documents. For processing they were divided into paragraphs; because of the prevalence of dialogue in novels, many of these paragraphs are actually single sentences of reported speech. The total number of paragraphs is

Table 1 List of words used to create the systemic networks

Group type	Words
Personal pronouns	I, me, my, mine, myself, we, us, our, ours, ourselves, you, your, yours, yourself, yourselves, they, their, theirs, them, themselves, he, him, his, himself, she, her, hers, herself, it, its, itself, one, one's
Adverbs	Afterwards, already, always, immediately, last, now, soon, then, yesterday, above, below, here, outside, there, under, again, almost, ever, frequently, generally, hardly, nearly, never, occasionally, often, rarely
Auxiliary verbs	Was, wasn't, had, were, hadn't, did, didn't, been, weren't, are, is, does, am, has, don't, haven't, doesn't, aren't, do, isn't, have, be, hasn't
Positive auxiliary verbs	Was, had, were, did, been, is, does, are, am, has, do, have, be
Adjectives	Good, old, little, own, great, young, long, such, dear, poor, new, whole, sure, black, small, full, certain, white, right, possible, large, fresh, sorry, easy, quite, blue, sweet, late, pale, pretty
Verbs	Said, know, see, think, say, go, came, make, come, went, seemed, made, take, looked, thought, saw, tell, took, let, going, get, felt, seen, give, knew, look, done, turned, like, asked

48,511. The longest novel contained 13,617 paragraphs (*Les Miserables*) and the shortest 736 (*The 39 Steps*).

We selected six different categories of words for experiments as shown in Table 1.

Figure 3 shows the systemic net of pronouns. In all of these figures, the thickness of each line indicates how often the corresponding path was taken as the result of a choice. Lines in blue represent the 'upper' choice, red the 'lower' choice, and black the situation where both choices occurred with approximately equal frequency.

The top-level choice (1) in this net is between pronouns where the point of view is internal to the story, and where the point of view is of an external narrator. This seems plausible, especially in the context of novels. Choice point 2 is largely between first-person and second-person pronouns, with apparently anomalous placement of 'me' and 'we'. Choice point 4 is between masculine pronouns and others, again entirely plausible given the preponderance of masculine protagonists in novels of this period. The remaining choices in this branch separate feminine, impersonal, and third-person plural pronouns. All of these choices are strongly

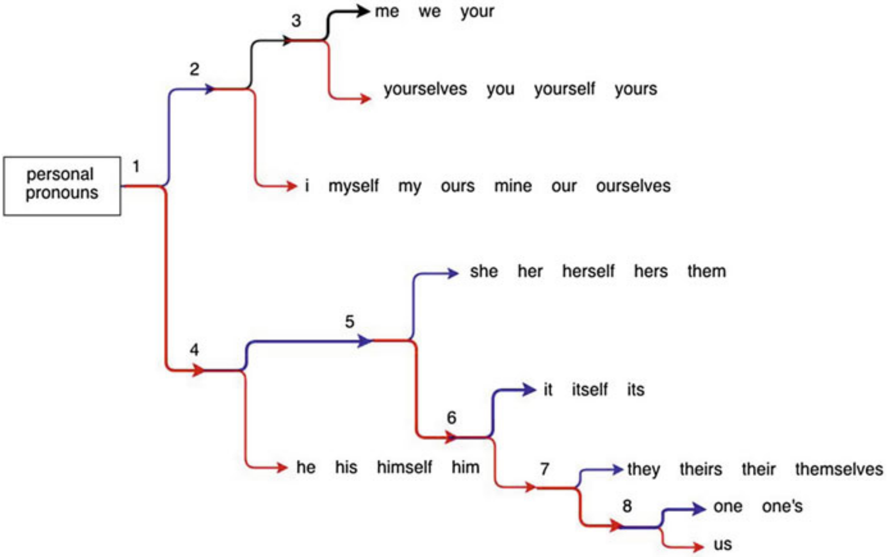

personal pronouns
1
2
3
me we your
yourselves you yourself yours
i myself my ours mine our ourselves
4
5
she her herself hers them
6
it itself its
he his himself him
7
they theirs their themselves
8
one one's
us

Fig. 3 Systemic net inferred for pronouns

disjunctive, weakening down the tree with choice point 7 the least disjunctive. It might be expected that, after the choice at point 1, choices might become more conjunctive as two or more people are mentioned. However, reported speech by one person is the most common paragraph structure in these novels, and many of these do not contain another pronoun reference (“He said ‘What’s for dinner?”).

Figure 4 shows the systemic net for auxiliary verbs. These might have separated based on their root verb (to be, to have, to do)[2] but in fact they separate based on tense. Choice point 1 is between past tense forms and present tense forms. Choices between verb forms are visible at the subsequent levels. Of course, auxiliary verbs are difficult to categorize because they occur both as auxiliaries, and as stand-alone verbs.

The set of auxiliary verbs is also difficult because many of them encapsulate a negative (‘hadn’t’), and negatives represent an orthogonal category of choices. Figure 5 shows that systemic net when only the positive auxiliary verbs are considered. Again, tense is the dominant choice.

Figure 6 shows the systemic net for adverbs from a limited set of three different kinds: time, place, and frequency. This systemic net seems unclear, but note that at least some branches agree with intuition, for example the lower branch from choice four.

There are a very large number of adjectives used in the corpus, most of them only rarely. However, it is interesting to consider how adjectives might be empirically

[2] And a computational linguist might have chosen this separation as the most ‘natural’.

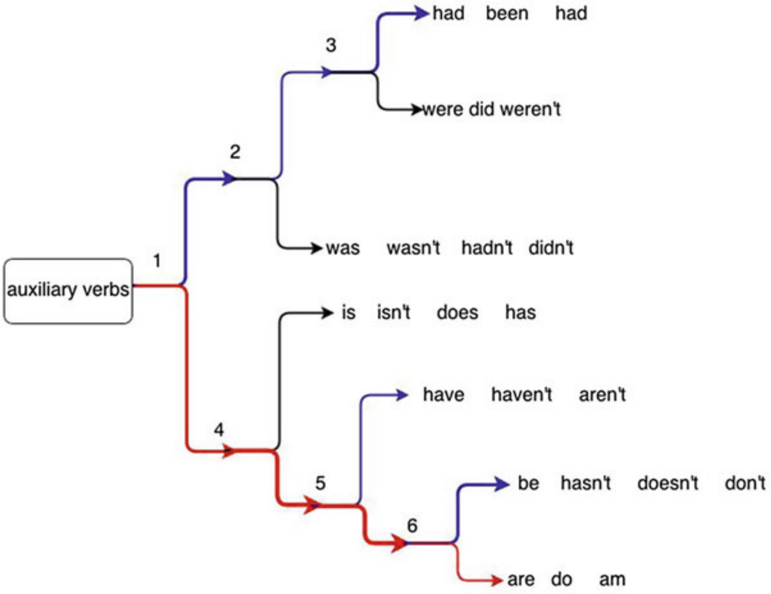
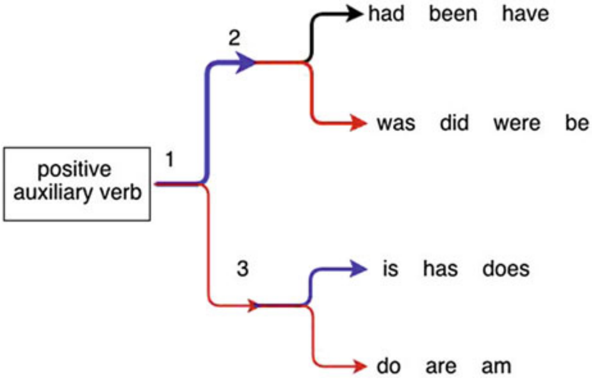


Fig. 4 Systemic net inferred for auxiliary verbs

Fig. 5 Systemic net inferred for positive auxiliary verbs



distinguished in fiction. (Note that this would not be the same net as the Appraisal Net described earlier, which might be inferrable from, say, a corpus of product reviews.) Figure 7 shows the systemic net for a limited set of adjectives of three kinds: appearance, color, and time. This net shows the typical structure for an extremely common word, in this case ‘good’ which appears as one outcome of the first choice. The sets of adjectives at each leaf are not those that would be conventionally grouped, but there are a number of interesting associations: ‘great’ and ‘large’ occur together, but co-occur with ‘black’ which is a plausible psychological association.

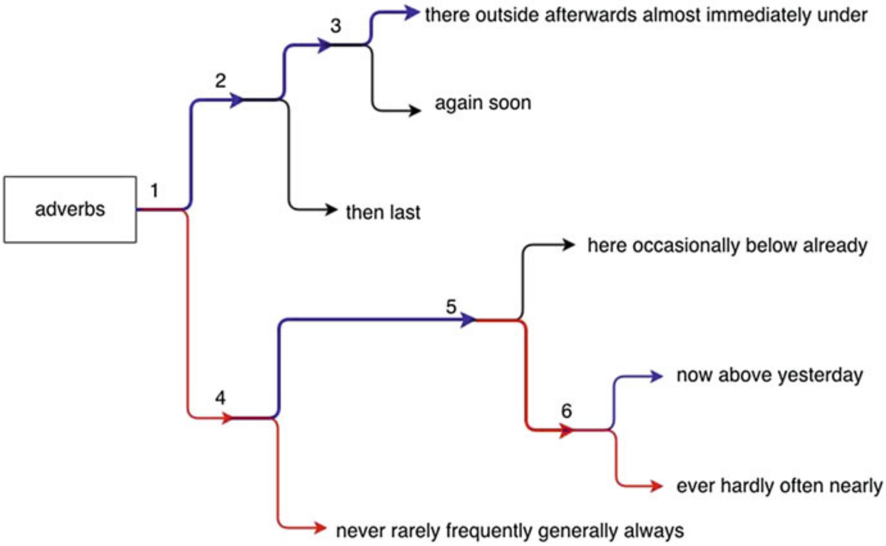

adverbs
1
2
3
there outside afterwards almost immediately under
again soon
then last
4
5
here occasionally below already
6
now above yesterday
ever hardly often nearly
never rarely frequently generally always

Fig. 6 Systemic net inferred for adverbs

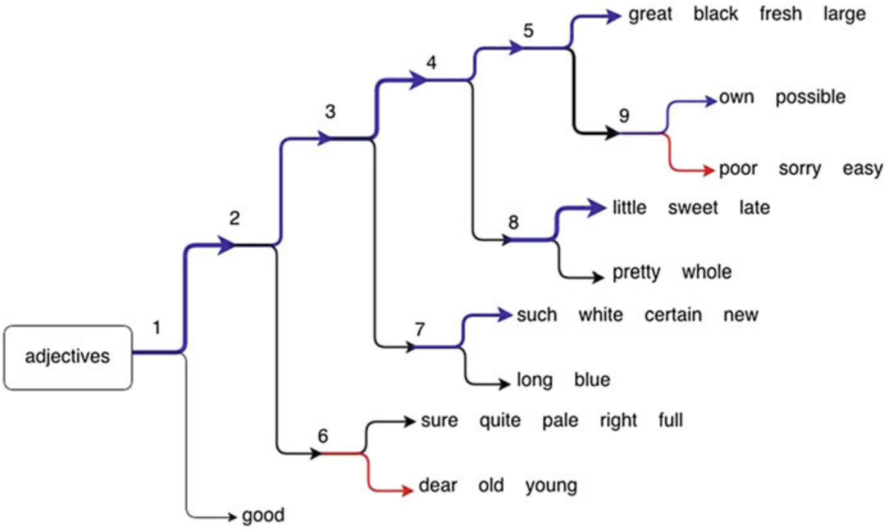


Fig. 7 Systemic net inferred for adjectives

These systemic nets look, from a human perspective, somewhere between plausible and peculiar. We now turn to more rigorous validation. Our goal is not so much that these nets should be explanatory from an intuitive perspective, but that they should be useful for analytic tasks (Fig. 8).

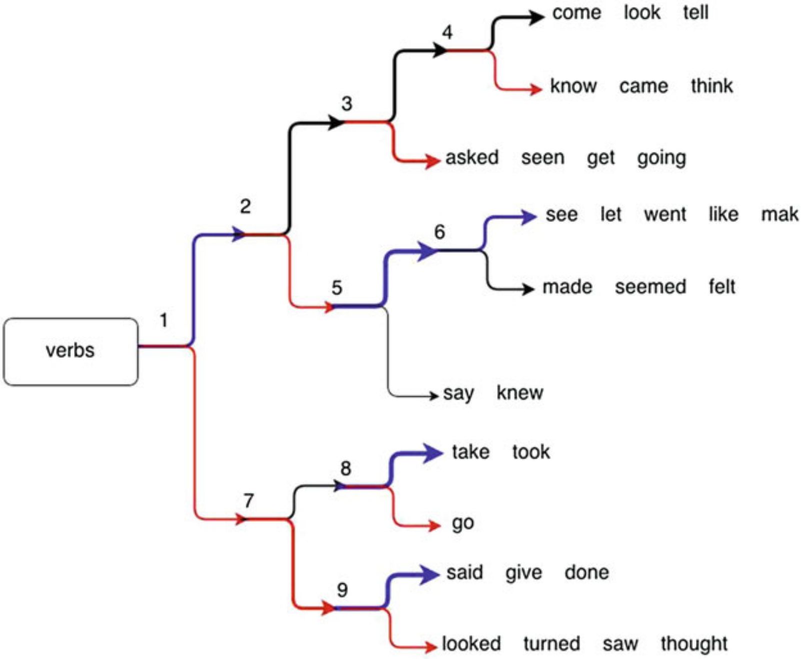


Fig. 8 Systemic net inferred for verbs

5 Validation

To validate our technique for inferring systemic nets, we use the following methods:

- Face validation. The systemic nets should involve choices that appear sensible and realistic. Note that this does not mean that they should match the hierarchy created to explain English grammar—such a grammar is an artificial construct intended to suggest consistent rules, and owing much to the grammar of Latin, rather than an accurate description of how English actually works.
- Comparison of document clustering based on word choices and based on systemic net choices. If choices reflect deeper structure, then documents should cluster more strongly based on choice structure than on word structure.
- Comparison of the performance of an example prediction task, authorship prediction, using word choices and systemic net choices. If choices reflect deeper structure, it should be easier to make predictions about documents based on choice structure than on word structure.
- Comparison with randomly created choice nets. Hierarchical clusterings with the same macroscopic structure as induced systemic nets should perform worse than the induced systemic nets.

5.1 Face Validation

The systemic nets shown in the previous section are not necessarily what a linguist might have expected, but it is clear that they capture regularities in the way words are used (especially in the domain of novels that was used, with their emphasis on individuals and their high rates of reported speech).

5.2 Clustering Using Word Choices Versus Net Choices

The difference between the systemic net approach and the bag-of-words approach is that they assume a different set of choices that led to the words that appear in each paragraph. The bag-of-words model implicitly assumes that each word was chosen independently; the systemic net model assumes that each word was chosen based on hierarchical choices driven by purpose, social setting, mental state, and language possibilities. Clustering paragraphs based on these two approaches should lead to different clusters, but those derived from systemic net choices should be more clear-cut. In particular, choices are not independent both because of hierarchy and because of the extrinsic constraints of the setting (novels, in this case)—so we expect to see clusters corresponding to registers.

We used two novels for testing purposes: *Robinson Crusoe* and *Wuthering Heights*, processed in the same way as our training data. Since these novels were not used to infer the systemic nets, results obtained using them show that the nets are capturing some underlying reality of this document class.

We compute the singular value decomposition of the paragraph-word matrix and the paragraph-choices matrix, both normalized by paragraph length. Plots show the resulting clustering of the paragraphs, with one test novel's paragraphs in red and the other in blue. In all of Figs. 9, 10, 11, and 12 the clustering derived from word frequencies is a single central cluster. In some of them, there appears to be

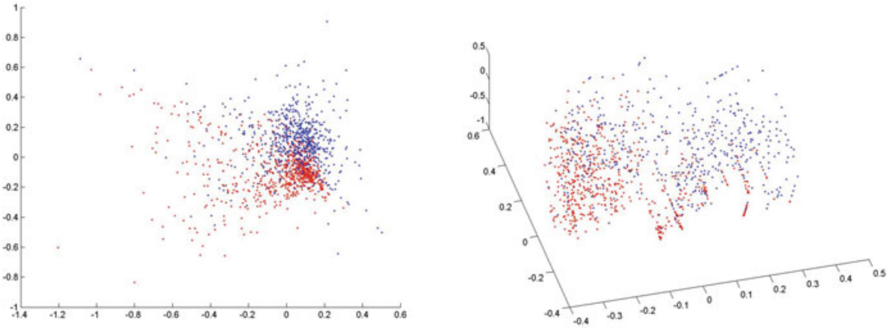


Fig. 9 SVD using pronouns, bag-of-words (left), choices (right)

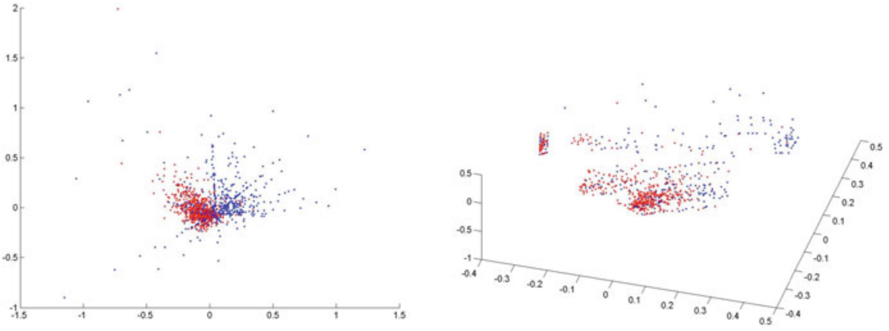


Fig. 10 SVD using auxiliary verbs, bag-of-words (left), choices (right)

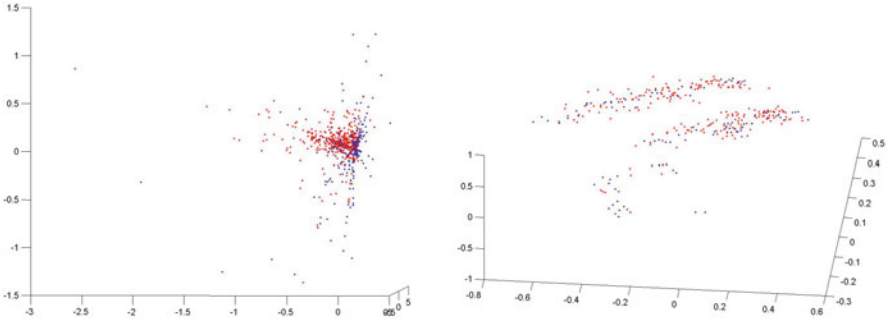


Fig. 11 SVD using adjectives, bag-of-words (left), choices (right)

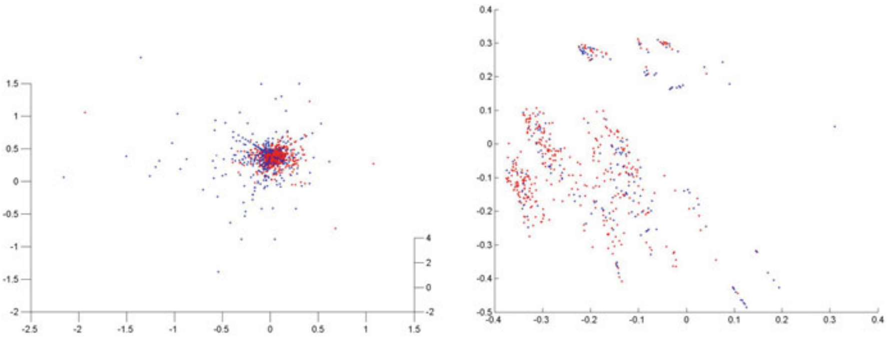


Fig. 12 SVD using verbs, bag-of-words (left), choices (right)

a separation between the two test documents, but these are illusions caused by overlays of points. In contrast, the clustering using choices shows strong clusters. These correspond to paragraphs that resulted from similar patterns of choices, that is to registers.

5.3 Authorship Prediction Using Word Choices Versus Net Choices

We argued that systemic nets are useful for applications where properties other than simple content are significant. To justify this claim we predict authorship *at the level of each individual paragraph* for our two test novels. This is a difficult task because paragraphs are so short; even humans would find it difficult to predict authorship at this level, especially without access to the semantics of the words used. Our goal is to show that the choice structure of the nets improves performance over simple use of bags of words. There are, of course, other ways to predict authorship, for example word n-grams or deep learning using LSTMs, but these are not directly comparable to systemic net approaches.

Again we use paragraph-word and paragraph-choice matrices as our data, and 5-fold cross-validated support vector machines with a radial basis kernel as the predictors. Results are shown for each of the word sets in Tables 2, 3, 4, 5, 6 and 7.

Across all of these word classes, authorship prediction based on word use hovers close to chance; in contrast, authorship prediction using systemic net choices range from accuracies of around 65%–75%, that is performance lifts of between 15 and 20 percentage points over prediction from word choices. And of these models is using only small numbers of words as signals of authorship. Clearly, the structural information coded in the systemic nets makes discrimination easier.

5.4 Inferred Nets Versus Randomly Generated Nets

Tables 8 and 9 compare the authorship prediction performance of the inferred systemic net and random networks constructed to have the same shape by dividing the words hierarchically into nested subsets of the same sizes as in the systemic net, but at random.

Table 2 Confusion matrices for personal pronouns; accuracy using words: 69.7%, accuracy using choices: 75.3%

	Predicted: words and choices			
Actual	RobCrusoe	WutHeights	RobCrusoe	WutHeights
RobCrusoe	694 (48%)	33 (2%)	584 (40%)	143 (10%)
WutHeights	407 (28%)	320 (22%)	216 (15%)	511 (35%)

Table 3 Confusion matrices for adverbs; accuracy using words: 51.3%, accuracy using choices: 63.4%

	Predicted: words and choices			
Actual	RobCrusoe	WutHeights	RobCrusoe	WutHeights
RobCrusoe	171 (12%)	556 (38%)	387 (27%)	340 (23%)
WutHeights	152 (10%)	575 (40%)	192 (13%)	535 (37%)

Table 4 Confusion matrices for auxiliary verbs; accuracy using words: 50.6%, accuracy using choices: 72.0%

	Predicted: words and choices			
Actual	RobCrusoe	WutHeights	RobCrusoe	WutHeights
RobCrusoe	435 (30%)	292 (20%)	553 (38%)	174 (12%)
WutHeights	426 (29%)	301 (21%)	233 (16%)	494 (34%)

Table 5 Confusion matrices for positive auxiliary verbs; accuracy using words: 51.4%, accuracy using choices: 67.6%

	Predicted: words and choices			
Actual	RobCrusoe	WutHeights	RobCrusoe	WutHeights
RobCrusoe	453 (30%)	292 (20%)	623 (43%)	104 (7%)
WutHeights	415 (29%)	312 (21%)	367 (25%)	360 (25%)

Table 6 Confusion matrices for adjectives; accuracy using words: 50.1%, accuracy using choices: 70.8%

	Predicted: words and choices			
Actual	RobCrusoe	WutHeights	RobCrusoe	WutHeights
RobCrusoe	295 (20%)	432 (30%)	490 (34%)	237 (16%)
WutHeights	294 (20%)	433 (30%)	187 (13%)	540 (37%)

Table 7 Confusion matrices for verbs; accuracy using words: 50.1%, accuracy using choices: 67.5%

	Predicted: words and choices			
Actual	RobCrusoe	WutHeights	RobCrusoe	WutHeights
RobCrusoe	297 (20%)	430 (30%)	476 (33%)	251 (17%)
WutHeights	296 (20%)	431 (30%)	221 (15%)	506 (35%)

Table 8 Personal pronouns: systemic network versus random nets

Number of	NNMF systemic network	Random nets		
paragraphs	Accuracy	min	mean	max
1	75.3%	69.3%	75.4%	82%
3	84.1%	68.1%	72.1%	76.2%
6	88.9%	66.3%	70%	71.5%

Table 9 Adjectives systemic network versus random nets

Number of	NNMF systemic network	Random nets		
paragraphs	Accuracy	min	mean	max
1	70.8%	70.2%	75.2%	79.4%
3	72.3%	66.9%	71.5%	73%
6	74.8%	64.5%	69.4%	72.3%

The performance of the random network is approximately the same as the inferred network at the level of single paragraph prediction. This is clearly a small sample size effect: choices that differentiate authors well are also available in the random network by chance. However, as the number of paragraphs available to make the prediction increases, the predictive performance of the systemic net continues to improve while that of the random network remains flat.

5.5 Combining Systemic Nets

We have built our systemic nets starting from defined word sets. In principle, a systemic net for all words could be inferred from a corpus. However, such a net would represent, in a sense, the entire language generation mechanism for English, so it is unlikely that it could be reliably built, and would require an enormous corpus.

However, it is plausible that the systemic nets we have built could be composed into larger ones, joining them together with an implied conjunctive choice at the top level. We now investigate this possibility.

One way to tell if such a composition is meaningful is to attempt the authorship prediction task using combined systemic nets. The results are shown in Table 10. The combined nets show a lift of a few percentage points over the best single net.

These results hint, at least, that complex systemic nets can be built by inferring nets from smaller sets of words, which can be done independently and perhaps robustly; and then composing these nets together to form larger ones. Some care is clearly needed: if the choice created by composing two nets interacts with the choices inside one or both of them, then the conjunctive composition may be misleading. This property is known as selectional restriction, and is quite well understood, so that it should be obvious when extra care is needed. For example, composing a net for nouns and one for adjectives using a conjunctive choice is unlikely to perform well because the choice of a noun limits the choice of adjectives that ‘match’ it.

Table 10 Prediction accuracy using combined word sets, best single systemic network, and combinations of systemic networks

	Words	Best single	Combined
Pronouns + adverbs	69%	75.3%	77.4%
Pronouns + adverbs + verbs	73.1%	75.3%	80.2%
Pronouns + adverbs + verbs + adjectives	80.37%	75.3%	80.44%

Table 11 Salafist-Jihadist words

Arabic	أمريكا	رب	ظالمين	رسول	عدو	يهود
English	America	God	Oppressors	Prophet	Enemy	Jews

6 Applying Systemic Nets for Intelligence Analysis

We now turn to applying the systemic net construction technique to text datasets that have intelligence value, documents created with islamist purposes of varying intensity.

A model of jihadist intensity developed by Koppel et al. [8] was designed to distinguish different strands of Islamic thoughts. We use the model (or set of words) describing a Salafist-jihadist orientation. The words were originally in Arabic, and were translated into English by the second author, a native Arabic speaker. The resulting list consists of 144 words. Table 11 shows some of words.

We show that a structure derived from three English-language islamist propaganda magazines, *Inspire*, *Dabiq* and *Azan*, generalizes elegantly to two islamist forums, showing that there is a widely held mindset (or set of distinctions) shared in this worldwide community.

7 Measuring Islamist Language

One way to measure the jihadi intensity of a given document is simply to sum the frequency with which relevant words occur. This approach has been widely used by, for example, Pennebaker to measure a number of properties [11] and the LIWC package has made this technology available to many researchers. This approach has been used to measure deception [20], informative language [23], imaginative language [23], and propaganda [3, 27, 28].

We examine three different jihadist magazines, all with the same professed goals, but originating from different countries, and from groups with different ideologies. Inspire is produced by Al Qaeda in the Arabian Peninsula. The first nine issues were edited by the American jihadist, Anwar al-Awlaki. Since his death, others, so far unidentified, have taken over. Dabiq is produced by ISIS (Daish) in Syria. Azan is produced by the Taliban in Pakistan. All three types of magazines are high in production values, use many visual images, and aim to imitate the look and feel of mainstream Western magazines. All of these magazines appear as pdfs; more recent issues have become so complex that it is impossible to extract the textual content using OCR but the text of 12 issues of Inspire, 5 issues of Azan, and 5 issues of Dabiq have been extracted. Skillicorn used the magazine data to do an empirical assessment of the intensity of propaganda across the three magazines [3].

A document-word matrix for the magazines was constructed by counting the frequencies of words from the jihadi language model, normalized as discussed

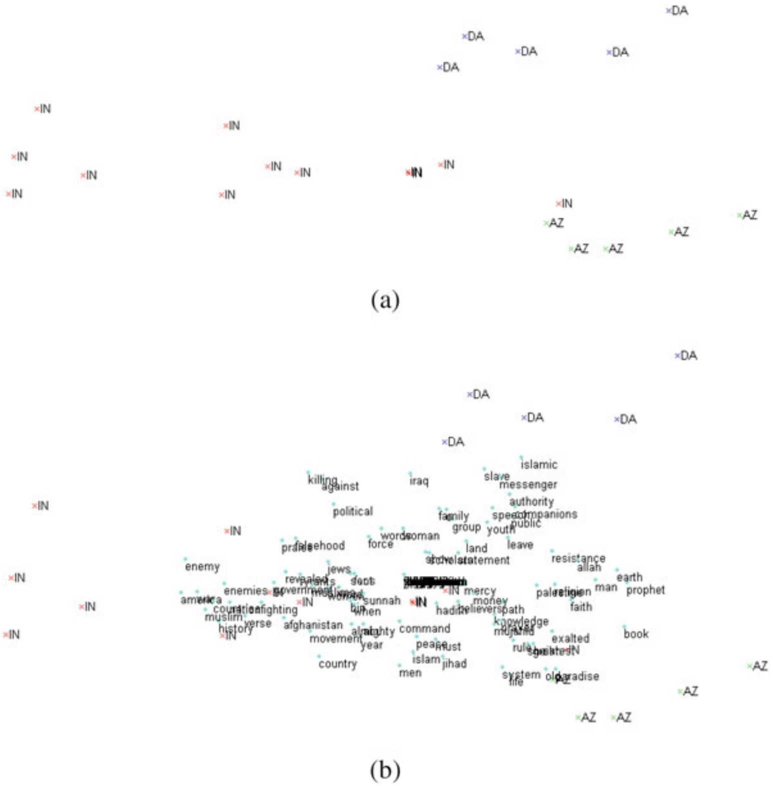


Fig. 13 Similarity among magazines using SVD (IN: Inspire, AZ: Azan, DA: Dabiq. (**a**) A plot of the magazines based on the jihadi language model. (**b**) A plot of the magazines based on the jihadi language model overlaid with the words, i.e. the U and V matrices plotted together

above. The SVD plots of the document-word matrix based on the jihadi words are shown in Fig. 13. Both Azan and Dabiq cluster strongly. Inspire does not cluster as well, suggesting that it does not have a consistent style or content focus. Figure 13b shows the magazines overlaid with the words they use—words and magazines can be considered to be pulled towards one another whenever a particular word is heavily used in a particular magazine issue. Some of the words most associated with Dabiq, therefore, are: ‘Iraq’, ‘Islamic’, ‘slave’, and ‘authority’. This seems reasonable since ISIS is active in Syria and Iraq. ISIS also allows the practice of slavery. Some of the words most associated with Azan are: ‘Paradise’, ‘life’, ‘old’, and ‘system’. These words suggest that Azan’s message is more focused towards the afterlife.

8 A Systemic Net Based on Jihadi Language Model

We use the magazines to construct a systemic net derived from jihadi language. Figure 14 shows the resulting systemic net labelled with the choice points for later reference. The tree has six choice points and six leaves, because the words at node 7 do not split further.

The first choice in the tree reflects a clear distinction between political and religious words. A sample of the words associated with this choice is given in Table 12.

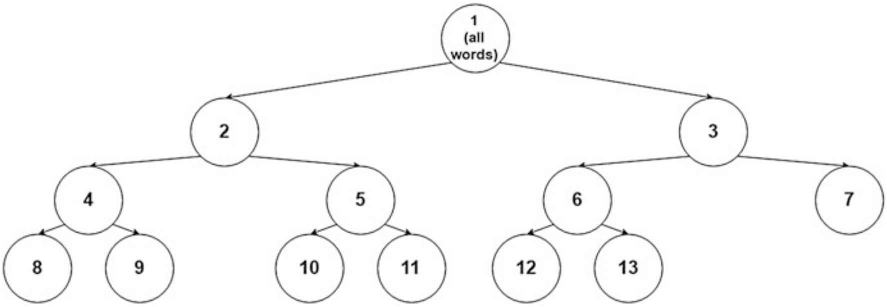


Fig. 14 A visualization of the systemic net

Table 12 First choice words

First choice	
Religious	Political
Prophet	Tyrants
Sheikh	Enemy
God	Fighting
Worlds	America
Paradise	Brotherhood
Exalted	Government
Earth	Palestinian
Monotheism	Oppressors
Goodness	Nation
Platform	Categories
Islamic	Act
Faith	Movement
Old	Nation
Owners	War
Allah	Doubt
Allowing	Country
Prayer	Ruler
Companions	Afghanistan

Table 13 Basic meaning behind each leaf node and the Google results obtained from using the words in each node

Leaf	Meaning	Google results
8	Used by inspire the most	Verses of Quran on Jihad–Islam
9	Used by Dabiq the most	Allah’s Quran—authenticity of the Quran Muhammad, Terrorist or Prophet?—Bible Probe Islam and antisemitism—Wikipedia, the free encyclopedia what every non-Muslim needs to know about Islam!—Bible.ca
10	Jihad focused	Islamic State and the Others · Raqqa is Being Slaughtered How Islam will dominate the world — - Duaat - WordPress.com Chapter 1: Muhammad and the Quran
11	Pure religion	Prophet Muhammad, pbuh - Some selected verses Does Islam regard non - Muslims with mercy and compassion The Book of Faith - Sahih Muslim - Sunnah.com
12	Al-qaeda and Afghanistan focused	Islam in Afghanistan - Wikipedia, the free encyclopedia Afghan Arabs - Wikipedia, the free encyclopedia Al-Qaeda - Infoplease
13	Teachings about Islam	Contemporary Islamist ideology authorizing genocidal murder Full text of “Islamic Books by Ibn Taymiyyah Maqdisi” Do the authentic teachings of Islam result in terrorism? Welcome to IONA masjid and learning center!—IONA Masjid !

The word sets resulting from some of the choices make immediate intuitive sense, while others do not. As a way to understand what each choice is capturing, we take the word sets from each leaf, and treat them as terms for a Google query. The top ranked documents associated with each of them are shown in Table 13. Most sets can be assigned a plausible meaning based on these results, the exception being node 8.

We can now compare how the magazines cluster based on bag of words versus based on choice sets. An SVD plot of the variation between the magazines based on choices is shown in Fig. 15. Compared to Fig. 13, Azan and Dabiq cluster more tightly based on choices than on words. There is no significant difference between the two approaches in how Inspire issues cluster. The color and shape coding of the magazines is based on which choices are made most often at choice points in the systemic net. Inspire tends to prefers the political branch over the religious branch, but does not cluster well. This reflects the wide variation in focus that has been previously noted, and perhaps the changes in editorship and authorship. Both Azan and Dabiq have consistent choice patterns across all issues, suggesting clarity of purpose, and a consistent editorial framework. Choices 7 and 9 are strongly associated with Dabiq, while Inspire tends to favor the opposite choices (6 and 8).

This analysis allows us to associate particular sets of magazines with particular patterns of word choice, and therefore to take a first step towards judging group

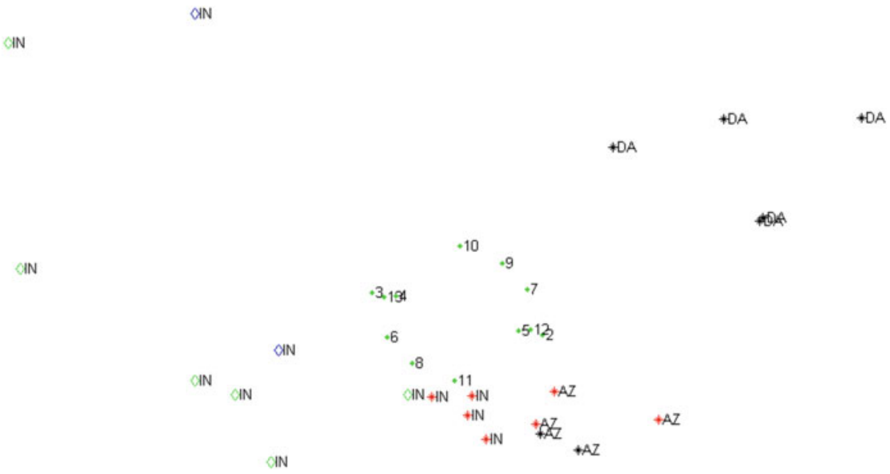


Fig. 15 SVD plot of the document-choice matrix of the magazines (diamonds: prefer political choice; asterisks: prefer religious choice; green and blue distinguish the outcome from choice point 2; black and red distinguish the outcome from choice point 3; numbers are the positions corresponding to the choices, i.e. the columns of the matrix)

intent. For example, it is possible to infer, in principle, whether a group's focus is internal or external, and if external what kind of target is likely to seem most attractive. This goes deeper than simply observing which words are frequent, because it associates words that are related in the sense that they occupy the same mental 'slot'.

9 Applying the Systemic Net to Islamist Forums

We apply the jihadi language systemic net that was inferred from the magazines to two new corpora, two islamist forums:

- Turn to Islam: which advertises itself as "correcting the common misconceptions about Islam".
- Islamic Awakening: which identifies itself as "dedicated to the blessed global Islamic awakening".

Turn to Islam (TTI) consists of 335,388 posts from 41,654 members collected between June 2006 and May 2013. Islamic Awakening (IA) consists of 201,287 posts from 3964 members collected between April 2004 and May 2012. Both data sets were collected by the University of Arizona Artificial Intelligence Lab [2]. The posts are primarily in English, but with a mixture of transliterated Arabic, some French, and a small number of words from other European languages.

Figure 16a shows the variation among a 10% uniformly random sample of posts from the TTI forum, based on the Jihadi language systemic net choices. The color and shape coding is the same as in Fig. 15. We can see a clear separation between posts making political versus religious word choices (diamonds vs asterisks). The separation also extends to the choice points in the second layer of the SFL net (blue/green and red/black). Figure 16b shows the cloud of points rotated so that the third dimension is visible, showing that the variation between red and black points is orthogonal to the variation between blue and green points. The striking point is that the choices inferred from word usage in the islamist magazines strongly and consistently cluster forum posts coming from a completely different context.

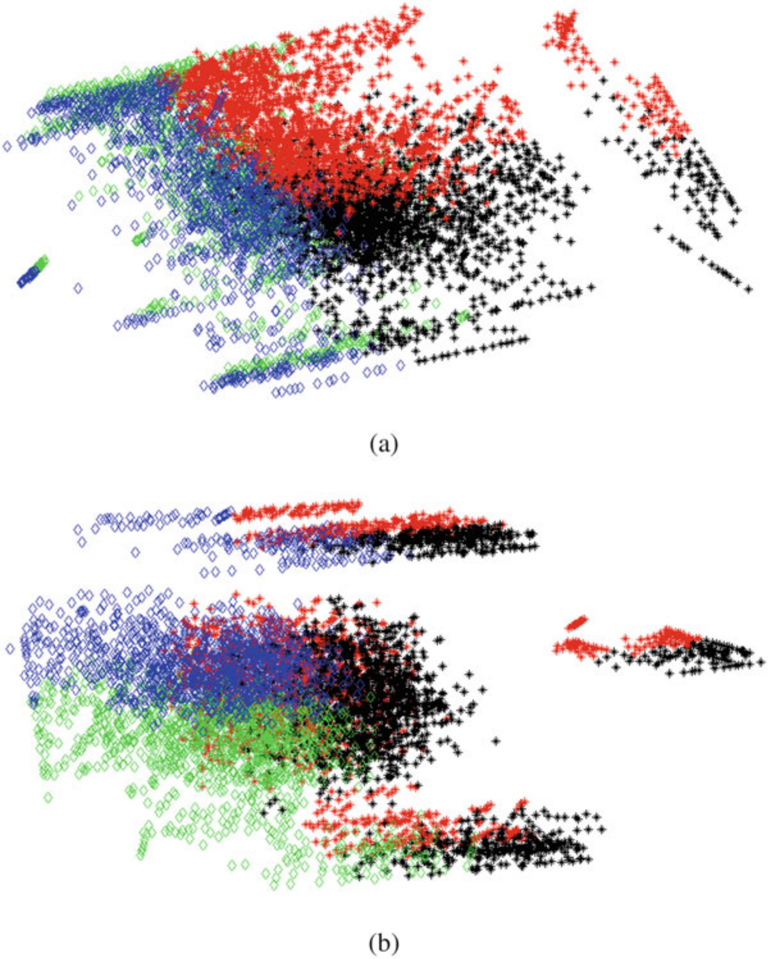
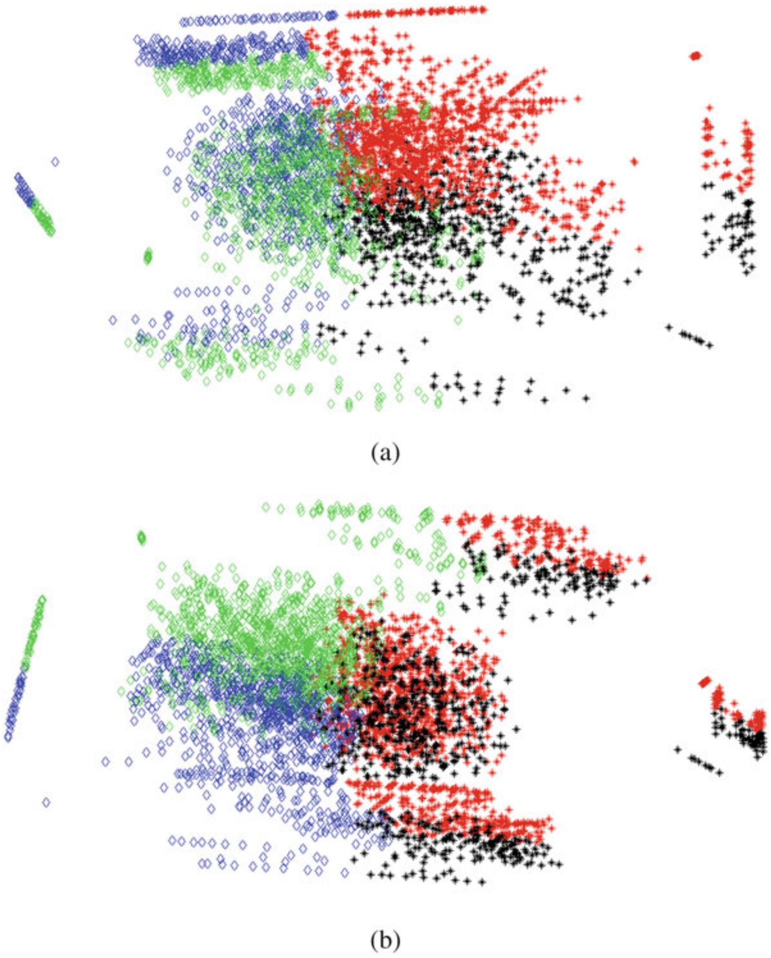


Fig. 16 SVD plot of the document-choice matrix of TTI posts (symbol and color coding as in Fig. 15). (**a**) First two dimensions. (**b**) Rotated view to show the third dimension



(a)

(b)

Fig. 17 SVD plot of the document-choice matrix of IA posts with the same symbol and color coding as Fig. 15. (**a**) First two dimensions. (**b**) Rotated view to show the third dimension

This suggests that there is a widely shared mindset in this community, interpreted broadly, that produces consistent language use across settings.

Figure 17 is the same analysis for the IA forums. Both TTI and IA cluster strongly and consistently based on the choice structure of the systemic net inferred from the magazines.

The word sets that result from choice point 5 are particularly interesting; they distinguish between two types of religious thinking, one that might be called purely religious and the other which is focused more on the jihad aspect of religion. The relevant words are shown in Table 14.

Figure 18 shows an SVD plot of TTI posts color-coded by Jihadi intensity which we obtain by adding two artificial documents that contain all of the words of the

Table 14 Choice point 5 words

Fifth choice point	
Pure religion	Jihad focused
Old	Behalf
Command	Hide
Exalted	Islamic
Folk	Monotheism
Mohammed	Authority
Allah	Owners
Peace	Jews
Believers	Resistance
Goodness	Companions
Faith	Woman
	Family
	Worlds
	Earth

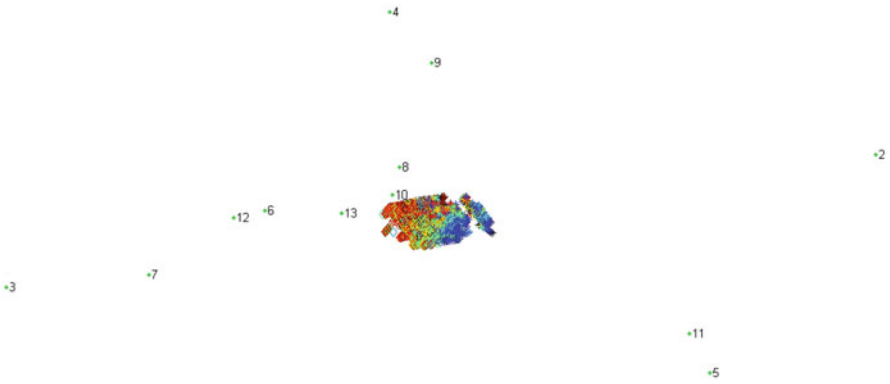


Fig. 18 SVD plot of TTI document-choice matrix. The posts are color coded based on jihadi intensity from blue (least intense) to red (most intense)

model at frequencies one standard deviation above, and one standard deviation below the mean, and using the line between them as a gradient of jihadi intensity. Blue points are the least jihadi and red are the most. Choice points from the systemic net are also included, and it is clear that the word sets that distinguish jihadi intensity most directly are those at nodes 10 and 11. Therefore, the choice made at choice point 5 could be also used, by itself, as a predictor of jihadi intensity.

10 Discussion

The strength of the religious vs political choice at the first choice point of the systemic net suggests that there is a fundamental differentiator among those who engage in islamist discussions or writings. It appears that islamist ideology can be plausibly separated into two threads, and this generalizes over different contexts and widely differing authors. It remains an open question whether the authors themselves are consciously aware of this; and whether an understanding of the dichotomy could be leveraged to increased the effectiveness of propaganda vehicles such as the magazines (and, for some posters, the forums). There is also a strong distinction in the religious domain between word choices that are, as it were, purely religious and those that are religious but with a jihadist subtext.

We have also demonstrated the effectiveness of systemic nets, and the choices they capture. The structure inferred from large, well-written islamist magazines generalizes very well to a completely different domain: short, informal posts in online forums.

Methodologically, we have shown that inferring systemic nets from data produces structures that reflect underlying language patterns, even though the word choice sets do not necessarily have a direct interpretation. The ability to infer systemic nets automatically, even if they are possibly not as accurate as those inferred by humans, opens up the SFL approach to many more application domains, of which intelligence analysis is just one.

References

1. S. Argamon, S. Dhawle, M. Koppel, J.W. Pennebaker, Lexical predictors of personality type, in *Proceedings of the Joint Annual Meeting of the Interface and Classification Society of North America* (2005)
2. Azure, Dark web forums. http://www.azsecure-data.org/dark-web-forums.html. AZSecure-data.org version (Accessed May 4th, 2016)
3. H. Borko, M. Bernick, Automatic document classification. J. ACM **10**(3), 151–162 (1963)
4. H. Chen, F.-Y. Wang, Artificial intelligence for homeland security. IEEE Intell. Syst. **20**(5), 12–16 (2005)
5. T. Coffman, S. Greenblatt, S. Marcus, Graph-based technologies for intelligence analysis. Commun. ACM **47**(3), 45–47 (2004)
6. D. Cook, L.B. Holder, Graph-based data mining. IEEE Intell. Syst. **15**(2), 32–41 (2000)
7. J.L. Creasor, D.B. Skillicorn, *QTagger: Extracting Word Usage from Large Corpora* (Queen's University, School of Computing, Kingston, 2012). Technical Report 2012-587
8. G.S. Davidson, B. Hendrickson, D.K. Johnson, C.E. Meyers, B.N. Wylie, Knowledge mining with VxInsight : discovery through interaction. J. Intell. Inf. Syst. **11**(259-285) (1998)
9. E.C. Davies, A retrospective view of Systemic Functional Linguistics, with notes from a parallel perspective. Funct. Linguistics **1**(1), 4 (2014)
10. J. Galloway, S. Simoff, Network data mining: discovering patterns of interaction between attributes, in *Advances in Knowledge Discovery and Data Mining*. Springer Lecture Notes in Computer Science, vol. 3918 (2006), pp. 410–414
11. Google WebAPI (2004). www.google.com/apis
12. M.A.K. Halliday, J.J Webster, *Bloomsbury Companion to Systemic Functional Linguistics*. Continuum Companions (Bloomsbury Academic, London, 2009)
13. M. Herke-Couchman, J. Patrick, Identifying interpersonal distance using systemic features, in *Proceedings of AAAI Workshop on Exploring Attitude and Affect in Text: Theories and Applications* (Springer, Netherlands, 2004), pp. 199–214

14. H. Kanayama, T. Nasukawa, H. Watanabe, Deeper sentiment analysis using machine translation technology, in *Proceedings of the 20th International Conference on Computational Linguistics* (2004)
15. A. Kappagoda, The use of systemic-functional linguistics in automated text mining, in *DSTO Defence Science and Technology Organisation DSTO-RR-0339* (2009). Technical report
16. C.E. Lamb, D.B. Skillicorn, Detecting deception in interrogation settings, in *IEEE International Conference on Intelligence and Security Informatics* (2013), pp. 160–162
17. M. Lazaroff, D. Snowden, Anticipatory models for counter-terrorism, in *Emergent Information Technologies and Enabling Policies for Counter-terrorism*, ed. by R.L. Popp, J. Yen, chapter 3, IEEE Press Series on Computational Intelligence (2006), pp. 51–73
18. D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
19. C. Matthiessen, M.A.K. Halliday, *Systemic Functional Grammar: A First Step into the Theory* (Macquarie University, Macquarie Park, 1997)
20. M.L. Newman, J.W. Pennebaker, D.S. Berry, J.M. Richards, Lying words: predicting deception from linguistic styles. Personal. Soc. Psychol. Bull. **29**(5), 665–675 (2003)
21. J. Patrick, The scamseek project—text mining for financial scams on the internet, in *Data Mining: Proceedings of the 4th Australasian Workshop on Data Mining*. Springer LNCS 3755 (2006), pp. 295–302
22. J. Qin, J. Xu, D. Hu, M. Sageman, H. Chen, Analyzing terrorist networks: a case study of the global Salafi Jihad network, in *Intelligence and Security Informatics, IEEE International Conference on Intelligence and Security Informatics, ISI 2005, Atlanta, GA, USA, May 19-20*. Lecture Notes in Computer Science LNCS 3495 (Springer, Berlin, 2005), pp. 287–304
23. P. Rayson, A. Wilson, G. Leech, Grammatical word class variation within the British National Corpus sampler. Lang. Comput. **36**(1), 295–306 (2001)
24. M. Sageman, *Understanding Terror Networks* (University of Pennsylvania Press, Philadelphia, 2004)
25. A.P. Sanfilippo, A.J. Cowell, S.C. Tratz, A.M. Boek, A.K. Cowell, C. Posse, L.C. Pouchard, Content analysis for proactive intelligence: marshalling frame evidence, in *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence* (2006), pp. 919–924
26. S. Scott, S. Matwin, Text classification using WordNet hypernyms, in *Natural Language Processing Systems: Proceedings of the Conference. Association for Computational Linguistics Somerset, New Jersey* (1998), pp. 38–44
27. D.B. Skillicorn, Lessons from a jihadi corpus, in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE, Piscataway, 2012), pp. 874–878
28. D.B. Skillicorn, E. Reid, Language use in the jihadist magazines Inspire and Azan. Secur. Inform. **3**(1), 9 (2014)
29. R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (2013), pp. 1631–1642
30. Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**, 24–54 (2010)
31. C. Whitelaw, S. Argamon, Systemic functional features in stylistic text classification, in *Proceedings of AAAI Fall Symposim on Style and Meaning in Language, Art, Music, and Design, Washington, DC* (2004)
32. C. Whitelaw, N. Garg, S. Argamon, Using appraisal taxonomies for sentiment analysis, in *Second Midwest Computational Linguistic Colloquium (MCLC 2005)* (2005)
33. Y. Zhang, S. Zeng, L. Fan, Y. Dang, C.A. Larson, H. Chen, Dark web forums portal: Searching and analyzing jihadist forums, in *IEEE International Conference on Intelligence and Security Informatics* (2009), pp. 71–76