# Automatic Analysis and Annotation of Literary Texts

R. Basili (†), A. Di Stefano (‡), R. Gigliucci (‡), A. Moschitti (†), and M. Pennacchiotti (†)

(†) Dipartimento di Informatica, Sistemi e Produzione
{basili,moschitti,pennacchiotti}@info.uniroma2.it

(‡) Dipartimento di Studi Filologici, Linguistici e Letterari
University of Rome Tor Vergata,
{alicedistefano,robertogigliucci}@tiscali.it

**Abstract.** In this work a machine learning oriented perspective on computer aided support to literary analysis is presented. A representation of narrative phenomena is proposed and an automatic annotation model for such phenomena is trained on texts provided by a critic. As a short-term research task, we studied how the observable textual piece of evidence impact on the learning agent capabilities, over a specific literary work, i.e. "*Gli Indifferenti*" by Alberto Moravia. Although preliminary, the results are good and confirm the viability of the approach for larger scale studies.

## 1 Introduction and Motivation

It has been noticed ([1]) that computer based literary critic is still relying on studies of *concordances* as traditionally intended since the 13th century. All the intermediate digital representations (storage, indexes, data structures or records) are not capitalized although they can play the notion of a new literary *monster* (like the Cheiron centaur) as a new meaningful, artistic and hermeneutic macro unit. It is indeed true that the digital representation, its metadata and its digital derivatives (e.g. indexes, parse trees, semantic references to external dictionaries) are new and more complex forms of "*concordances*" and should be used by the literary scholar in cooperation with the original content. New processes of narrative analysis should thus take all of this into account by exploiting the fruitful interactions among the parts of the monster within suitable software architectures (that are thus more complex than digital archives/catalogs). The long term enterprise of this research is thus to design novel platforms for supporting studies of narrative texts that are:

- more *computer-centred*, as they work at a higher level of abstraction
- *interactive* with the scholar, as the software is proactive with respect to the literary work
- multifunctional and integrated as they support incremental refinement of internal knowledge of the opera along with more interaction with the expert takes place.

In this work an early exploration of these ideas has been carried out. According to a preliminary (and machine learning oriented) perspective, we studied a representation of narrative phenomena as an explicit knowledge model able to provide a high abstraction level about the literary work. Over this model we designed an automatic (autonomous and proactive) indexing model for such narrative phenomena able to capitalize examples in the text provided by the literary expert and learn to recognize them in unlabeled texts. As an intermediate research task, we studied at which extent the observable textual evidence supports the learning of a narrative phenomenon in a specific work (i.e. "*Gli Indifferenti*" by Alberto Moravia, [2]). In order to achieve this goal, we evaluated the impact of several levels of textual information (e.g. from orthographic properties to morphological, syntactic and semantic features) on the inductive capability of the resulting agent. The design (Sections 2 and 3) and the results of our experiments (Section 4) will be discussed throughout this paper.

## 2 An ontology of narrative elements

The critical study of a literary work usually starts from the identification and the analysis of specific narrative phenomena, such as the use of schemas or peculiar descriptions. A computational tool able to support a critical study should identify and model these aspects: what is needed is thus a representation of the narrative phenomena as an explicit knowledge model. The target model can be used both to manually annotate the literary work and to feed a machine learning algorithm, thus supporting proactively the critical analysis. For instance, a knowledge model could represent and hierarchically organize the use of figurative language, dividing tropes and schemas, and successively classifying different rhetorical figures.

The task of finding and modelling interesting narrative phenomena should be intended as a cooperation between a pool of critics skilled on a specific author and a pool on knowledge engineers. On the one hand, the experts should identify the most interesting narrative events. On the other hand, the engineers should be responsible of organizing the events in an unambiguous explicit knowledge model. The final result will then be a reference *ontology of narrative elements* that can be used both for annotation and learning.

During the annotation phase, a skilled critic is requested to carefully analyse the opera and annotate all the text fragments referring to one or more narrative elements (for example all tropes). Therefore, a formal model for the notion of *narrative annotation* is also needed. A narrative annotation defines precisely how a *text fragment* is meaningful to a narrative phenomenon and how this is linguistically expressed. The following simplifying assumptions can be made:

- A text fragment that instantiates a specific occurrence of a narrative phenomenon is a sequence of consecutive words in the texts.
- Specific types of text fragments instantiating different phenomena can be nested one in the other (they can overlap only in this way).

In this view, a simple mark-up language, such as XML, can be used as a narrative annotation model, in which *elements* identify narrative phenomena and whose *attributes* serve as description of the specific phenomenon expressed in the fragment. In order to keep the annotation task conform to the standards used by the web-literary community, the XML-based TEI formalism [3] (today acknowledged as a standard formalism for sharing literary works over digital media) has been adopted as mark-up language. We extended the TEI tagset with specific *narrative elements* that capture all the interesting type of narrative phenomena defined by an expert. We defined two main elements: the `NAR` tag and the `SUB-NAR` tag. The first tag is intended to capture *narrative macro-structures*, that is an entire text fragment with generic narrative attributes. The `SUB-NAR` tag is more specific and is always nested in the `NAR` tag: it describes fragments with all the properties of the embedding `NAR` tag plus the attributes of the embedded `SUB-NAR`. This annotation strategy is used to better capture the nesting behaviour of many narrative phenomena. Every specific narrative event (`NAR`) is described using XML attributes, whose definition depends on the different target literary work. The next section discusses the main choices made for *"Gli Indifferenti"*.

### 2.1 A case study on "Gli Indifferenti" by Alberto Moravia

*"Gli Indifferenti"* is an Italian novel written by Alberto Moravia in 1929, which describes the drama of a middle-class Italian family, set in Rome during the Fascist regime. Moravia's style is lucid, direct and rational, characterized, in a linguistic perspective, by extreme simplicity. Moravia's debut, *"Gli Indifferenti"* has been chosen as a case study for its literary value and for its exemplification of narrative phenomena that will recurr in its following work. Notwithstanding, the novel is unique in its theatrical structure: many dialogues and actions are confined in space and time, as actually it happens on a theatre stage. Moreover, much attention is devoted to highly detailed elements such as lights, clothing, objects and characters' movements. The plot crosses different and fairly recognizable environments: rooms of family houses or the city open spaces. Consequently, to distinguish and isolate actions and descriptions in the novel is relatively easy. A distinguishing trait of the Moravia's prose, specifically devoted to the representation of reality, is the didascalic and realistic introductory description of people and places. Text fragments targeted to describe people and places have thus a narrative interest. Accordingly, a computational model that automatically spots these fragments in the novel has an impact on the scholar work. The "*ontology of narrative elements*" for "Gli indifferenti" defines *descriptions*, in terms of four main classes, denoted by a `OGG` (object) tag: *External Place, Internal Place, Male Person* and *Female Person*. Each description has two distinct properties. The first is *typology*, i.e. the possible use of figurative language in the description. When figures are used, the description is said to have a *symbolic* typology, otherwise it is *objective*. The second is *expressiveness*, that characterizes the prose complexity.

"Sicuro", rispose Leo accendendo una sigaretta; "forse non mi vuoi?" <NAR OGG="pm" TIPO="o" EXP="s">Curvo, seduto sul divano, egli osservava la fanciulla con una attenzione avida;</NAR> <NAR OGG="pf" TIPO="o" EXP="s"><SUB_NAR PDV="Leo">gambe dai polpacci storti, ventre piatto, una piccola valle di ombra fra i grossi seni, braccia e spalle fragili, e quella testa rotonda cosi' pesante sul collo sottile.</SUB_NAR></NAR>

**Fig. 1.** An example of TEI narrative annotation from "Gli Indifferenti" (chapter 1).

In the annotation model, a *description phenomenon* is identified as a text macro-structure denoted by the `NAR` tag, whose specific attributes are: `OGG`, to identify one of the four classes, `TIPO` to define typology and `EXP` to identify the expressiveness.

Another interesting aspect of a description is the *viewpoint*: sometimes the description is presented by the author through the eyes of one of the characters. Finally, the ontology also defines descriptions in which an *action* takes place. The point of view and the presence of an action are formalized in the annotation model through the `SUB-NAR` tag, as they can both be viewed as narrative phenomena nested in more generic descriptive events. Each character of the novel is thus a possible value for a viewpoint (`PDV`) attribute of the `SUB-NAR` tag while `ACTION` is a further Boolean attribute.

The whole original Italian version of Moravia's novel includes sixteen chapters and is made by roughly 91.000 words. It has been annotated by a pool of critics, resulting in a complete electronic TEI version of the work, browsable and accessible through the Web. An example of annotation is shown in Fig. 1

## 3 A machine learning approach to narrative phenomenon recognition

The annotated novel has been used to apply machine learning algorithms to recognize and automatically annotate specific narrative fragments. An example-driven algorithm has been designed using as *training set* the fragments annotated by the experts. By relying on textual information (i.e. *features*) as observed in these example fragments, the algorithm learns a narrative model, and automatically annotates the set of unseen (*test*) fragments.

### 3.1 Automatic recognition of narrative sections

As the minimal subpart of a narrative fragment can be assumed to span over a paragraph, the problem of annotating narrative fragments can be divided in the two subtasks:

1. the *detection of interesting paragraphs*, i.e. those including target *narrative descriptions*; and
2. the *classification of relevant paragraphs* according to the target *description type*, e.g. *External Place*, *Internal Place*, *Male Person* and *Female Person*.

The automatic detection of interesting paragraphs can be carried out by a binary classifier, hereafter called *IPC*. An IPC selects only the fragments which truly contain *descriptions*. Once the set of relevant fragments are available, labelling consists in applying a pool of $n$ binary description classifiers (one for each description type). Such classifiers can be learned with the ONE-vs-ALL approach [4] and, in case of more than one label is accepted, the label with the highest probability (e.g. the highest score among the binary classifiers) is retained. The result is a description type multiclassifier ($DTC$).

In literature many classification approaches (binary and multiclass classifiers) have been described; among others, Support Vector Machines have shown satisfactory accuracy when few and noisy training data is available.

### 3.2 Binary classification with Support Vector Machines (SVMs)

SVMs represent linguistic objects, e.g. narrative descriptions, by means of feature vectors in the $\Re^n$ vector space. Accordingly, each representation expresses a set of features extracted from the source paragraphs. Each feature is characterized by a weight, i.e. the vector component. In the simplest case weights are 1 or 0 (if a feature appears or not in the text).

Given the vectors $\boldsymbol{x}$ of positive (i.e. interesting) and negative (i.e. uninteresting) paragraphs, SVMs learn the hyperplane which separates positive and negative examples with the highest margin. More formally, by applying the *Structural Risk Minimization principle* [5], SVMs compute the linear function $H(\boldsymbol{x}) = \boldsymbol{w} \times \boldsymbol{x} + b = 0$, where $\boldsymbol{x} \in \Re^n$ is the hyperplane variable, $\boldsymbol{w} \in \Re^n$ is the gradient and $b \in \Re$ is a constant. A new object $\boldsymbol{x'}$ is classified as an instance of the target linguistic phenomenon *iff* $H(\boldsymbol{x'}) > 0$.

The most critical design phase for SVMs is the definition of features effective to encode the target phenomenon in the classification function. The next Section illustrates the different sets of linguistic features adopted to characterize the description of narrative sections.

### 3.3 Representing Narrative Phenomena

The detection of narrative fragments needs different kind of information (i.e. features). Specifically, the complexity of description types can only be captured by a mix of different linguistic layers, ranging from orthography/morphology to syntax and semantics. Accordingly, different sets of linguistic features for the SVMs have been designed. This distinction has been kept explicit during the learning and testing phases to study the contribution of individual linguistic layers.

**Orthographic features** (`orth`). They represent the bag of words model of the novel at the orthographic (i.e. surface) level. That is, each *plain word* (adjectives, adverbs, nouns and verbs) appearing in the text is considered as a distinct feature (we found 19.274 unique words). The value of each feature is the word frequency in the target paragraph.

**Morphological features** (`morph`). Morphological features include two different feature subsets:

- **Lemmas** (`lemmas`). These features represent the morphological (i.e. lemma) level. Each lemma in the novel is a feature: for example, *ando'* (*went*) and *andato* (*gone*) are represented by the same feature *andare* (*to go*). 8.703 different lemmas were found in the novel. The value of these features is the frequency of individual lemmas in each paragraph.
- **Part of Speech** (`POS`). These features capture morphosyntactic properties, i.e. the Part of Speech of individual words. The rate of male/female nouns and adjectives in a paragraph are examples of these features. Also the rate of different verb forms with respect to all verb forms (e.g. finite, conjunctive, conditional, gerund forms) and the rate of different verb tenses (present, simple past, etc.) are computed as separate features.
- **Named Entities** (`NE`). This single feature represents the rate of *person* Named Entities in a paragraph.

**Syntactic features** (`synt`). Simple binary syntactic relations are observed. The POS bigrams considered are (male and female) adjective-noun pairs and (male and female) noun-verb pairs. The value of these features are given by the number of each bigram over the total number of bigrams in the fragment.

**Semantic features** (`sem`). These features try to capture the semantic properties of the lexicon. Both verbs and nouns are generalized through MultiWordNet [6] and mapped to specific *semantic classes*. Classes are somewhat dependent on the novel as they are assumed to provide relevant information to the narrative classes. Semantic noun classes are: *furniture*, *clothing*, *body parts*,*external places* and *internal places*. Semantic verb classes are derived from the WordNet Lexical files [7], including 15 abstract classes such as *communication*, *emotion* and *creation* verbs. Each class represents a feature. Using MultiWordNet, lemmas are generalized in each class: each feature is then valued as the rate between the class frequency and the lemma in a fragment.

The extraction of morphological and syntactic properties of words and lemmas is carried out by Chaos ([8]), a modular dependency-based syntactic parser for Italian and English texts, developed at the AI-NLP laboratory of the University of Roma, Tor Vergata.

## 4 Preliminary Results

In these experiments we tested the accuracy in retrieving interesting paragraphs (description detection) and classification accuracy of the description type, i.e. *External Place*, *Internal Place*, *Male Person* and *Female Person*. Additionally, we tested the impact of different feature, i.e. orthographic, morphological, syntactic and semantic features, on classification accuracy.

### 4.1 Experimental Set-Up

As referring corpus we used the electronic version of *Gli Indifferenti* (91.000 words), authored by Alberto Moravia, in which all narrative sections related to

description of places and people were manually annotated. In particular, 395 paragraphs out of a total of 2326 (17%) were assigned by human annotators to descriptive sections. From such 395 paragraphs, 51 were assigned to external places, 113 to internal places, 156 to female people and 75 to male people. For the experiments, we used the SVM-light software [9] (available at `svmlight .joachims.org`) with the default linear kernel and the default value for the trade-off (between training error and margin) parameter (`-c` option). We noted also that the cost-factor parameter (i.e. `-j` option) is not critical, i.e. a value of 2 optimizes the accuracy for almost all classifiers. We divided such corpus in 30% for testing and 70% for training, thus the interesting paragraph binary classifier (IPC) was tested on 698 whereas the description type multiclassifier (DTC) was tested on 118 instances.

The IPC performance was evaluated by means of the $F_1$ measure[1] whereas DTC performance was measured by means of Accuracy (1 - error rate). Additionally, the individual type classifier performance was derived with $F_1$.

### 4.2 Classification results

Table 1 shows the $F_1$ of IPC according to different feature sets. Columns 2, 3, 4, 5 and 6 show the result when the orthographic (*orth*), lemmas (*lemmas*), morphological (*morph*), syntactic (*synt*) and semantic (*sem*) features are, respectively, used. Columns 7, 8, 9 and 10 report the *orth-morph*, *orth-synt*, *morph-synt* and *all* feature class combinations, respectively.

| | *orth* | *lemmas* | *morph* | *synt* | *sem* | *orth-morph* | *orth-synt* | *morph-synt* | *all* |
|---|---|---|---|---|---|---|---|---|---|
| $F_1$ | 84.6 | 85.7 | 87.2 | 18.6 | 62.2 | 88.6 | 83.6 | 86.1 | 88.6 |

**Table 1.** $F_1$ of diverse linguistic features for the automatic detection of interesting paragraphs.

Regarding the description type classification, Table 2 shows the accuracy of DTC according to different information levels. Columns from 2 to 10 report the results when the *orth*, *lemmas*, *all*, *synt*, *sem*, *orth-morph*, *orth-synt*, *morph-synt* and *all* feature sets are used.

| | *orth* | *lemmas* | *morph* | *synt* | *sem* | *ortho-morph* | *ortho-synt* | *morpho-synt* | *all* |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy | 73.7 | 72.8 | 73.7 | 22.2 | 45.6 | 78.1 | 73.7 | 75.4 | 79.0 |

**Table 2.** Accuracy produced by diverse linguistic features for the automatic classification of the description type.

As a conclusive evaluation, Table 3 reports the $F_1$ of the individual description type classifiers, i.e. *External Place*, *Internal Place*, *Female Person* and *Male Person*.

[1] $F_1$ assigns equal importance to Precision $P$ and Recall $R$, i.e. $F_1 = \frac{2P \cdot R}{P+R}$.

| Category | Precision | Recall | $F_1$ |
|---|---|---|---|
| *External Place* | 61.54 | 57.14 | 59.26 |
| *Internal Place* | 77.78 | 63.64 | 70.00 |
| *Female Person* | 67.24 | 86.67 | 75.73 |
| *Male Person* | 57.14 | 72.73 | 64.00 |
| Accuracy | 79.0 | | |

**Table 3.** $F_1$ of the individual description type classifier using all linguistic features.

### 4.3 Discussion of results and Conclusions

Our preliminary experiments on paragraph detection (see Table 1) provide material for a more in depth discussion. First, the detection of interesting paragraphs seems to be rather simple. Using the orthographic features alone, i.e. the traditional *bag-of-words* without stemming, the system achieves an $F_1$ of 84.6%. This is a quite high result, since it is similar to the one obtained on the easier task of document categorization.

Secondly, the representational power of lemmas ($F_1$ of 85.7%) is higher than the one of the orthographic features, i.e. simple words (in agreement with Information Retrieval literature). This suggests that in language with rich morphology (e.g. Italian), lemmatization is important. Moreover, when the morphology features are added to the lemmas (i.e. when we use the *morph* feature set) the $F_1$ increases of 1.5%. Note that the orthographic features contain the morphological and the lemma information but given the low number of training examples it is more convenient to keep them separated. Indeed, as lemmas have a higher probability to be matched than words, the system recall increases, whereas the separated morphological features help to preserve the precision.

The syntactic features used alone or in conjunction with other feature sets seem to penalize system accuracy. The major reason is that the basic features accuracy is already quite high: thus, the error rate of the syntactic parser (higher than the classification system error rate) decreases the overall accuracy.

Finally, when the orthographic features are used in conjunction with the *morph* set the system achieves the highest $F_1$, 88.6%, which is the same obtained with *all* feature sets.

The results on description type classification (see Table 2) are slightly different from the previous outcomes. First, *lemmas* produce a lower accuracy than *orth* (72.8% vs. 73.7%). This is explained by the higher importance of word morphology for this task. For example, to distinguish between male and female descriptions the whole word surface is relevant; lemmas may not provide enough information for this (e.g. *bello* and *bella* both transform in *bello*).

Second, as for IPC, the *orth* and *morph* combination provides a higher accuracy, 78.1% whereas in contrast to the IPC results, the syntactic information seems to improve the *morph* set (73.7% vs. 75.4%). Possible reasons are:

- 73.7% is more than 10 percent points lower than the one provides by IPC, i.e. this value is comparable with the syntactic parser accuracy: thus, parsing errors do not impact remarkably.

– the syntactic parser is applied to a subset of all possible paragraphs, i.e. those related to descriptions. Thus, its errors can be easily generalized by the learning classification algorithm, which can learn to recognize and avoid them.

Finally, the highest performance is achieved with *all* feature set, suggesting that the syntactic and the semantic information can impact remarkably, as type classification is a more conceptual task.

Finally, Table 3 shows that the $F_1$ of the individual classifiers, i.e. *External Place*, *Internal Place*, *Female Person* and *Male Person*, are strictly proportional to the size of the available training data. As expected, the categories are quite symmetric. The related classification problems thus show similar complexities. Moreover, the individual classifier $F_1$ are quite lower than the combined classifier accuracy, i.e. 79%. This is not surprising if we consider that we transformed binary classification in multiclassification by means of a voting strategy (according to maximum score) which usually improves the classification accuracy of individual voters.

In conclusion, in this paper we have presented a novel and viable approach to the analysis of literary work. The experimental results suggest that the discovery of interesting narrative phenomena can be automated with a good accuracy. The approach is thus already applicable to a new framework for literary study that better harmonises computer-based analysis and the expert's preferences, opening more powerful perspectives for proactive human-machine "*literary*" interfaces.

**Acknowledgments**

Authors want to acknowledge the invaluable discussions had with prof. Andrea Gareffi that inspired most of the principles underlying the presented work.

## References

1. Rockwell, G.: What is text analysis, really? Literary and Linguistic Computing **18, n. 2** (2003) 201–219
2. Moravia, A.: Gli Indifferenti. Bompiani (1929)
3. Sperber-McQeen, C., Burnard, L., eds.: TEI P4: Guidelines for Electronic Text Encoding and Interchange. 4 edn. Oxford (2002)
4. R., R., A., K.: In defense of one-vs-all classification. Journal of Machine Learning Research **5** (2004) 101–114
5. V., V.: The Nature of Statistical Learning Theory. Springer (1995)
6. Bentivogli, L., Pianta, E., Girardi, C.: Multiwordnet: developing an aligned multilingual database. In: First International Conference on Global WordNet, Mysore, India (2002)
7. Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM **38** (1995) 39–41
8. Basili, R., Zanzotto, F.M.: Parsing engineering and empirical robustness. Natural Language Engineering **8/2-3** (2002)
9. Joachims, T.: Making large-scale SVM learning practical. In Schölkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods - Support Vector Learning. (1999)