Segmentation and Clustering of Textual Sequences: a Typological Approach

Christelle Cocco, Raphaël Pittier, François Bavaud and Aris Xanthos
University of Lausanne, Switzerland
{Christelle.Cocco,Raphael.Pittier,
Francois.Bavaud,Aris.Xanthos}@unil.ch

Abstract

The long term goal of this research is to develop a program able to produce an automatic segmentation and categorization of textual sequences into discourse types. In this preliminary contribution, we present the construction of an algorithm which takes a segmented text as input and attempts to produce a categorization of sequences, such as narrative, argumentative, descriptive and so on. Also, this work aims at investigating a possible convergence between the typological approach developed in particular in the field of text and discourse analysis in French by Adam (2008) and Bronckart (1997) and unsupervised statistical learning.

1 Introduction

An increasing amount of research has been conducted concerning text genre detection using POS (part-of-speech) tags since the work of Biber (1988). For instance, Malrieu and Rastier (2001) describe how to classify texts according to genres (comedy, tragedy, drama...) or discourses (literary, legal, political...) using POS-tags.

POS-tags can be determined in an unsupervised way (see *e.g.* Schmid (1994)) and their distribution happens to differ according to types of texts, such as narrative, explicative and so on. Hence, developing automatic discourse type detection, which is of interest to the linguistic community, seems practicable.

Thus, the purpose of the present study is to cluster clauses of a text into discourse types, *i.e.* to develop a tool for type detection with a limited quantity of annotated texts. We limit ourselves to the use of simple bag-of-words models on which fuzzy and K-means clustering are applied.

Specifically, the aim is twofold: firstly, the construction of a program which takes a segmented text as input and produces a categorization of sequences of clauses by clustering, based principally on POS-tags; secondly, the comparison of this clustering with the typology proposed by a human expert, corresponding to discourse types. Thus, this preliminary work aims at investigating a possible convergence between unsupervised statistical learning on the one hand, and the typological approach developed in particular in the field of French linguistics by Adam (2008) and in language psychology by Bronckart (1997) on the other hand.

As a first step, sample texts were manually annotated, that is segmented (section 2.1) and classified (section 2.2). Then, the clauses resulting from the previous segmentation were clustered on the basis of their POS distribution (sections 2.3 and 2.4). It appeared that the latter vary across the typological classes proposed by the expert (section 3.1) which were compared to those resulting from fuzzy and K-means clustering processes (sections 3.2 and 4). Future developments are proposed in section 5.

2 Method

2.1 Segmentation

The first step of this research was to create a corpus of annotated texts. For that purpose, a human expert has been working on 19th century French short stories by Maupassant. Only one genre is examined, because distributions of POS-tags vary with genre as mentioned in the introduction. For the same reason, only one author is considered (see *e.g.* Koppel and Schler (2003)) in this preliminary work. Annotation was carried out by means of XML tags, which is becoming a standard practice in this field (see *e.g.* Daoust et al. (2010)).

It transpired that segmentation into sentences

Proceedings of Recent Advances in Natural Language Processing, pages 427–433,
Hissar, Bulgaria, 12-14 September 2011.

Texts	♯ sentences	♯ clauses	♯ tokens		♯ types		% discourse types according to the human expert					
			with punct.	without punct.	wordforms	tags	nar	dial	descr	expl	arg	inj
”Un Fou?”	150	316	2’635	2’185	764	28	33.54	14.56	10.44	14.56	18.67	8.23
”L’Orient”	88	189	1’750	1’488	654	27	28.04	25.93	20.11	19.05	4.23	2.65
”Un Fou”	266	400	3’140	2’574	837	29	44.75	1.75	13.25	11.75	17.00	11.50
Total	504	905	7’525	6’247	2’255	30	37.35	11.27	13.70	14.25	14.92	8.51

Table 1: Statistics of the three annotated texts by Maupassant. 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 simple tokens without punctuation and figures, considering compounds as separated tokens. Number of wordform types. Number of POS-tag types. Percentage of clauses for each discourse type (nar=narrative, dial=dialogal, descr=descriptive, expl=explicative, arg=argumentative, inj=injunctive).

was not sufficiently fine-grained for the envisioned analysis, so the expert was instructed to segment the texts at the clause level.

2.2 Classification by a human expert

To be able to compare the results of the automatic clustering with a classification according to the typological approach developed in particular by Adam (2008; 2005) and Bronckart (1997), the expert was then asked to classify the clauses into six types. In fact, Adam proposes a classification of textual sequences into five types: narrative, argumentative, descriptive, explicative and dialoged sequences. However, we decided to add an injunctive type, following Bronckart. The expert decision to classify clauses was based partly upon formal criteria, such as punctuation, typical words, tense of verbs and semantics; and partly upon his linguistic and literary knowledge. Table 1 shows descriptive statistics about annotated texts.

An important issue inherent in this task is that the typological structure of the text is hierarchical rather than linear. This means that a sequence of a given type may contain sequences of other types. The number of inclusions is not limited. For the purpose of annotation, the use of XML tags appears to be appropriate, since it allows us to describe trees. However, taking into account the full hierarchical structure represents an additional difficulty for the automatic clustering procedure; in this first approach, the problem is treated as linear, *i.e.* only the leaves of the tree structure are considered (for the clauses). For instance, in the extract given in table 2, the first three clauses are regarded as narrative; the forth as injunctive; the fifth as argumentative; and the others as explicative.

```
<div type=”narratif”>
<e>Je le trouvai tantôt couché sur un divan,
 en plein rêve d’opium.</e>
<e>Il me tendit la main sans remuer le corps,</e>
<e>et me dit :</e><cr/>
  <div type=”dialogal”>
    <div type=”injonctif”>
    <e>Reste là, parle,</e>
    </div>
    <div type=”argumentatif”>
    <e>je te répondrai de temps en temps,</e>
      <div type=”explicatif”>
      <e>mais je ne bougerai point,</e>
      <e>car tu sais qu’une fois la drogue avalée</e>
      <e>il faut demeurer sur le dos.</e><cr/>
      </div>
    </div>
  </div>
</div>
```

Table 2: Annotated extract of ”L’Orient” by Maupassant. <e> refers to clause.

2.3 Automatic fuzzy clustering

The general principle is to perform a maximally unsupervised classification (clustering) to be compared with the limited database of annotated clauses created by the expert. As a consequence, only POS-tags (*e.g.* noun, adjective, verb present, demonstrative pronoun, and so on) are used to cluster clauses.

In more detail, this program involves several steps. Firstly, the text is divided into n clauses (based on the manual annotation). Secondly, POS-tags are attributed to all the words of each clause with TreeTagger (Schmid, 1994), yielding a distribution over POS-tags. Thus a contingency table between clauses and POS-tags is obtained.

As a next step, clauses are categorized with the thermodynamic clustering procedure, a variant of fuzzy K-means, which amounts to minimizing a free energy term, made up of an energy (the within-cluster dispersion) and an entropy (the clause-cluster mutual information). In a nutshell, fuzzy clustering aims at assigning each clause to the various clusters in a probabilistic fashion. At each iteration step, the membership z_i^g of sentence

i in group g is defined by the following equation (Rose et al., 1990; Bavaud, 2009):

$$z_{ig} = \frac{\rho_g \exp(-\beta D_i^g)}{\sum_{h=1}^{m} \rho_h \exp(-\beta D_i^h)} \qquad (1)$$

where $\rho_g = \sum_{i=1}^n f_i z_{ig}$ is the relative weight of group g and f_i is the relative weight of clause i, D_i^g is the chi-squared dissimilarity between clause i and the centroid of group g, and β is the inverse temperature parameter controlling the number of groups (a larger β implies more groups). At the outset, centroids are chosen randomly (uniformly distributed memberships).

In addition, the user must choose the initial number m of groups, the number N_{max} of maximum iterations and the relative temperature t_{rel} defining the inverse temperature $\beta := 1/\left(t_{\text{rel}} \times \Delta\right)$, where $\Delta := \frac{1}{2}\sum_{ij} f_i f_j D_{ij}$ is total inertia and D_{ij} is the chi-squared dissimilarity between clauses i and j.

Moreover, groups whose profiles are close enough are aggregated, thus reducing the initial number of groups m to the final number of groups M (Bavaud, 2009). In that case, memberships of sentences of similar groups are added in the following way: $z_{i[g\cup h]} = z_{ig} + z_{ih}$. Two groups are considered close if $\theta_{gh}/\sqrt{\theta_{gg}\theta_{hh}} \geq 1 - 10^{-5}$ where $\theta_{gh} = \sum_{i=1}^n f_i z_{ig} z_{ih}$ measures the overlap between groups g and h (Bavaud, 2010).

Also, a factorial correspondence analysis (FCA) is performed to produce a low dimensional representation of the chi-squared dissimilarities D_{ij} between clauses (and between POS-tags).

At the end of the process, each clause is attributed to the most probable group and the results are plotted in 2D (figures 3 and 4).

Moreover, observing the dependency of the effective number of groups as well as evaluation measures (figures 1 and 2) provides a guidance for determining suitable values of the temperature.

2.4 K-means clustering

We also compared the above fuzzy algorithm to the well-known K-means method (see *e.g.* Manning and Schütze (1999)). As for the former, chi-squared dissimilarities are calculated in the algorithm. Two versions are investigated, a weighted and a non-weighted (*i.e.* uniform weights for each clause) approaches.

In K-means, the number m of groups (and not the relative temperature) must be chosen *a priori*. We have concentrated on $m = 6$ (the number of groups in the expert classification) as well as on values of m corresponding to performance peaks in the fuzzy version (see figures 1 and 2).

2.5 Evaluation criteria

Regarding the evaluation, the aim is to compare automatic clustering and expert classification. In addition to χ^2 statistic which measures the dependence between the two classifications, a certain number of similarity indices between partitions exist, among which the Jaccard index, noted J, seems to be a good indicator (Denœud and Guénoche, 2006; Youness and Saporta, 2004):

$$J = \frac{\sum_i \sum_j n_{ij}^2 - n}{\sum_i n_{i\bullet}^2 + \sum_j n_{\bullet j}^2 - \sum_i \sum_j n_{ij}^2 - n} \qquad (2)$$

where n_{ij} is the number of clauses belonging to the unsupervised cluster i and the manual class j.

Another interesting measure is the corrected Rand index (Denœud and Guénoche, 2006):

$$RC = \frac{r - \text{Exp}(r)}{\text{Max}(r) - \text{Exp}(r)} \qquad (3)$$

with $r = \frac{\sum_{i,j} n_{ij}(n_{ij}-1)}{2}$,
$\text{Exp}(r) = \frac{\sum_i n_{i\bullet}(n_{i\bullet}-1)\sum_j n_{\bullet j}(n_{\bullet j}-1)}{2n(n-1)}$,
$\text{Max}(r) = \frac{\sum_i n_{i\bullet}(n_{i\bullet}-1)+\sum_j n_{\bullet j}(n_{\bullet j}-1)}{4}$.

3 Results

3.1 Relevance of the method

To ensure that the choice of using POS-tags is relevant in this context, the dependence between the classification of clauses made by the human expert and the POS-tags they contain must be established. Table 3 reports the corresponding independence ratios ($R_{w,c}$ in Li et al. (2008)) for the three annotated texts by Maupassant. An independence ratio greater than 1 shows a mutual attraction, whereas if it is less than 1, it shows a mutual repulsion. Furthermore, stars in this table indicate the most significant chi2 term-category dependance for each POS-tag with 2 degrees of freedom in relation to $\chi^2_{1-0.001}[2] = 10.83$ (Yang and Pedersen, 1997; Li et al., 2008). It appears that a number of POS-tags are relevant for the types investigated, such as adjectives for the descriptive type ($q = 1.62$ and $\text{chi2} = 27.88$), simple past tense for the narrative type ($q = 2.60$ and $\text{chi2} = 110.55$) or future tense for the dialogal type ($q = 4.63$ and $\text{chi2} = 62.10$). Satisfactorily enough, the value of the chi-square on the contingency table between POS-tags and discourse types ($\text{chi2} = 752.6$ with $\text{df} = 145$) is large, denoting a highly significant link between classes and POS-tags ($p < 10^{-15}$). Moreover, research into genre detection using POS-tags reports interesting results (Karlgren and Cutting, 1994; Kessler et al., 1997; Malrieu and Rastier, 2001), which are, to some extent, relevant for type detection.

	nar	dial	descr	expl	arg	inj
ABR	**2.92**	0.00	0.00	0.00	0.00	0.00
ADJ	0.78	1.07	**1.62***	1.10	0.85	0.75
ADV	0.96	1.02	0.71	1.17	1.04	**1.39**
DET:ART	0.91	0.97	1.15	0.83	**1.22**	0.93
DET:POS	**1.27**	0.76	0.95	0.80	0.93	0.77
INT	**1.34**	**1.34**	0.00	1.05	0.95	0.94
KON	0.93	1.03	0.75	1.19	**1.25**	0.84
NAM	1.00	1.15	1.11	1.03	0.33	**2.15**
NOM	0.92	0.89	**1.20**	0.87	1.15	1.03
NUM	**1.51**	0.52	1.05	0.93	0.74	0.00
PRO	**2.92**	0.00	0.00	0.00	0.00	0.00
PRO:DEM	0.69	0.97	0.95	**1.52**	1.42	0.58
PRO:IND	0.68	1.34	1.08	**1.45**	1.33	0.00
PRO:PER	**1.30***	1.03	0.58	1.05	0.86	0.67
PRO:REL	0.70	1.14	1.25	**1.28**	1.01	1.07
PRP	0.96	0.99	**1.18**	0.98	1.04	0.77
PRP:det	0.59*	1.45	1.31	0.65	1.19	**1.78**
PUN	0.95	0.99	1.15	0.80	1.00	**1.34**
PUN:cit	0.00	4.11*	0.80	0.00	0.23	**4.91**
SENT	**1.16**	0.96	0.79	1.08	0.83	1.05
VER:cond	1.29	0.97	0.00	0.87	**1.83**	0.00
VER:futu	0.53	**4.63***	0.39	0.44	0.17	1.37
VER:impf	1.44	0.38	**2.06***	0.34	0.50	0.12
VER:infi	1.07	0.78	0.89	**1.50**	0.91	0.53
VER:pper	**1.26**	0.91	0.98	0.90	0.80	0.57
VER:ppre	**1.42**	1.09	1.05	0.78	0.31	0.81
VER:pres	0.81	0.98	0.71	1.34	1.07	**1.79***
VER:simp	**2.60***	0.10	0.32	0.00	0.28	0.00
VER:subi	0.53	0.00	0.59	**5.26***	0.00	0.00
VER:subp	0.32	**3.58**	0.00	1.61	0.64	1.67

Table 3: Independence ratio for the three texts by Maupassant (q): numbers indicate the ratio of the observed counts to their expected values under independence. The strongest mutual attraction for each POS-tag is in bold characters. Stars in cells point out the most significant chi-squared per POS-tag ($\alpha = 0.001$).[2]

3.2 Results with automatic fuzzy clustering

Figures 1 to 6 present the results for the method described above. The number of groups after aggregation and the corrected Rand index as a function of the relative temperature are shown for the text ”Un Fou?” in figure 1. These curves are obtained with an initial number of groups $m = 316$ corresponding to the number of clauses $n = 316$ and a number of maximum iterations of $N_{\text{max}} = 400$. The entire process is executed around 20 times for each relative temperature (with randomly chosen initial memberships) and so, values in graphics represent the mean of these 20 simulations. With the same parameters, figure 2 shows the evolution of the Jaccard index with the rela-

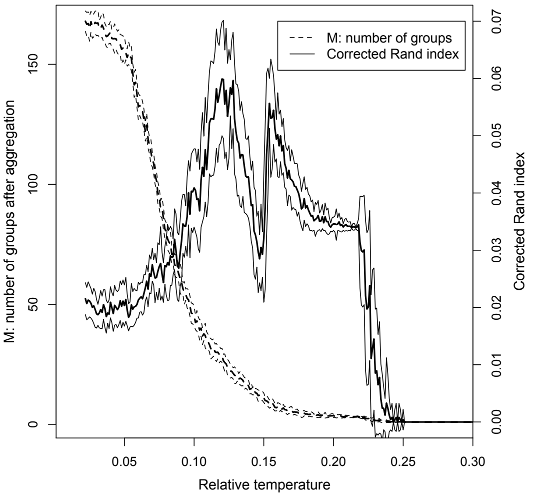

Figure 1: ”Un Fou?” by Maupassant: number of groups and corrected Rand index as a function of the relative temperature. For each curve, the thick line represents the mean and the two thin lines represent the standard deviation.

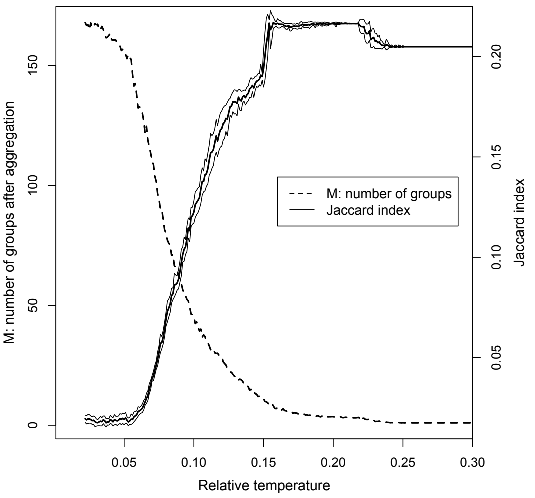

Figure 2: ”Un Fou?” by Maupassant: Jaccard index according to the relative temperature. The curve of number of groups is given for reminder.

[2] A complete explanation about the signification of POS-tags in the table is available on `http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html`

tive temperature. In figure 1, the two remarkable peaks for the corrected Rand index correspond to around 26 and 8.5 groups after aggregation. Figure 2 shows that Jaccard index increases when the number of groups decreases until there is only one group. However, the maximum of Jaccard index appears around 8 groups as does the second maximum of the corrected Rand index. It is obvious that the two indexes give different results. On the one hand, the Jaccard index takes a non-zero value in presence of single group, an artefact due to the absence of correction for self-similarity in (2). On the other hand, the corrected Rand index can take negative values, which means that results are worse than chance.

Similar studies were made for ”L'Orient” and ”Un Fou”. For the former, the corrected Rand index decreases when the relative temperature increases, with two small local maxima for around 96 and 30 goups. For the latter, corrected Rand index is always negative, except with small relative temperatures which correspond to 180 groups. And for the both texts, Jaccard index increases while the number of groups decrease monotically until it becomes maximal for one group.

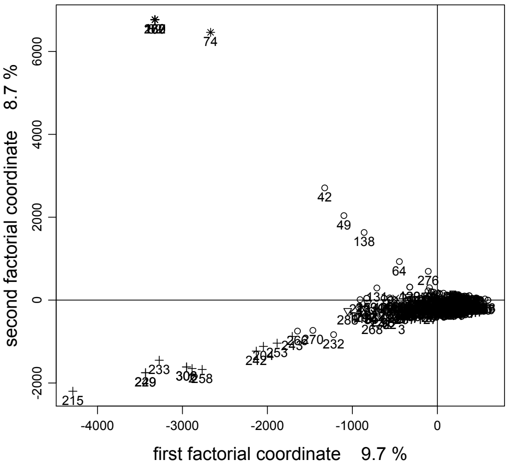

Figure 3: Clustering of clauses of ”Un Fou?” by Maupassant (each symbol belongs to one of the eight clusters and numbers correspond with the position of clauses in the text).

Finally, an example for ”Un Fou?” is given in figures 3 to 6 with the following parameters: $m = 316$, $N_{\max} = 400$ and $t_{rel} = 0.157$ designed to produce $M = 8$ groups after aggregation, because the two evaluation indexes have interesting value for this number of group. In figure 3, all clauses ($n = 316$) are represented in a 2D plot. Figure 5 represents dissimilarities between POS-tags in the same space as figure 3. For all these figures, dissimilarities are not well represented, since the expressed inertia is only 18.4%.

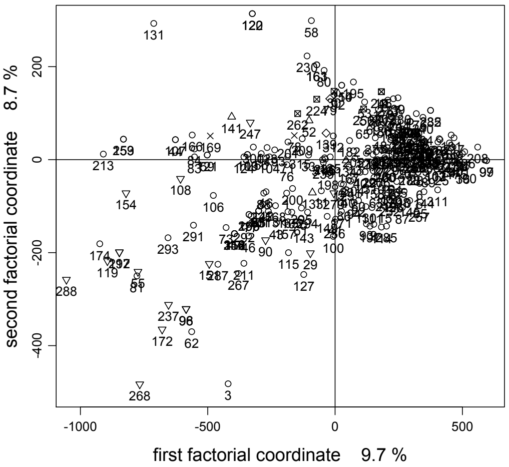

Figure 4: Zoom of figure 3.

4 Preliminary evaluation

		Classes identified by the expert						Total
		arg	descr	dial	expl	inj	nar	
Clusters	1	48	30	33	34	15	101	261
	2	4	0	2	1	0	0	7
	3	0	0	1	1	0	0	2
	4	0	0	2	0	0	0	2
	5	0	2	2	0	8	0	12
	6	2	0	1	0	3	0	6
	7	5	0	5	4	0	3	17
	8	0	1	0	6	0	2	9
Total		59	33	46	46	26	106	316

Table 4: Cross-counts between unsupervised and manual classification.

Table 4 shows cross-counts between automatic clusters and classes assigned by the human expert corresponding to the analysis of figures 3 to 6. The chi square reveals a strong relation between automatic clustering and expert classification (chi2 $=$ 137.28 with df $= 35$ and $p < 10^{-13}$). For this table, other evaluation criteria are less satisfactory ($RC = 0.06$ and $J = 0.22$).

In addition to the results obtained above with the fuzzy clustering, K-means (respectively fuzzy clustering) was performed on the three texts for 6 groups (respectively with a relative temperature yielding around 6 groups) and on ”Un Fou?” and ”L'Orient” for a number of groups (respectively relative temperature) corresponding to the best

	Method	♯S	m	M		$N_{\max}$	N_{eff}		chi2		df		J		RC	
				mean	sd		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
”Un Fou?”	NW K-means	300	6	-	-	200	14.26	4.49	155.96	28.34	25	-	.18	.01	.06	.02
	W K-means	300	6	-	-	200	16.18	5.60	152.46	28.21	25	-	.15	.01	.07	.02
	fuzzy t_{rel}=0.165	40	316	6.1	1.71	400	182	49.1	104.26	14.70	25.5	8.53	.21	.00	.04	.01
	NW K-means	89	26	-	-	200	12.62	3.16	338.04	19.58	125	-	.07	.01	.05	.01
	W K-means	225	26	-	-	200	11.68	2.78	335.66	16.56	125	-	.07	.00	.05	.01
	fuzzy t_{rel}=0.125	40	316	23.8	2.66	400	195.1	73.2	265.29	20.42	113.9	12.83	.17	.01	.06	.01
”L’Orient”	NW K-means	300	6	-	-	200	12.66	3.76	63.09	8.94	25	-	.15	.01	.02	.01
	W K-means	300	6	-	-	200	12.42	3.81	69.95	13.75	25	-	.14	.01	.04	.02
	fuzzy t_{rel}=0.18	40	189	6.48	0.82	400	161.2	61.5	48.24	5.83	27.38	4.08	.20	.00	-.01	.00
	NW K-means	144	30	-	-	200	9.22	2.25	225.97	19.95	145	-	.06	.01	.03	.01
	W K-means	207	30	-	-	200	8.76	1.79	229.74	19.69	145	-	.06	.00	.04	.01
	fuzzy t_{rel}=0.117	40	189	30.8	3.04	400	290.8	63.2	220.92	29.93	149	15.20	.13	.01	.01	.01
”Un Fou”	NW K-means	300	6	-	-	200	17.00	6.22	63.20	16.45	25	-	.16	.01	-.04	.01
	W K-means	300	6	-	-	200	17.80	6.43	70.61	19.33	25	-	.15	.01	.00	.01
	fuzzy t_{rel}=0.188	40	400	6.08	1.07	400	89.1	27.5	142.89	11.72	25.38	5.36	.24	.01	-.04	.02

Table 5: Comparison of results between K-means (non-weigthed (NW) and weighted (W)) and fuzzy clustering algorithms. ♯S denotes the number of random starts on which the mean and the standard deviation (sd) are computed for evaluation criteria and other values. N_{eff} refers to the effective number of iterations needed to stabilize the group centroid positions.

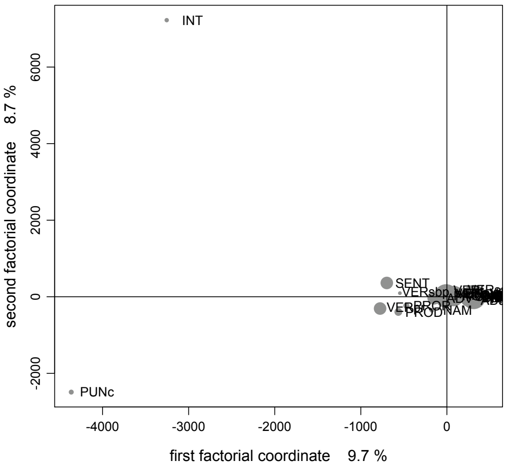

Figure 5: Representation of POS-tags of ”Un Fou?” by Maupassant.

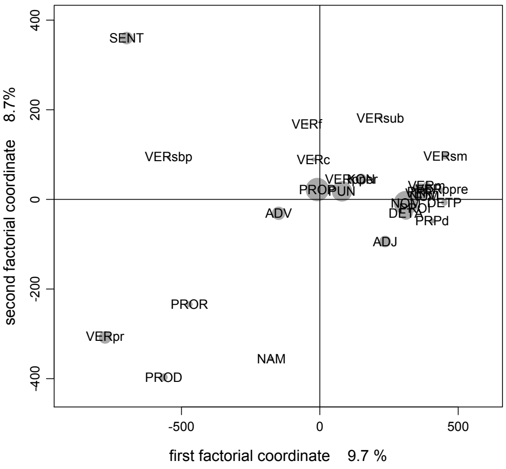

Figure 6: Zoom of figure 5.

performances. Evaluation criteria for the three texts and all methods are summarized in table 5. It is obvious that the three evaluation criteria do not imply the same conclusions. For instance, chi2 values indicates that for ”Un Fou?” and 6 groups, the non-weighted K-means induces the most promising classification. Regarding the Jaccard index, the fuzzy clustering seems to involve better results. As for the corrected Rand, it shows that weighted K-means improves the clustering. However, a certain number of regularities transpire. For ”Un Fou?”and 26 groups, the weighted K-means never implies the best results according to the three criteria, while for ”L’Orient” and ”Un Fou”, the non-weighted K-means never induces the best classification. Finally, the corrected Rand index is low, or even negative, for ”Un Fou”. Perhaps it is due to the fact that this text is partly different, even if it is a short story as the both other texts. Indeed, this text is made of a journal part.

To conclude, all these preliminary results must be considered with caution, in regard to the small size of the sample, annotated by a unique expert.

5 Work in progress

Despite encouraging first results, demonstrating the dependence between the clusters obtained based on POS-tags and the linguistic types assessed by the human expert, the limitations are obvious, and further improvements have to be explored. First of all, it will be interesting to apply a bi- or trigram model to replace individual POS-tags. Besides this, using only POS-tags might reveal itself no sufficient, and calling for considering the inclusion of typical words which discriminate, in a certain proportion, the different discourse types. And, in the same line, feature selection between POS-tags could improve results (see *e.g.* Yang and Pedersen(1997); Li et al. (2008)). It is also crucial to consider and exploit the hierarchical structure of discourse types. One way to do this could be to take into account the dominance of one type over others in a part of the hierarchical structure. Moreover, the use of other measures of clause dissimilarities, alternative to the chi-squared distances, may improve clustering results. Furthermore, combining fuzzy clustering and K-means as in the "simulated annealing" approach of Rose et al. (1990) should be explored. Finally, the possibility of automatically segmenting the text into clauses should be considered.

References

Jean-Michel Adam. 2005. *La linguistique textuelle: Introduction à l'analyse textuelle des discours*. Armand Colin, Paris.

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

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. *ECML PKDD 2010: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, J.L. Balcázar et al. (Eds.), 6321:103–118.

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

Jean-Paul Bronckart. 1997. *Activité langagière, textes et discours: pour un interactionisme socio-discursif*. Delachaux et Niestlé, Lausanne; Paris.

François Daoust, Yves Marcoux and Jean-Marie Viprey. 2010. L'annotation structurelle. *JADT 2010: 10th International Conference on Statistical Analysis of Textual Data*.

Lucile Denœud and Alain Guénoche. 2006. Comparison of Distances Indices Between Partitions. *Data Science and Classification*, 21–28.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. *Proceedings of the 15th conference on Computational linguistics*, 2. Kyoto, Japan.

Brett Kessler, Geoffrey Nunberg and Hinrich Schütze. 1997. Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 32–38. Madrid, Spain.

Moshe Koppel and Jonathan Schler. 2003. Exploiting Stylistic Idiosyncrasies for Authorship Attribution. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 69–72.

Yanjun Li, Congnan Luo and Soon M. Chung. 2008. Text Clustering with Feature Selection by Using Statistical Data. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):641–652.

Denise Malrieu and François Rastier. 2001. Genres et variations morphosyntaxiques. *Traitement Automatique des langues*, 42(2):548–577.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

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.

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. *Proceedings of the International Conference on New Methods in Language Processing*, 44–49.

Yiming Yang and Jan O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, 412–420.

Genane Youness and Gilbert Saporta. 2004. Une méthodologie pour la comparaison de partitions. *Revue de statistique appliquée*, 52(1):97–120.