# Semantics from Narrative

Fionn Murtagh *, ** and Adam Ganz ***

* Science Foundation Ireland, Wilton Place, Dublin 2, Ireland

** Department of Computer Science, *** Department of Media Arts
Royal Holloway, University of London, Egham TW20 0EX, England

## Abstract

We study two aspects of information semantics: (i) the collection of all relationships, (ii) tracking and spotting anomaly and change. The first is implemented by endowing all relevant information spaces with a Euclidean metric in a common projected space. The second is modelled by an induced ultrametric. A very general way to achieve a Euclidean embedding of different information spaces based on cross-tabulation counts (and from other input data formats) is provided by Correspondence Analysis. From there, the induced ultrametric that we are particularly interested in takes a sequential e.g. temporal ordering of the data into account. We exemplify this with filmscript related to cinema and television.

**Keywords:** text analysis, Correspondence Analysis, contiguity-constrained hierarchical clustering, filmscript, semantic analysis

## 1. Analysis of Narrative

### *1.1. Introduction*

The data mining and data analysis challenges addressed are the following.

(i) Great masses of data, textual and otherwise, need to be exploited and decisions need to be made. Correspondence Analysis handles multivariate numerical and symbolic data with ease.

(ii) Structures and interrelationships evolve in time.

(iii) We must consider a complex web of relationships.

(iv) We need to address all these issues from data sets and data flows. We will look at how this works, using the Casablanca film script, and continue with scripts from a television series.

### *1.2. The Changing Nature of Movie and Drama*

McKee (1999) bears out the great importance of the film-script: “50% of what we understand comes from watching it being said”. And: “A screenplay waits for the camera. [...] Ninety percent of all verbal expression has no filmic equivalent”.

An episode of a television series costs US$2-3 million per one hour of television. Generally screenplays are written speculatively or commissioned, and then prototyped by the full production of a pilot episode. Increasingly, and especially availed of by the young, television series are delivered via the Internet.

Originating in one medium – cinema, television, game, online – film and drama series are increasingly migrated to another. So scriptwriting must take account of digital multimedia platforms. This has been referred to in computer networking parlance as "multiplay" and in the television media sector as a "360 degree" environment.

Cross-platform delivery motivates interactivity in drama. So-called reality TV has a considerable degree of interactivity, as well as being largely unscripted.

There is a burgeoning need for us to be in a position to model the semantics of film script, – its most revealing structures, patterns and layers. With the drive towards interactivity, we also want to leverage this work towards more general scenario analysis. Potential applications are to business strategy and planning; education and training; and science, technology and economic development policy.

### *1.3. Correspondence Analysis: a Semantic Analysis Plat-form*

For McKee (1999), film-script text is the "sensory surface of a work of art" and reflects the underlying emotion or perception. Our data mining approach models and tracks these underlying aspects in the data. Our approach to textual data mining has a range of novel elements.

Firstly, a novelty is our focus on the orientation of narrative through Correspondence Analysis (Benzécri, 1979; Murtagh, 2005; Le Roux and Rouanet, 2004; Lebart and Salem, 1994), which maps scenes (and subscenes), and words used, in a near fully automated way, into a Euclidean space representing all pairwise interrelationships. Such a space is ideal for visualization. Interrelationships between scenes are captured and displayed, as well as interrelationships between words, and mutually between scenes and words.

The starting point for analysis is frequency of occurrence data, typically the ordered scenes crossed by all words used in the script.

If the totality of inter-relationships is one facet of semantics, then another is anomaly or change as modelled by a clustering hierarchy. If, therefore, a scene is quite different from immediately previous scenes, then it will be incorporated into the hierarchy at a high level. This novel view of hierarchy will be discussed further in section 1.5 below.

We draw on these two vantage points on semantics – viz. totality of interrelationships, and using a hierarchy to express change.

Among further work that we report on in Murtagh et al. (2009) is the following. We devise a Monte Carlo approach to test statistical significance of the given script's patterns and structures as opposed to randomized alternatives (i.e. randomized realizations of the scenes). Alternatively we examine caesuras and breakpoints in the film-script, by taking the Euclidean embedding further and inducing an ultrametric on the sequence of scenes.

### *1.4. Casablanca Narrative: Illustrative Analysis*

The well known Casablanca movie serves as an example for us. Film-scripts, such as for Casablanca, are partially structured texts. Each scene has metadata and the body of the scene contains dialog and possibly other descriptive data. The Casablanca script was half completed when production began in 1942. The dialog for some scenes was written while shooting was in progress. Casablanca was based on an unpublished 1940 screenplay (Burnett and Alison, 1940). It was scripted by J.J. Epstein, P.G. Epstein and H. Koch. The film was directed by M. Curtiz and produced by H.B. Wallis and J.L. Warner. It was shot by Warner Bros. between May and August 1942.

As an illustrative first example we use the following. A data set was constructed from the 77 successive scenes crossed by attributes – Int[erior], Ext[erior], Day, Night, Rick, Ilsa, Renault, Strasser, Laszlo, Other (i.e. minor character), and 29 locations. Many locations were met with just once; and Rick's Café was the location of 36 scenes. In scenes based in Rick's Café we did not distinguish between "Main room", "Office", "Balcony", etc. Because of the plethora of scenes other than Rick's Café we assimilate these to just one, "other than Rick's Café", scene.

In Fig. 1, 12 attributes are displayed; 77 scenes are displayed as dots (to avoid over-crowding of labels). Approximately 34% (for factor 1) + 15% (for factor 2) = 49% of all information, expressed as inertia explained, is displayed here. We can study interrelationships between characters, other attributes, scenes, for instance closeness of Rick's Café with Night and Int (obviously enough).

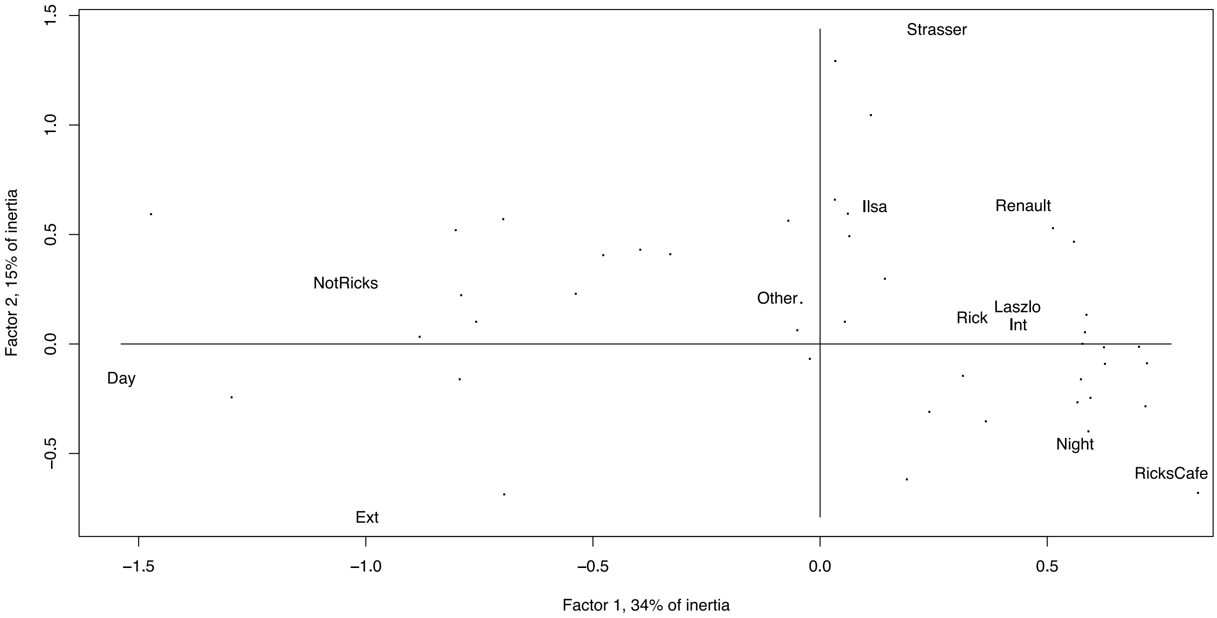


*Figure 1: Correspondence Analysis of the Casablanca data derived from the script. The input data is presences/absences for 77 scenes crossed by 12 attributes. The 77 scenes are located at the dots, which are not labelled here for clarity*

### *1.5. The Geometry and Topology of Information*

Some underlying principles are as follows. We start with the cross-tabulation data, scenes x attributes. Scenes and attributes are embedded in a metric space. This is how we are probing the *geometry of information*, which is a term and viewpoint used by van Rijsbergen (2004).

We come now to a different principle: that of the *topology of information*. The particular topology used is that of hierarchy. Euclidean embedding provides a very good starting point to look at hierarchical relationships. An innovation in our work is as follows: the hierarchy takes sequence, e.g. timeline, into account. This captures, in a more easily understood way, the notions of novelty, anomaly or change.

Fig. 2, left, illustrates this situation, where the anomalous or novel point is to the right. The further away the point is from the other data then the better is this approximation (Murtagh, 2004). The strong triangular inequality, or ultrametric inequality, holds for tree distances: see Fig. 2 (right).

Fig. 3 uses a sequence-constrained complete link agglomerative algorithm. It shows up scenes 9 to 10, and progressing from 39, to 40 and 41, as major changes. The sequence constrained algorithm, i.e. agglomerations are permitted between adjacent segments of scenes only, is described in Murtagh (2005). The agglomerative criterion used, that is subject to this sequence constraint, is a complete link one.

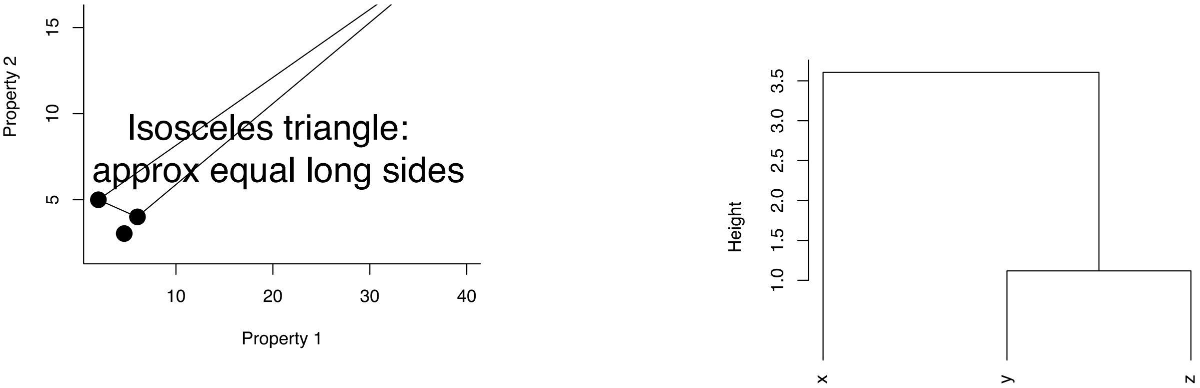


*Figure 2: Left: a stylized view – consider a query point on the far right, and consider the approximate closeness relationship. Right: The strong triangular inequality defines an ultrametric: every triplet of points satisfies the relationship: d(x, z) ≤ max{d(x, y), d(y, z)} for distance d. In addition the symmetry and positive definiteness conditions hold for any pair of points*

### *1.6. Our Platform for Analysis of Semantics*

Correspondence analysis supports the following: (i) analysis of multivariate, mixed numerical/ symbolic data; (ii) web of interrelationships; and (iii) evolution of relationships over time.

Correspondence Analysis is in practice "a tale of three metrics" (Murtagh, 2005). The analysis is based on embedding a cloud of points from a space governed by one metric into another. Furthermore, the cloud offers vantage points of both observables and their characterizations, so – in the case of film-script – for any one of the metrics we can effortlessly pass between the space of filmscript scenes and attribute set. The three metrics are as follows.

- Chi squared, $\chi^2$, metric – appropriate for profiles of frequencies of occurrence.
- Euclidean metric, for visualization, and for static context.
- Ultrametric, for hierarchic relations and, as we use it in this work, for dynamic context.

In the analysis of semantics, we distinguish two separate aspects.

1. Context – the collection of all interrelationships: The Euclidean distance makes a lot of sense when the population is homogeneous. All interrelationships together provide context, relativities – and hence meaning.
2. Hierarchy tracks anomaly: Ultrametric distance makes a lot of sense when the observables are heterogeneous, discontinuous. The latter is especially useful for determining: anomalous, atypical, innovative cases.

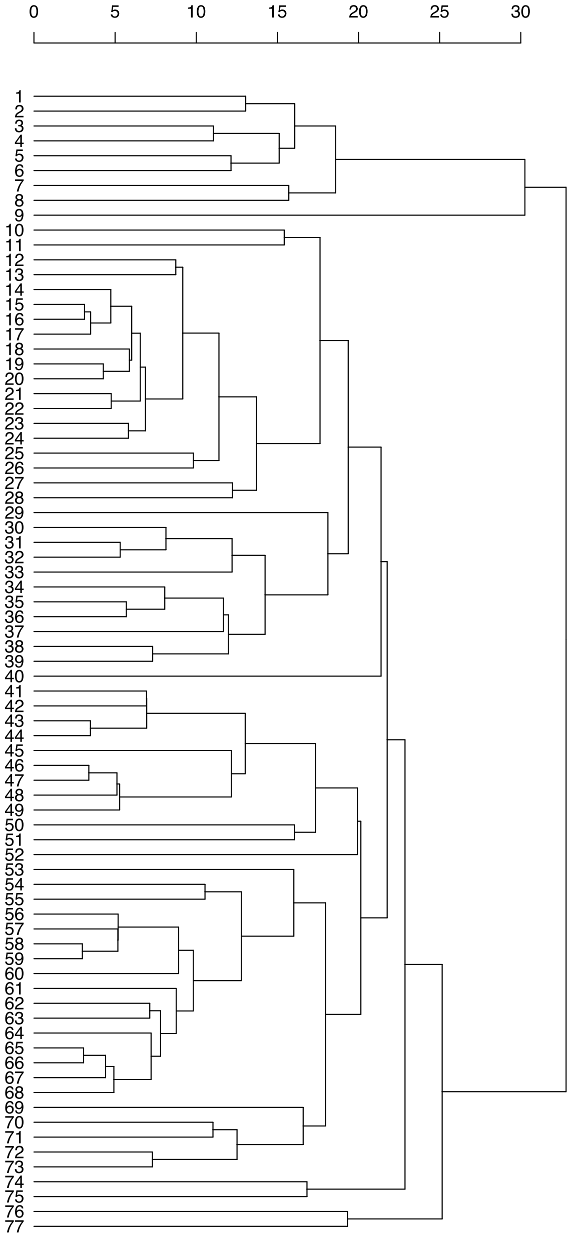


*Figure 3: 77 scenes clustered. These scenes are in sequence: a sequence-constrained agglomerative criterion is used for this. The agglomerative criterion itself is a complete link one. See Murtagh (1985) for properties of this algorithm*

## 2. Deeper Look at Semantics in Casablanca Script

### *2.1. Text Data Mining*

The Casablanca script has 77 successive scenes. In total there are 6.710 words in these scenes. We define words as consisting of at least two letters. Punctuation is first removed. All upper case is set to lower case. We use from now on all words. We analyze frequencies of occurrence of words in scenes, so the input is a matrix crossing scenes by words.

### *2.2. Analysis of a Pivotal Scene, Scene 43*

As a basis for a deeper look at Casablanca we have taken comprehensive but qualitative discussion by McKee (1999) and sought quantitative and algorithmic implementation.

Casablanca is based on a range of miniplots. For McKee its composition is “virtually perfect”.

Following McKee (1999), we carry out an analysis of Casablanca’s “Mid-Act Climax”, Scene 43, subdivided into 11 “beats”. McKee divides this scene, relating to Ilsa and Rick seeking black market exit visas, into 11 “beats”.

1. Beat 1 is Rick finding Ilsa in the market.
2. Beats 2, 3, 4 are rejections of him by Ilsa.
3. Beats 5, 6 express rapprochement by both.
4. Beat 7 is guilt-tripping by each in turn.
5. Beat 8 is a jump in content: Ilsa says she will leave Casablanca soon.
6. In beat 9, Rick calls her a coward, and Ilsa calls him a fool.
7. In beat 10, Rick propositions her.
8. In beat 11, the climax, all goes to rack and ruin: Ilsa says she was married to Laszlo all along. Rick is stunned.

Fig. 4 shows the evolution from beat to beat rather well. 210 words are used in these 11 “beats” or subscenes. Beat 8 is a dramatic development. Moving upwards on the ordinate (factor 2) indicates emotional distance between Rick and Ilsa. Moving downwards indicates rapprochement.

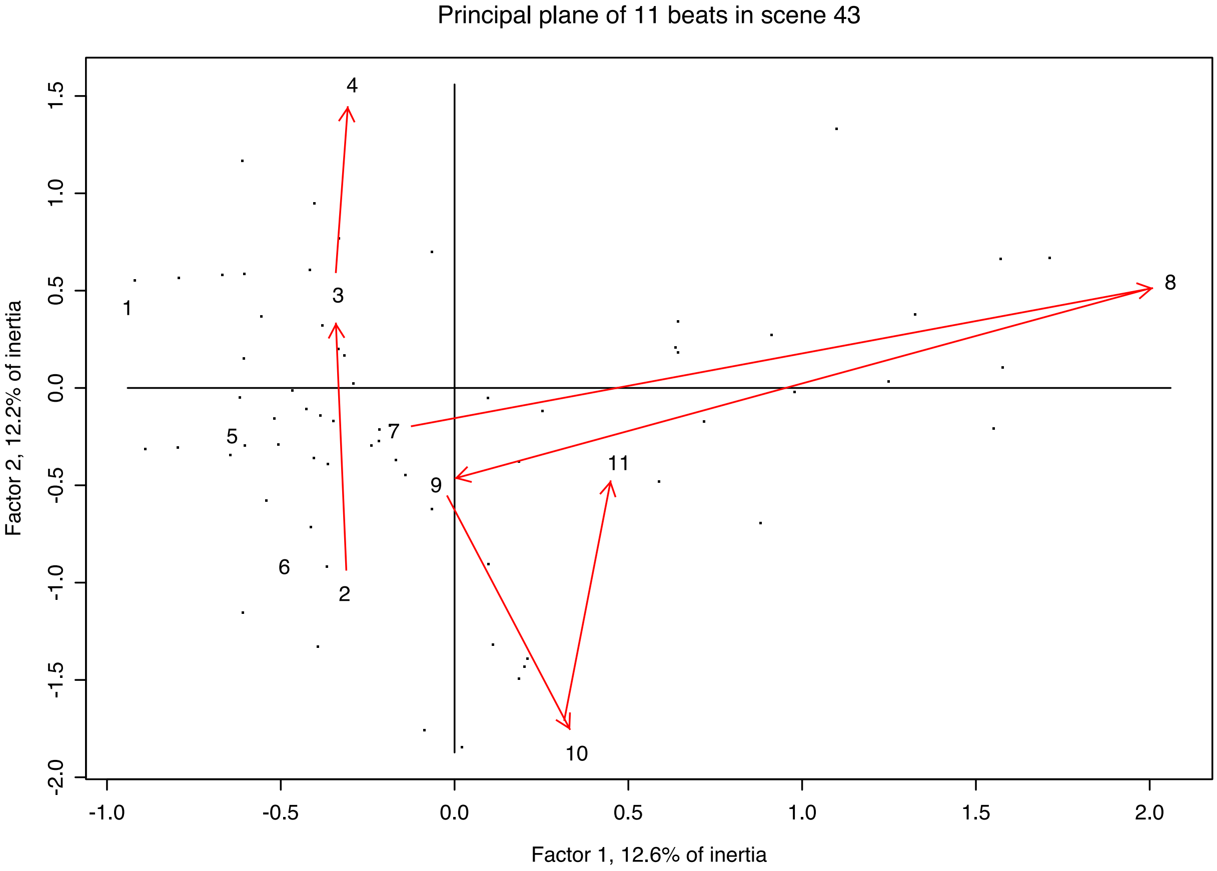


*Figure 4: Correspondence Analysis principal plane – best Euclidean embedding in two dimensions – of scene 43. This scene is a central and indeed a pivotal one in the movie Casablanca. It consists of eleven sub- scenes, which McKee terms “beats”. We discuss in the text the evolution over sub-scenes 2, 3 and 4; and again over sub-scenes 7, 8, 9, 10, and 11*

In the full-dimensional space we can check some other of McKee’s guidelines. Lengths of beat get shorter leading up to climax: word counts of final five beats in scene 43 are: 50, 44, 38, 30 and 46. A style analysis of scene 43 based on McKee can be Monte Carlo tested against 999 uniformly randomized sets of the beats. In the great majority of cases (against 83% and more of the randomized alternatives) we find the style in scene 43 to be characterized by: small variability of movement from one beat to the next; greater tempo of beats; and high mean rhythm.

The planar representation in Fig. 4 accounts for approximately 12.6% + 12.2% = 24.8% of the inertia, and hence the total information. We will look at the evolution of this scene, scene 43, using hierarchical clustering of the full-dimensional data – but based on the relative orientations, or correlations with factors. This is because of what we have found in Fig. 4, viz. change of direction is most important.

Fig. 5 shows the hierarchical clustering, based on the sequence of beats. Input data are of full dimensionality so there is no approximation involved. Note the caesura in moving from beat 7 to 8, and back to 9. There is less of a caesura in moving from 4 to 5 but it is still quite pronounced.

Further discussion of these results can be found in Murtagh et al. (2009).

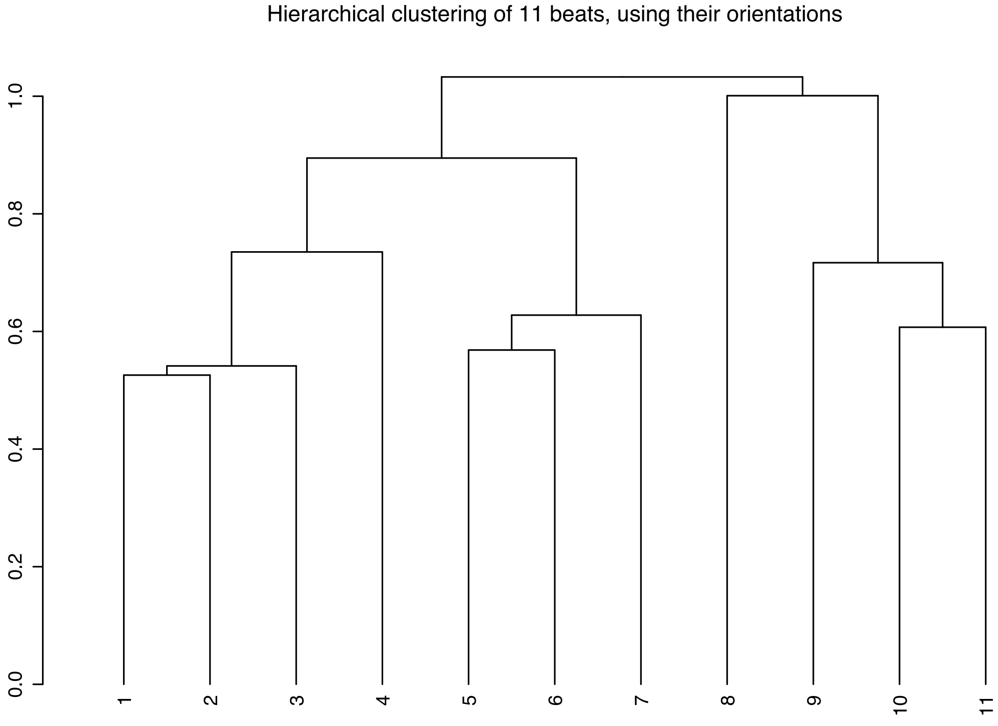


*Figure 5: Hierarchical clustering of sequence of beats in scene 43 of Casablanca. A sequence-constrained complete link agglomerative clustering algorithm is used. The input data is based on the full dimensionality Euclidean embedding provided by the Correspondence Analysis. The relative orientations (defined by correlations with the factors) are used as input data*

## 3. Television: CSI, Crime Scene Investigation

### *3.1. Data Used*

We took three CSI (Crime Scene Investigation, Las Vegas – Grissom, Sara, Catherine et al.) television scripts from series 1:

- 1X01, Pilot, original air date on CBS Oct. 6, 2000. Written by Anthony E. Zuiker, directed by Danny Cannon.
- 1X02, Cool Change, original air date on CBS, Oct. 13, 2000. Written by Anthony E. Zuiker, directed by Michael Watkins.

- 1X03, Crate ‘N Burial, original air date on CBS, Oct. 20, 2000. Written by Ann Donahue, directed by Danny Cannon.

Note the differences between writers and directors in most cases. We will refer to these scripts as CSI 101, CSI 102 and CSI 103. All film-scripts were obtained from TWIZ TV (www.twiztv.com).

We also took another three scripts from series 3, which we will not further discuss here. See Murtagh et al. (in press).

An example of a very short scene, scene 25 from CSI 101, follows.

```
[INT. CSI - EVIDENCE ROOM -- NIGHT]
(WARRICK opens the evidence package and takes out the shoe.)
(He sits down and examines the shoe. After several dissolves, WARRICK opens the lip
of the shoe and looks inside. He finds something.)
WARRICK BROWN: Well, I’ll be damned.
(He tips the shoe over and a piece of toe nail falls out onto the table. He picks
it up.)
WARRICK BROWN: Tripped over a rattle, my ass.
```

We see here scene metadata, characters, dialog and action information, all of which we use. Frontpiece, preliminary or preceding storyline information, and credits were ignored by us. All punctuation was ignored. All upper case was converted to lower case. There was no pruning of stopwords (e.g., “the”, “and”, etc.). In CSI 101 the top words and their frequencies of occurrence were:

the 443; to 239; grissom 195; you 176; and 166; gil 114; catherine 105; of 89; he 85; nick 80; in 79; on 79; it 78; at 76; ted 66; sara 65; warrick 65; ...

### 3.2. *Capturing and Displaying Text Sequence Semantics through Spatial Embedding: using CSI 101 Pilot*

In the Pilot, CSI 101, there are 50 scenes, with word counts ranging from 146 words to 676 words. In all there are 9.934 words. There are 1.679 unique words greater than 1 letter in length, with lower case replacing upper case, and with punctuation ignored. We will use this 1.679 unique word set.

The frequency of occurrence data crossing the 50 scenes and 1679 words is mapped, using Correspondence Analysis, into a space of intrinsic dimensionality 49: if $n$, $m$ are respectively the numbers of rows or scenes, and columns or words, then the inherent dimensionality is $\min(n - 1, m - 1)$; the reason why 1 is subtracted from both is that the cloud of scenes and the cloud of words are both centred, giving a linear dependence. The origin is the average, expressing the hypothetical scene, or the hypothetical word, carrying no information.

In Fig. 6, the scenes and words are located in the same embedding. The figure is interpreted in a visually natural, Euclidean, way, which is not the same as when we are presented with a frequency of occurrence data array. Defined on the basis of the frequency of occurrence array, we have the $\chi^2$ distance between scenes and/or words. The output display in Fig. 6 is a best planar view of a space endowed with the Euclidean metric. Both scenes and words have a “built-in” normalization, given by the $\chi^2$ metric.

One important fact to keep in mind is that this is a best planar view of what is, in reality, a 49-dimensional space. The quality of the approximation involved in this is seen in the percentage inertia explained by these factors. Inertia explained by a factor r is the sum over all scenes of:

mass times the projection squared on the axis. Quite typically for Correspondence Analysis, the extent of approximation is low in percentage terms. This is because less important factors or axes are "explained by", or determined by, isolated, very particular, words (which thereby also determine the information content of particular scenes).

The relationship between scenes and words in Fig. 6 is ultimately given by the dual space relationships (Murtagh, 2005): each scene is located at the centre of gravity of all words; and each word is located at the centre of gravity of all scenes. As noted above, this establishes a semantic property for the location of any scene – since the scene's location is determined by the word set. Similarly each word's location is determined by the scene set, hence establishing its semantics.

In practice, Fig. 6 presents a very useful view of relationships in our scenes x words data. We can look for polarities in the data; or anomalous scenes or words; or clusters or other configurations of scenes with reference to words or vice versa. But it is an approximation to the full dimensional reality. Therefore for some purposes we prefer to use the full dimensionality Euclidean representation furnished by the factor space. In such a case there is no low dimensional projection involved, and no loss of information.

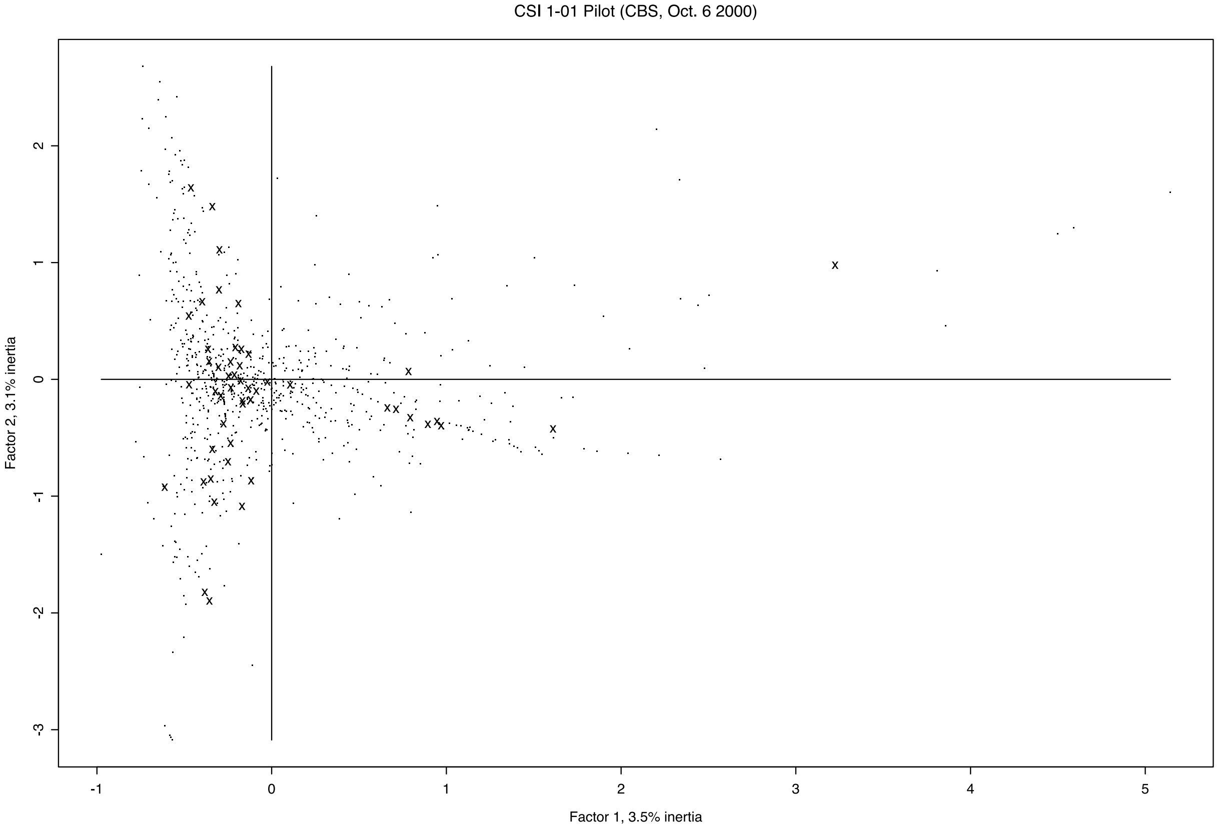


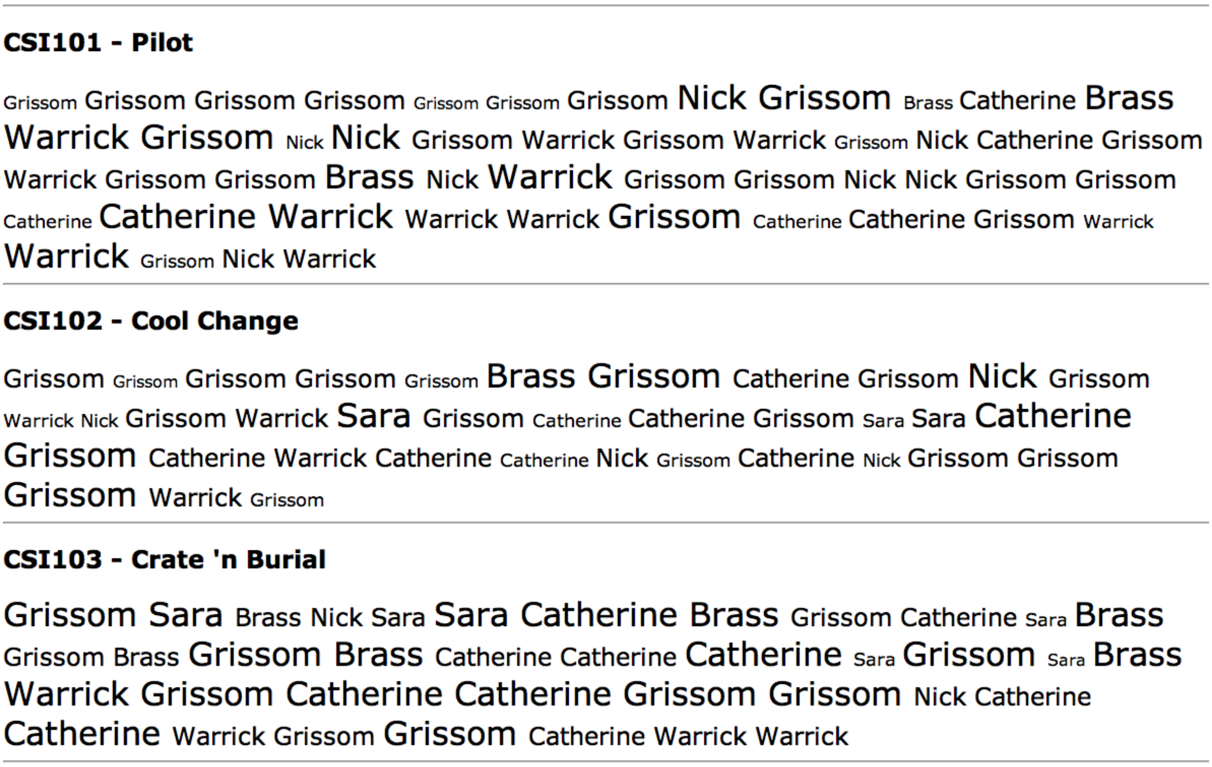
*Figure 6: Correspondence Analysis principal factor plane of projections of 50 scenes (each represented with an x), and 1.679 characterizing words (each represented with a dot). In this planar view of the two clouds, the cloud of scenes and the cloud of words, we eschew labels for clarity*

### 3.3. Tracking of Characters

In the sequence of scenes balance must be maintained as well as tempo-related contrast. In such areas as contrasts between interior and exterior scenes, day and night, and the presence or absence of principal and secondary characters, the filmscript must reflect vital aids and hints to the viewer, provoking both continuity of understanding by the viewer and discontinuity to trigger heightened attention.

We will look at the principal characters in the CSI scripts and television series programs: Gil Grissom, Warrick Brown, Nick Stokes, Catherine Willows, Jim Brass and Sara Sidle. We will refer to them by the first or family name mainly used: Grissom, Warrick, Nick, Catherine, Brass and Sara.

The Correspondence Analysis allows us to easily seek the principal character who is closest to each scene. In the plot of scenes crossed by all words used in the filmscript, which naturally contains the character names, we look for proximity – in the full dimensional Euclidean, factor space, so no approximation is involved – between the character and the scenes. The relative importance is expressed by size in Fig. 7. This relative importance is a scaled version of the log (base 10) of the squared Euclidean distance. (Using the distance or squared distance, and taking the log, there is clearly no effect on monotonicity of proximity. We take the log for improved visual appearance.)



**CSI101 - Pilot**

Grissom Grissom Grissom Grissom Grissom Grissom Grissom Nick Grissom Brass Catherine Brass Warrick Grissom Nick Nick Grissom Warrick Grissom Warrick Grissom Nick Catherine Grissom Warrick Grissom Grissom Brass Nick Warrick Grissom Grissom Nick Nick Grissom Grissom Catherine Catherine Warrick Warrick Warrick Grissom Catherine Catherine Grissom Warrick Warrick Grissom Nick Warrick

**CSI102 - Cool Change**

Grissom Grissom Grissom Grissom Grissom Brass Grissom Catherine Grissom Nick Grissom Warrick Nick Grissom Warrick Sara Grissom Catherine Catherine Grissom Sara Sara Catherine Grissom Catherine Warrick Catherine Catherine Nick Grissom Catherine Nick Grissom Grissom Grissom Warrick Grissom

**CSI103 - Crate 'n Burial**

Grissom Sara Brass Nick Sara Sara Catherine Brass Grissom Catherine Sara Brass Grissom Brass Grissom Brass Catherine Catherine Catherine Sara Grissom Sara Brass Warrick Grissom Catherine Catherine Grissom Grissom Nick Catherine Catherine Warrick Grissom Grissom Catherine Warrick Warrick

*Figure 7: In CSI 101, CSI 102 and CSI 103, there were respectively 50, 37 and 38 scenes. We show the most important character (personality), among the six principal characters, for each scene in succession. The size used in the display, expressing relative importance for the scene, is defined via proximity between scene and character name, as explained in the text*

We can see at a glance how Grissom pervades these films; whether characters reappear as the most crucial players implying intertwining of different actors; how the central roles of male and female characters alternate and so on. We could of course collect statistics of appearance and present such results as histograms or pie charts, or a time series. However, the motivation for our tag clouds is to have a range of properties of the filmscript presented simultaneously.

## 4. Conclusions

We have reported on how the structure of narrative, as expressed by a filmscript, can be analysed.

Apart from applications in the fast converging world of television, games, and Internet, we are also pursuing applications to the semantic analysis of general narrative. This includes blog data and we have also started to work on twitter data.

## References

Benzécri J.P. (1979). *L'Analyse des Données, Tome I Taxinomie, Tome II Correspondances*, 2nd ed. Paris: Dunod.

Burnett M. and Allison J. (1940). *Everybody Comes to Rick's*. Screenplay.

Le Roux B. and Rouanet H. (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Dordrecht: Kluwer.

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Paris: Dunod.

Lebart L., Salem A. and Berry L. (1998). *Exploring Textual Data*. Dordrecht: Kluwer.

McKee R. (1999). *Story: Substance, Structure, Style, and the Principles of Screenwriting*. London: Methuen.

Murtagh F. (1985). *Multidimensional Clustering Algorithms*. Würzburg: Physica-Verlag.

Murtagh F. (2004). On ultrametricity, data coding, and computation. *Journal of Classification*, **21**: 167-184.

Murtagh F. (2005). *Correspondence Analysis and Data Coding with R and Java*. Boca Raton (FL): Chapman & Hall/CRC.

Murtagh F., Ganz A. and McKie S. (2009). The structure of narrative: the case of film scripts. *Pattern Recognition*, **42**: 302-312.

Murtagh F., Ganz A., McKie S., Mothe J. and Englmeier K. (in press). Text sequence visualization using planar maps, hierarchical clustering and tag clouds: the case of film-scripts, *Information Visualization*.

van Rijsbergen C.J. (2004). *The Geometry of Information Retrieval*. Cambridge: Cambridge University Press.