# Discourse Type Clustering using POS *n*-gram Profiles and High-Dimensional Embeddings

**Christelle Cocco**
Department of Computer Science and Mathematical Methods
University of Lausanne
Switzerland
Christelle.Cocco@unil.ch

## Abstract

To cluster textual sequence types (discourse types/modes) in French texts, K-means algorithm with high-dimensional embeddings and fuzzy clustering algorithm were applied on clauses whose POS (part-of-speech) *n*-gram profiles were previously extracted. Uni-, bi- and trigrams were used on four 19th century French short stories by Maupassant. For high-dimensional embeddings, power transformations on the chi-squared distances between clauses were explored. Preliminary results show that high-dimensional embeddings improve the quality of clustering, contrasting the use of bi- and trigrams whose performance is disappointing, possibly because of feature space sparsity.

## 1 Introduction

The aim of this research is to cluster textual sequence types (named here discourse types)[1], such as narrative, descriptive, argumentative and so on in French texts, and especially in short stories which could contain all types.

For this purpose, texts were segmented into clauses (section 2.1). To cluster the latter, *n*-gram POS (part-of-speech) tag profiles were extracted (section 2.3). POS-tags were chosen because of their expected relation to discourse types.

Several authors have used POS-tags among other features for various text classification tasks, such as Biber (1988) for text type detection, Karlgren and Cutting (1994) and Malrieu and Rastier (2001) for genre classification, and Palmer et al. (2007) for situation entity classification. The latter is an essential component of English discourse modes (Smith, 2009). Moreover, previous work in discourse type detection has shown a dependency between POS-tags and these types (Cocco et al., 2011).

In this paper, K-means algorithm with high-dimensional embeddings and fuzzy clustering algorithm were applied on uni-, bi- and trigram POS-tag profiles (section 2.4) and results were evaluated (section 2.5). Finally, results are given in section 3.

## 2 Method

### 2.1 Expert assessment

The human expert, a graduate student in French linguistics, annotated 19th century French short stories by Maupassant, using XML tags. Each text was first segmented into clauses, whose length is typically shorter than sentences. Then, texts were annotated retaining the following six discourse types: narrative, argumentative, descriptive, explicative, dialogal and injunctive.[2] They resulted from an adaptation of the work of Adam (2008a; 2008b) in text and discourse analysis, as well as Bronckart (1996) in psycholinguistics, concerning textual sequence types. The former does not consider the injunctive type.

Let us briefly describe these types (Adam, 2008a; Adam, 2008b; Bronckart, 1996), together with the criteria finally adopted by the human expert for this time-consuming task.

[1] **Sequence** type is an appropriate name, because it refers to text **passage** type. However, it will be further mentioned as discourse types, a frequent French term. In English, a standard term is: discourse modes.

[2] Regarding English, there are five discourse modes according to Smith (2009): narrative, description, report, information and argument.



*Proceedings of the EACL 2012 Student Research Workshop*, pages 55–63,
Avignon, France, 26 April 2012. 

**Narrative** type corresponds to told narrative. One of the principal linguistic markers of this type is the presence of past historic tense. However, when referring to repeated actions, imperfect tense is generally used. **Argumentative** type corresponds to texts whose aim is to convince somebody of an argument. An important linguistic marker of this type is the presence of argumentative connectors such as *mais* "but", *cependant* "however", *pourtant* "yet" and so on. **Explicative** type aims to explain something unknown, such as encyclopaedic knowledge, and answers to the question "Why?". A typical linguistic marker of this type is the presence of phraseological phrases, such as *(si)...c'est parce que/c'est pour que* "(if)...it is because/in order to". **Descriptive** type represents textual parts where the time of the story stops and where characteristic properties of a subject, animated or not, are attributed. Several linguistic markers are relevant for this type: use of imperfect tense (except when the narrative part is in present tense); a large number of adjectives; spatio-temporal organizers; and stative verbs. **Dialogal** type is a verbal exchange. However, in this project, direct speech is considered as dialogal too. Typical linguistic markers of this type are quotes, strong punctuation and change of spatio-temporal frame. Finally, **injunctive** type is an incentive for action. This type has linguistic markers such as use of imperative tense and exclamation marks. In our corpus, this type is always included in a dialogal segment.

Discourse types are generally nested inside each other resulting in a hierarchical structure. For instance, an injunctive sequence of one clause length can be included in a dialogal sequence, which can in turn be included in a longer narrative sequence matching the entire text. In the simplified treatment attempted here, the problem is linearized: only the leaves of the hierarchical structure will be considered.

### 2.2 Corpus

The corpus consists of four 19$^{\text{th}}$ century French short stories by Maupassant: "L'Orient" , "Le Voleur", "Un Fou?" and "Un Fou". Descriptive statistics about these texts are given in table 1. These values are based on unigram counts. For bigram and trigram counts, clauses shorter than two and three words respectively were removed. For the first text, "L'Orient", three clauses were removed for trigrams; for "Le Voleur", one clause was removed for trigrams; and for "Un Fou?", thirteen clauses for trigrams. An extra step was made for "Un Fou", because of its very different structure w.r.t. the three other texts. Indeed, the majority of this text is written as a diary. Dates, which could not be attributed to a discourse type, were consequently removed, reducing the number of clauses from 401 to 376 for unigrams. Then, two clauses were removed for bigrams because they were too short, and again ten for trigrams.

### 2.3 Preprocessing

Before applying clustering algorithms, annotated texts were preprocessed to obtain a suitable contingency table, and dissimilarities between clauses were computed. Firstly, each text was POS-tagged with TreeTagger (Schmid, 1994) excluding XML tags. Secondly, using the manual clause segmentation made by the human expert, distributions over POS-tag *n*-grams were obtained for each clause, resulting in a contingency table.

Then, chi-squared distances between clauses were computed. In order to accomplish this, coordinates of the contingency table (with $n_{ik}$ denoting the number of objects common to clause $i$ and POS-tag *n*-gram $k$, $n_{i\bullet} = \sum_k n_{ik}$ and $n_{\bullet k} = \sum_i n_{ik}$) are transformed in this manner:

$$y_{ik} = \frac{e_{ik}}{f_i \sqrt{\rho_k}} - \sqrt{\rho_k} \tag{1}$$

where $e_{ik} = n_{ik}/n$ are the relative counts, $f_i = e_{i\bullet} = n_{i\bullet}/n$ (row weights) and $\rho_k = e_{\bullet k} = n_{\bullet}k/n$ (column weights) are the margin counts. Finally, the squared Euclidean distances between these new coordinates

$$D_{ij} = \sum_k (y_{ik} - y_{jk})^2 \tag{2}$$

define the chi-squared distances.

### 2.4 Algorithms

Two algorithms were applied on these distances.

**K-means with high-dimensional embedding**

Firstly, the well-known K-means (see *e.g.* Manning and Schütze (1999)) was performed in a weighted version (*i.e.* longer clauses are more important than shorter ones), by iterating the following pair of equations:

$$z_i^g = \begin{cases} 1 & \text{if} \quad g = \underset{h}{\operatorname{argmin}}\, D_i^h \\ 0 & \text{else.} \end{cases} \tag{3}$$

| Texts | ♯ sent. | ♯ clauses | ♯ tokens | | ♯ types | | % discourse types according to the expert | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | with punct. | w/o punct. | word | tag | arg | descr | dial | expl | inj | nar |
| L'Orient | 88 | 189 | 1'749 | 1'488 | 654 | 27 | 4.23 | 20.11 | 25.93 | 19.05 | 2.65 | 28.04 |
| Le Voleur | 102 | 208 | 1'918 | 1'582 | 667 | 29 | 4.81 | 12.02 | 13.94 | 4.81 | 2.88 | 61.54 |
| Un Fou? | 150 | 314 | 2'625 | 2'185 | 764 | 28 | 18.15 | 10.51 | 14.65 | 14.65 | 8.28 | 33.76 |
| Un Fou | 242 | 376 | 3'065 | 2'548 | 828 | 29 | 17.82 | 13.83 | 1.86 | 11.70 | 12.23 | 42.55 |

Table 1: Statistics of the annotated texts by Maupassant. For the text "Un Fou", dates were initially removed from the text. Number of sentences as considered by TreeTagger (Schmid, 1994). Number of clauses as segmented by the human expert. Number of tokens including punctuation and compounds as tagged by TreeTagger. Number of tokens without punctuation and numbers, considering compounds as separated tokens. Number of wordform types. Number of POS-tag types. The last columns give the percentage of clauses for each discourse type (arg = argumentative, descr = descriptive, dial = dialogal, expl = explicative, inj = injunctive, nar = narrative).

$$D_i^g = \sum_j f_j^g D_{ij} - \Delta_g \tag{4}$$

where $z_i^g$ is the membership of clause $i$ in group $g$ and $D_i^g$ is the chi-squared distance between the clause $i$ and the group $g$ as resulting from the *Huygens principle*. In the equation 4, $f_j^g = (f_i z_{ig})/\rho_g = p(i|g)$, $D_{ij}$ is the chi-squared distances between clauses given by the equation 2 and $\Delta_g = 1/2 \sum_{jk} f_j^g f_k^g D_{jk}$ is the inertia of group $g$. In addition, $\rho_g = \sum_i f_i z_{ig} = p(g)$ is the relative weight of group $g$.

At the outset, the membership matrix $Z$ was chosen randomly, and then the iterations were computed until stabilisation of the matrix $Z$ or a number of maximum iterations $N_{\max}$.

Besides the K-means algorithm, Schoenberg transformations $\varphi(D)$ were also operated. They transform the original squared Euclidean distances $D$ into new squared Euclidean distances $\varphi(D)$ (Bavaud, 2011) and perform a high-dimensional embedding of data, similar to those used in Machine Learning. Among all Schoenberg transformations, the simple componentwise power transformation was used, *i.e.*

$$\varphi(D_{ij}) = (D_{ij})^q \tag{5}$$

where $0 < q \leq 1$.

In a nutshell, the K-means algorithm was applied on the four texts, for uni-, bi- and trigrams POS-tags, with $q$ in equation 5 varying from 0.1 to 1 with steps of 0.05. Given that the aim was to find the six groups annotated by the human expert, the K-means algorithm was computed with a number of groups $m = 6$. Moreover, $N_{\max} = 400$ and for each $q$, calculations were run 300 times, and then the averages of the relevant quantities (see section 2.5) were computed.

**Fuzzy clustering**

Secondly, the same algorithm which was used in a previous work (Cocco et al., 2011) was applied here, *i.e.* the fuzzy clustering algorithm.

In brief, it consists of iterating, as for the K-means, the membership $z_i^g$ of clause $i$ in group $g$ defined in the following way (Rose et al., 1990; Bavaud, 2009):

$$z_i^g = \frac{\rho_g \exp(-\beta D_i^g)}{\sum_{h=1}^{m} \rho_h \exp(-\beta D_i^h)} \tag{6}$$

until stabilisation of the membership matrix $Z$ (randomly chosen at the beginning as uniformly distributed over the $m$ groups) or after $N_{\max}$ iterations. $D_i^g$ is given by equation 4 and $\rho_g$ is the relative weight of group $g$. Moreover, it turns out convenient to set $\beta := 1/(t_{\text{rel}} \times \Delta)$, the "inverse temperature" parameter, where $\Delta := \frac{1}{2} \sum_{ij} f_i f_j D_{ij}$ is the inertia and $t_{\text{rel}}$ is the relative temperature which must be fixed in advance.

The values of $\beta$ controls for the bandwidth of the clustering, *i.e.* the number of groups: the higher $\beta$, the larger the final number of groups $M$ (see figure 9). As a matter of fact, depending of $\beta$ values, group profiles are more or less similar. Also, group whose profiles are similar enough are aggregated, reducing the number of groups from $m$ (initial number of groups chosen at the beginning) to $M$. This aggregation is made by adding memberships of clauses: $z_i^{[g \cup h]} = z_i^g + z_i^h$. Two groups are considered similar enough if $\theta_{gh}/\sqrt{\theta_{gg}\theta_{hh}} \geq 1 - 10^{-5}$, with $\theta_{gh} = \sum_{i=1}^{n} f_i z_i^g z_i^h$ which measures the overlap between $g$ and $h$ (Bavaud, 2010). Finally, each clause is attributed to the most probable group.

For the application in this project, fuzzy clustering algorithm was computed on the four texts,

for uni- bi- and trigrams POS-tags. At the outset, the initial number of groups $m$ was equal to the number of clauses for each text (see table 1 and section 2.2), with a relative temperature $t_{\text{rel}}$ from 0.022 to 0.3 with steps of 0.001 (except for the text "Un Fou" with $t_{\text{rel min}} = 0.02$, $t_{\text{rel max}} = 0.3$ and $t_{\text{rel step}} = 0.01$). Besides this, $N_{\text{max}} = 400$ and for each $t_{\text{rel}}$, algorithm was run 20 times, and finally the averages of the relevant quantities (see section 2.5) were computed.

### 2.5 Evaluation criteria

The clustering obtained by the two algorithms (K-means with high-dimensional embedding and fuzzy clustering) were compared to the classification made by the human expert. As clustering induces anonymous partitions, traditional indices such as precision, recall and Cohen's Kappa cannot be computed.

Among the numerous similarity indices between partitions, we have examined the *Jaccard index* (Denœud and Guénoche, 2006; Youness and Saporta, 2004):

$$J = \frac{r}{r+u+v} \tag{7}$$

whose values vary between 0 and 1, and the *corrected Rand index* (Hubert and Arabie, 1985; Denœud and Guénoche, 2006):

$$RC = \frac{r - \text{Exp}(r)}{\text{Max}(r) - \text{Exp}(r)} \tag{8}$$

whose the maximal value is 1. When this index equals 0, it means that similarities between partitions stem from chance. However, it can also take negative values when number of similarities is lower than the expectation (*i.e.* chance).

Both indices are based upon the contingency table $n_{ij}$, defined by the number of objects attributed simultaneously to group $i$ (w.r.t. the first partition) and to group $j$ (w.r.t. the second partition). Moreover, in both indices, $r = \frac{1}{2}\sum_{ij} n_{ij}(n_{ij}-1)$ is the number of pairs simultaneously joined together, $u = \frac{1}{2}(\sum_j n_{\bullet j}^2 - \sum_{ij} n_{ij}^2)$ (respectively $v = \frac{1}{2}(\sum_i n_{i\bullet}^2 - \sum_{ij} n_{ij}^2)$) is the number of pairs joined (respectively separated) in the partition obtained with algorithm and separated (respectively joined) in the partition made by the human expert, $\text{Exp}(r) = \frac{1}{2n(n-1)}\sum_i n_{i\bullet}(n_{i\bullet}-1)\sum_j n_{\bullet j}(n_{\bullet j}-1)$ is the expected number of pairs simultaneously joined together by chance and $\text{Max}(r) = \frac{1}{4}\sum_i n_{i\bullet}(n_{i\bullet}-1) + \sum_j n_{\bullet j}(n_{\bullet j}-1)$.

## 3 Results

On the one hand, results obtained with the K-means algorithm and power ($q$) transformations for uni-, bi- and trigrams are presented in figures 1 to 8. On the other hand, results obtained with fuzzy clustering for uni- bi- and trigrams are only shown for the text "Le Voleur" in figures 9 to 13. For the three other texts, results will be discussed below.

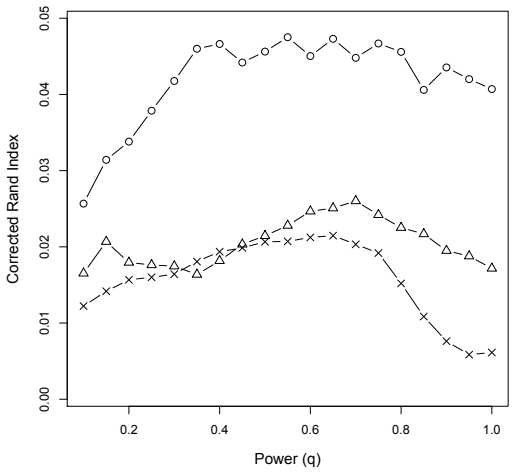


Figure 1: "L'Orient" with K-means algorithm: corrected rand index as a function of power (q) (○ = unigrams, △ = bigrams and × = trigrams). The standard deviation is approximatively constant across $q$ ranging from a minimum of 0.018 and a maximum of 0.024 (unigrams); 0.0099 and 0.015 (bigrams); 0.0077 and 0.013 (trigrams).

A first remark is that corrected Rand index and Jaccard index behave differently in general. This difference is a consequence of the fact that Jaccard index does not take into account the number of pairs simultaneously separated in the two partitions, a fact criticised by Milligan and Cooper (1986).

Regarding the texts "L'Orient", "Le Voleur" and "Un Fou?" with K-means algorithm and the corrected Rand index (figures 1, 3 and 5), unigrams give the best results. Moreover, power transformations (equation 5) tend to improve them. For instance, for the text "L'Orient" (figure 1), the best result is $RC = 0.048$ with $q = 0.55$, and for the text "Un Fou?" (figure 5), the best

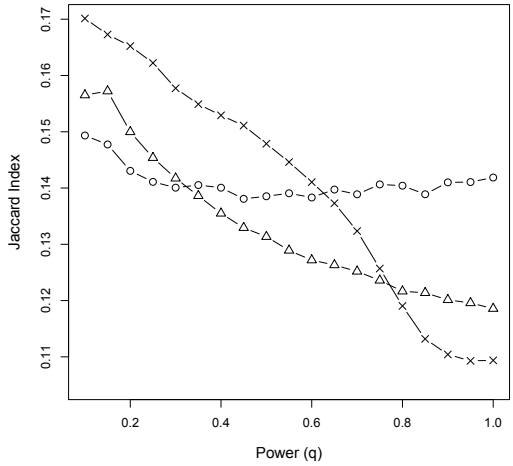


Figure 2: "L'Orient" with K-means algorithm: Jaccard index as a function of power (q) ($\circ$ = unigrams, $\triangle$ = bigrams and $\times$ = trigrams).

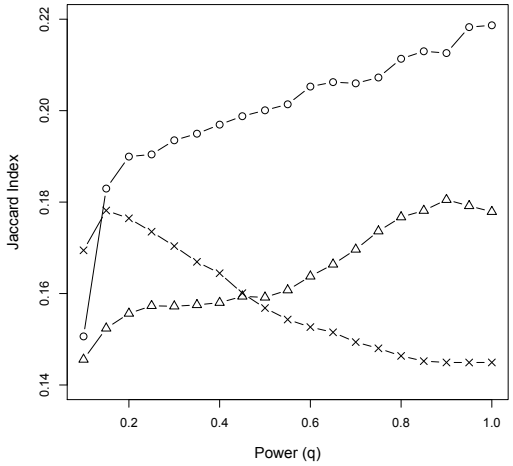


Figure 4: "Le Voleur" with K-means algorithm: Jaccard index as a function of power (q) ($\circ$ = unigrams, $\triangle$ = bigrams and $\times$ = trigrams).

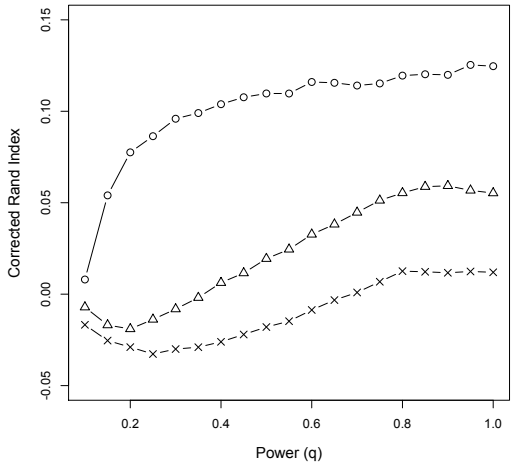


Figure 3: "Le Voleur" with K-means algorithm: corrected rand index as a function of power (q) ($\circ$ = unigrams, $\triangle$ = bigrams and $\times$ = trigrams).

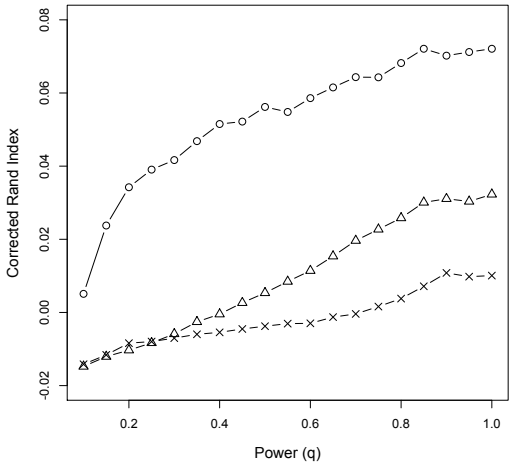


Figure 5: "Un Fou?" with K-means algorithm: corrected rand index as a function of power (q) ($\circ$ = unigrams, $\triangle$ = bigrams and $\times$ = trigrams).

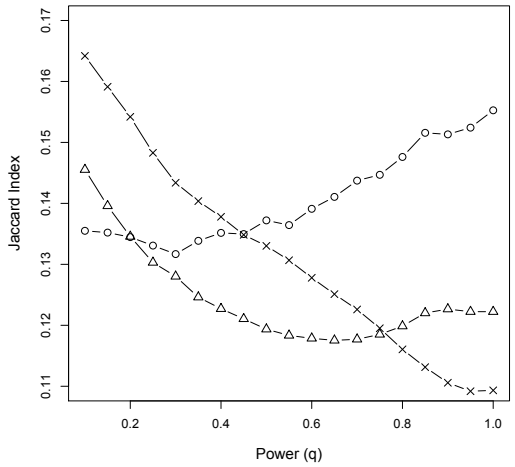


Figure 6: "Un Fou?" with K-means algorithm: Jaccard index as a function of power (q) (○ = unigrams, △ = bigrams and × = trigrams).

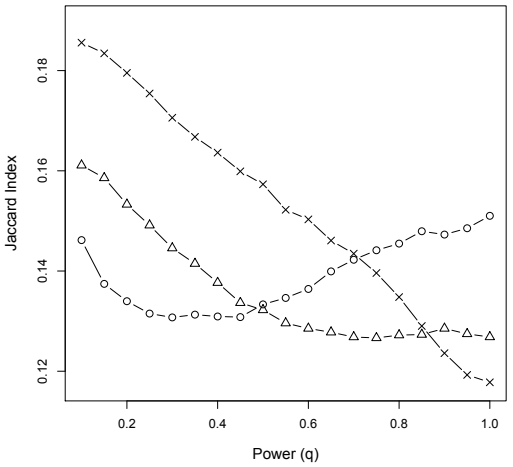


Figure 8: "Un Fou" with K-means algorithm: Jaccard index as a function of power (q) (○ = unigrams, △ = bigrams and × = trigrams).

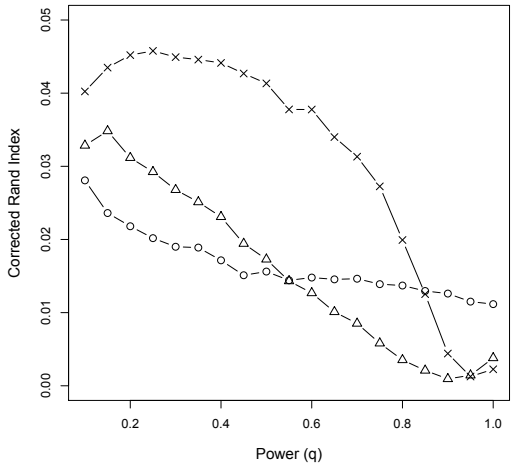


Figure 7: "Un Fou" with K-means algorithm: corrected rand index as a function of power (q) (○ = unigrams, △ = bigrams and × = trigrams).

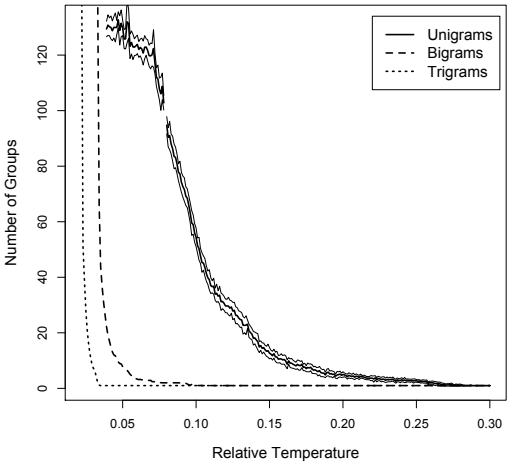


Figure 9: "Le Voleur" with fuzzy clustering algorithm: average number of groups as a function of the relative temperature. For unigrams, the thick line indicates the average and the two thin lines represent the standard deviation. The other curves depict the average of the number of groups.

result is $RC = 0.072$ with $q = 0.85$.

Regarding the fuzzy clustering algorithm, figure 9 shows, for the text "Le Voleur", the relation between the relative temperature and the number of groups for uni- bi- and trigrams, *i.e.* number of groups decreases when relative temperature increases. Figure 10 (respectively figure 12) presents the corrected Rand index (respectively the Jaccard index) as a function of relative temperature, while figure 11 (respectively figure 13) shows, for each relative temperature, the average number of groups on the x-axis and the average corrected Rand index (respectively Jaccard index) on the y-axis, over 20 clusterings. There is a remarkable peak for this text ($RC = 0.31$ (respectively $J = 0.48$)), when $t_{\text{rel}} = 0.145$ (respectively 0.148), corresponding to $M = 14.4$ (respectively 13.4). The same phenomenon appears with the text "Un Fou?", when $t_{\text{rel}} = 0.158$ and $M = 7.8$. However, the peak for the Jaccard index is less important and it is not the highest value. More-

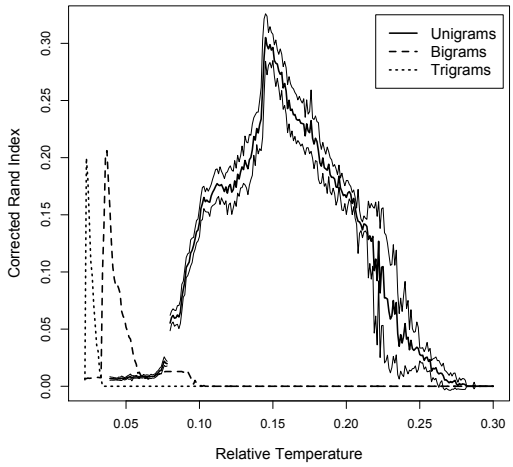


Figure 10: "Le Voleur" with fuzzy clustering algorithm: corrected Rand index as a function of relative temperature.

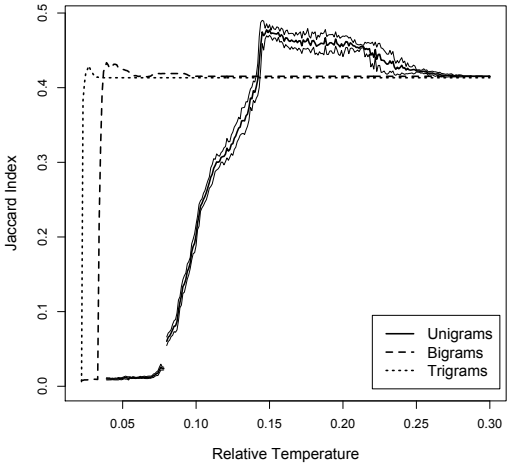


Figure 12: "Le Voleur" with fuzzy clustering algorithm: Jacccard index as a function of relative temperature.

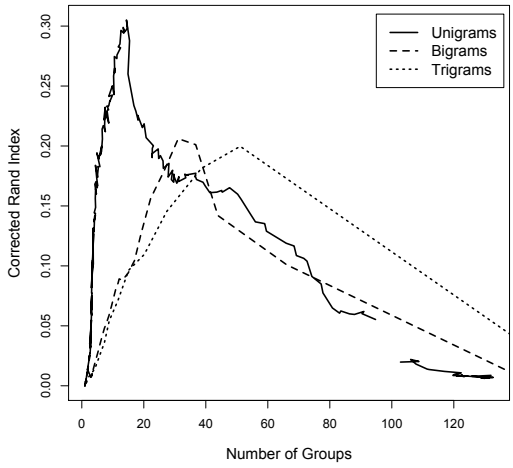


Figure 11: "Le Voleur" with fuzzy clustering algorithm: corrected Rand index as a function of number of groups.

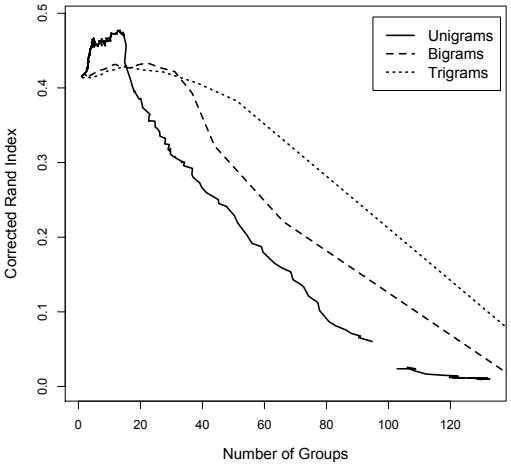


Figure 13: "Le Voleur" with fuzzy clustering algorithm: Jaccard index as a function of number of groups.

over, for the latter text, there is a higher peak, which occurs only with the corrected Rand index, for $t_{\text{rel}} = 0.126$ and $M = 24.5$.

For the two other texts, there are some peaks, but not as marked as in other texts. Besides, for these two texts, corrected Rand index takes negative values, especially for "Un Fou". While the reason for these different behaviours is not known, it should be noted that the structure of these texts is different from that of the two other texts. Indeed, "Un Fou" is written as a diary and uses mainly the present tense, also in narrative and descriptive parts; "L'Orient" contains several long monologues mainly using the present tense too.

On figure 12, it appears that Jaccard index is constant when one group remains, and the same phenomenon appears for all texts. Indeed, from the distribution of table 2, one finds from equation 7: $r = 8\,939$, $u = 0$ and $v = 12\,589$, implying $J = 0.415$.

Overall, it is clear that results differ depending on texts, no matter which algorithm or evaluation criterion is used. Furthermore, they are always better for "Le Voleur" than for the three

| arg | descr | dial | expl | inj | nar |
|---|---|---|---|---|---|
| 10 | 25 | 29 | 10 | 6 | 128 |

Table 2: Types distribution for the text "Le Voleur".

other texts.

Finally, in most case, unigrams give better results than bi- and tri-grams. The relatively disappointing performance of bi- and trigrams (w.r.t. unigrams) could be accounted for by the sparsity of the feature space and the well-known associated "curse of dimensionality", in particular in clustering (see *e.g.* Houle et al. (2010)). Results are clearly different for "Un Fou", and the reason of this difference still needs to be investigated.

Certainly, as the sample is small and there is a unique annotator, all these results must be considered with caution.

## 4 Conclusion and further development

A first conclusion is that the use of POS-tag *n*-grams does not seem to improve the solution of the problem exposed here. In contrast, high-dimensional embedding seems to improve results. Concerning evaluation criteria, results clearly vary according to the selected index, which makes it difficult to compare methods. Another point is that even choosing only short stories of one author, text structures can be very different and certainly do not give the same results.

These results are interesting and in general better than those found in a previous work (Cocco et al., 2011), but this is still work in progress, with much room for improvement. A next step would be to combine fuzzy clustering with high-dimensional embedding, which can both improve results. Moreover, it could be interesting to add typical linguistic markers, such as those mentioned in section 2.1, or stylistic features. It would also be possible to use lemmas instead of or with POS-tags, if more data could be added to the corpus. Besides, *Cordial Analyseur* [3] could be used instead of TreeTagger, because it provides more fine-grained POS-tags. However, as for *n*-grams, it could imply a sparsity of the feature space. Another idea would be to perform a supervised classification with cross-validation. In this case, it would be interesting to investigate feature selection (see *e.g.* Yang and Pedersen (1997)). Also, the hierarchical structure of texts (cf. section 2.1) should be explored. Only the leaves were considered here, but in reality, one clause belongs to several types depending on the hierarchical level examined. Therefore, it could be relevant to consider the dominant discourse type instead of the leaf discourse type. Similarly, since in our corpus, injunctive type is always included in dialogal type, the former could be removed to obtain a larger dialogal class. In addition, it would be useful to find a better adapted measure of similarity between partitions. Finally, an important improvement would be to obtain more annotated texts, which should improve results, and a second human expert, which would permit us to assess the difficulty of the task.

[3] `http://www.synapse-fr.com/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm`


## Acknowledgments

I would like to thank François Bavaud and Aris Xanthos for helpful comments and useful discussions; Guillaume Guex for his help with technical matters; and Raphaël Pittier for annotating the gold standard.


## References


Jean-Michel Adam. 2008a. *La linguistique textuelle: Introduction à l'analyse textuelle des discours*. Armand Colin, Paris, 2nd edition.

Jean-Michel Adam. 2008b. *Les textes: types et prototypes*. Armand Colin, Paris, 2nd edition.

François Bavaud. 2009. Aggregation invariance in general clustering approaches. *Advances in Data Analysis and Classification*, 3(3):205–225.

François Bavaud. 2010. Euclidean distances, soft and spectral clustering on weighted graphs. In José Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6321 of *Lecture Notes in Computer Science*, pages 103–118. Springer, Berlin; Heidelberg.

François Bavaud. 2011. On the Schoenberg transformations in data analysis: Theory and illustrations. *Journal of Classification*, 28(3):297–314.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Jean-Paul Bronckart. 1996. *Activité langagière, textes et discours: Pour un interactionisme socio-discursif*. Delachaux et Niestlé, Lausanne; Paris.

Christelle Cocco, Raphaël Pittier, François Bavaud, and Aris Xanthos. 2011. Segmentation and clustering of textual sequences: a typological approach.

In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 427–433, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Guy de Maupassant. 1882. Le voleur. *Gil Blas*, June 21. `http://un2sg4.unige.ch/athena/selva/maupassant/textes/voleur.html`. Thierry Selva. Accessed 2011, July 6.

Guy de Maupassant. 1883. L'orient. *Le Gaulois*, September 13. `http://un2sg4.unige.ch/athena/selva/maupassant/textes/orient.html`. Thierry Selva. Accessed 2011, March 5.

Guy de Maupassant. 1884. Un fou?. *Le Figaro*, September 1. `http://un2sg4.unige.ch/athena/maupassant/maup_fou.html`. Thierry Selva. Accessed 2011, February 7.

Guy de Maupassant. 1885. Un fou. *Le Gaulois*, September 2. `http://un2sg4.unige.ch/athena/selva/maupassant/textes/unfou.html`. Thierry Selva. Accessed 2011, April 26.

Lucile Denœud and Alain Guénoche. 2006. Comparison of distance indices between partitions. In Vladimir Batagelj, Hans-Hermann Bock, Anuška Ferligoj, and Aleš Žiberna, editors, *Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 21–28. Springer Berlin Heidelberg.

Michael Houle, Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. 2010. Can shared-neighbor distances defeat the curse of dimensionality? In Michael Gertz and Bertram Ludäscher, editors, *Scientific and Statistical Database Management*, volume 6187 of *Lecture Notes in Computer Science*, pages 482–500. Springer, Berlin; Heidelberg.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th conference on Computational linguistics*, volume 2 of *COLING '94*, pages 1071–1075, Stroudsburg, PA, USA. Association for Computational Linguistics.

Denise Malrieu and Francois Rastier. 2001. Genres et variations morphosyntaxiques. *Traitement automatique des langues*, 42(2):547–577.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1st edition, June.

Glenn W. Milligan and Martha C. Cooper. 1986. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21(4):441–458.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. 2007. A sequencing model for situation entity classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 896–903, Prague, Czech Republic, June.

Kenneth Rose, Eitan Gurewitz, and Geoffrey C. Fox. 1990. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Carlota S. Smith. 2009. *Modes of Discourse: The Local Structure of Texts*. Number 103 in Cambridge studies in linguistics. Cambridge University Press, Cambridge, UK, digitally printed version edition.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420.

Genane Youness and Gilbert Saporta. 2004. Une Méthodologie pour la Comparaison de Partitions. *Revue de Statistique Appliquée*, 52(1):97–120.