Short Term Mobility Research Program - 2013

Dr. Martijn Wieling

University of Tübingen, Department of Quantitative Linguistics

Title of the work program

Modelli di variazione della lingua italiana attraverso le sue varietà d'uso
Models of variation of Italian across usage-based language varieties

The goal

Over the last two years, one of the lines of research of the ItaliaNLP Lab at ILC-CNR (www.italianlp.it) has been devoted to tasks aimed at identifying the language variety, the author, the text genre or the level of readability of a text by exploiting the distribution of a wide variety of multi-level linguistic features automatically extracted from texts. Besides differences at the level of the typology of selected features, these different tasks have been tackled within the same methodological paradigm, i.e. by combining NLP-enabled feature extraction and machine learning. Results achieved in the different tasks are described in several papers, listed below:

- automatic readability assessment
  - readability assessment of documents and sentences
    READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification
    http://www.aclweb.org/anthology-new/W/W11/W11-2308.pdf
  - readability assessment of documents across textual genres
    Genre–oriented Readability Assessment: a Case Study
    http://aclweb.org/anthology//W/W12/W12-5812.pdf

- textual genre classification
  Linguistic Profiling of Texts and Document Categorization by Genre and Readability. An Exploratory Study on Italian Fictional Prose
  Paper accepted at RANLP-2013 to appear in the Conference Proceedings

- native language identification
  Linguistic Profiling based on General–purpose Features and Native Language Identification
  http://aclweb.org/anthology-new/W/W13/W13-1727.pdf

Within this context, the goal of the proposed *Short Term Mobility* program was to investigate role and typology of features contributing to the different tasks, with a twofold aim: on the one hand, to understand what are the features which contribute most significantly to the classification of texts by readability, genre and native language of the L2 writer; on the other hand, to identify the optimal number of features to be used within the classifiers developed for each task.

A further line of research which started being investigated during Martijn Wieling visit was devoted to the study of patterns of lexical variation in Tuscany and of the underlying linguistic features, based on the dialectal data of the Lexical Atlas of Tuscany (*Atlante Lessicale Toscano*).

Methods and techniques

To pursue this goal, we decided to experiment with different methods used in the literature to carry out feature selection and ranking for classification. The following techniques have been identified as potentially relevant for the tasks at hand:

1. Hierarchical spectral partitioning of bipartite graphs
2. Grafting-based feature selection

1. Hierarchical spectral partitioning of bipartite graphs

It turned out that the bipartite approach is not suitable for the readability assessment, genre identification and native language identification tasks due to the characteristics of the datasets we have been dealing with and the high number of features used in the native language identification task. On the other hand, the method appears to work well with the lexical dialectal data.

2. Grafting-based feature selection

The idea behind grafting is that it iteratively selects the best features in the maximum entropy framework. In this case besides a classifier, also a ranking of features is obtained, one of the goals of the project. Given the ranking, it is straightforward to iterate over an increasing number of features and evaluate its performance on a test set. For grafting we have used the "tinyest" implementation made by Daniël de Kok (of the University of Groningen). His implementation is specific for ranking, so we used it as a binary classifier (score 0, the lowest, versus score 1, the highest). As readability is either simple or difficult, this classifier can directly be used. For genre-identification, four classes are distinguished. We therefore constructed four models (each one distinguishing one class from all others) and combined these in a second step, by assigning the genre which had the highest probability of being the right one (according to our four models). The ranking of the features is obtained by applying this approach to the whole training set. The optimal number of features can then be found by splitting the complete training set into 10 different sets of training data (90% of the original size) and test data (10% of the original size) and using the test data sets to evaluate the performance of the training data sets for an increasing number of features (adhering to the ranking obtained earlier). Consequently, this method is in principle able to reach both goals of this STM program.

Two other methods used in the literature were also experimented with (namely *Principal Component Analysis* and *Random Forests*) which led however to unsatisfactory results.

Used datasets

The following table contains a description of the datasets used for the different tasks:

Abbreviation name	Corpus	Coarse–grained genre	N. documents	N. words
Rep	*La Repubblica*, Italian newspaper Marinelli et al. (2003)	*Journalism*	321	232,908
2Par	*Due Parole*, easy–to–read Italian newspaper Piemontese (1996)	*Journalism*	322	73,314
ChildLit	*Children Literature* Marconi et al. 1994	*Literature*	101	19,370
AdLit	*Adult Literature* Marinelli et al. (2003)	*Literature*	327	471,421
ChildEdu	*Educational Materials* for Primary School Dell'Orletta et al. (2011)	*Educational*	127	48,036
AdEdu	*Educational Materials* for High School Dell'Orletta et al. (2011)	*Educational*	70	48,103
Wiki	*Wikipedia* articles from the Italian Portal "Ecology and Environment"	*Scientific prose*	293	205,071
ScientArt	*Scientific articles* on different topics (e.g. climate changes and linguistics)	*Scientific prose*	84	471,969

For native language identification, we used the TOEFL-11 learner essay corpus (Blanchard et al. 2013) distributed in the framework of the first shared task on Native Language Identification (NLI) which was organized in the framework of the NAACL-HLT 2013 Workshop on *Innovative Use of NLP for Building Educational Applications* (http://www.cs.rochester.edu/~tetreaul/naacl-bea8.html).

Experiments and results

TASK 1: ASSESSING DOCUMENT READABILITY

Experiment 1

Method: grafting
Training Data: readability corpora (all genres combined)
Test Data: all genres combined as well as individual genres (Journalism, Educational, ScientificProse, Literature)
Steps:

1) obtain the best L1 value using 10-fold cross validation on the training set;
2) use this L1 value with grafting to obtain the ranking of the 90+ features;
3) iteratively select an increasing number of features and evaluate the performance using 10-fold cross validation. Choose the number of features after which improvements are minimal.
4) evaluate the performance of the selected number of features on the test set.

Interesting results are obtained, showing that already using only 2 features (`MediaLunghezzaParole` and `MediaLunghezzaFrasi`, corresponding to the features used in traditional readability measures) results in a performance of about 70%.

The performance graphs on the basis of the feature count and the ranked lists (see "results-tt-all*" files) are stored in http://www.let.rug.nl/wieling/grafting-readability/results/

Experiment 2

Method: grafting
Training Data: genre-specific readability corpora
Test Data for the individual genres (Journalism, Educational, ScientificProse, Literature)
Steps:

1) obtain the best L1 value using 10-fold cross validation on the training set;
2) use this L1 value with grafting to obtain the ranking of the 90+ features;
3) iteratively select an increasing number of features and evaluate the performance using 10-fold cross validation. Choose the number of features after which improvements are minimal;
4) evaluate the performance of the selected number of features on the test set.

The performance graphs and the ranked lists are stored in http://www.let.rug.nl/wieling/grafting-readability/results/

The results of this second set of experiments show that genre-specific readability models reach a much higher accuracy with respect to the generic model of Experiment 1. However, the best results of Experiment 2 are achieved by using a higher number of features (varying across genres but always quite high) with respect to the best results of the Experiment 1 for which a restricted number of features is sufficient to reach the best results (which are however much lower if compared to those of Experiment 2). This demonstrates that there are genre-specific features which help identifying the appropriate readability level within each genre: these features, which vary in number and typology across genres, explain why a general model cannot exploit features which are genre-specific and limits itself to superficial and general complexity features which are valid across genres (i.e. `MediaLunghezzaParole` and `MediaLunghezzaFrasi`, corresponding to average word

length and average sentence length). The results obtained for each genre in terms of accuracy and number of features used are:

- Educational: 83.3% accuracy using 15 features;
- Journalism: 98.33% accuracy using 39 features;
- Literature: 80% accuracy using 37 features;
- Scientific Prose: 81.35% accuracy using 17 features.

TASK 2: ASSESSING SENTENCE READABILITY

Experiment 1

Method: grafting
Training Data: Focus on Journalism: Rep and 2Par readability corpora.
Four different training sets:

1) Rep: GOLD (i.e. manually revised with pruning of easy-to-read sentences); 2Par and Rep of equivalent size
2) Rep: NOGOLD (no correction of sentences: all sentences from Rep are marked as being difficult) but same size as GOLD; 2Par and Rep of equivalent size
3) Rep: NOGOLD big (whole corpus without correction); whole 2Par corpus
4) Rep: NOGOLD balanced (no correction; corpus size same as whole 2Par); whole 2Par corpus

Test Data: Rep: GOLD (i.e. manually revised with pruning of easy-to-read sentences); 2Par and Rep of equivalent size
Steps:

1) obtain the best L1 value using 10-fold cross validation on the training set;
2) use this L1 value with grafting to obtain the ranking of features;
3) iteratively select an increasing number of features and evaluate the performance using 10-fold cross validation. Choose the number of features after which improvements are minimal;
4) evaluate the performance of the selected number of features on the test set.

The performance graphs and the ranked lists are stored in http://www.let.rug.nl/~wieling/grafting-sentread/results/

It turned out that the results for sentence readability are best for the manually corrected small training set (so better than for the small non-gold set, but also both balanced and imbalanced big non-gold sets). The results achieved against the test set show an accuracy of 83.5% using more than 60 features.

Comparing these results with respect to those achieved in the previous task of document readability assessment, it turned out that sentence readability assessment requires a larger number of features to obtain the best accuracy. This demonstrates that assessing sentence readability is a much more complex task requiring an in depth analysis of the text with respect to measuring the readability of a document.

TASK 3: GENRE CLASSIFICATION

Experiment 1

Method: grafting
Training Data: genre corpora (readability levels collapsed together)
Test Data: individual genres (Journalism, Educational, ScientificProse, Literature)

Steps:

1) obtain the best L1 value using 10-fold cross validation on the training set;
2) use this L1 value with grafting to obtain the ranking of features;
3) iteratively select an increasing number of features and evaluate the performance using 10-fold cross validation. Choose the number of features after which improvements are minimal;
4) evaluate the performance of the selected number of features on the test set.

The performance graphs on the basis of the feature count and the ranked lists are stored in http://www.let.rug.nl/~wieling/grafting-genreclassification/

The final results are reported in http://www.let.rug.nl/~wieling/grafting-genreclassification/results/results.png where:

- the red line is based on ten-fold cross validation of the training set to determine what is the optimal number of features (presumably about 15 features, see http://www.let.rug.nl/~wieling/grafting-genreclassification/results/results-increasing-featurecount.txt).
- the blue line shows the test performance.

Experiments have also been performed aimed at investigating whether and to what extent the size of the training datasets influenced the results. It turned out that training with an imbalanced dataset is not a problem: it actually performs slightly better given the increased data. The results obtained with balanced and imbalanced training sets are stored respectively in:

- http://www.let.rug.nl/~wieling/grafting-genreclassification/old/results-balanced/
- http://www.let.rug.nl/~wieling/grafting-genreclassification/old/results-imbalanced/

The results obtained for each genre in terms of accuracy and number of features used are:

- Educational: 62.5% accuracy using 27 features;
- Journalism: 91.66% accuracy using 53 features;
- Literature: 93.33% accuracy using 16 features;
- Scientific Prose: 89.07% accuracy using 20 features.

Il can be noted that the results achieved on this task are comparable with those achieved in Experiment 2 of Task 1 “Document readability assessment” both in terms of accuracy and number of features needed.

TASK 4: NATIVE LANGUAGE IDENTIFICATION

Experiment 1

Differently from the previous tasks, in this case a really high number of features is available (more than 500,000). On the one hand this fact makes it more crucial to be able to reduce the number of features used, but on the other hand this turned out to be a problem from the computational point of view. Below we summarize the work done on this task during the STM period:

- training of all grafting (ranking) models for each of the 11 languages. The L1 value of 1e-6 was selected as this yielded a few thousand features as opposed to more than 100.000 (which would take at least 2 days of running per model);

- the individual models turned out to function well: each individual model distinguishes about 80 — 85% correctly and the increase also appears to level off for higher number of features (but we need at least a few thousand features);
- unfortunately, for an increased number of features (from about 1000 or so) the class probabilities for each individual model are generally either 0 or 1. Consequently, as soon as there are X classifiers classifying the document as being from the corresponding language, all the X probabilities of the models for this document will be 1. Obviously, this is undesirable and reduces performance. Because of the low number of features this does not play a role for the genre identification.

On the basis of what reported above, we think that some sort of dynamic grafting classification model would be needed: instead of the ranking application, a real classifier combined with a dynamic grafting approach. The first N features are selected using grafting per single feature (slow), then the next (say 10*N) features are selected using grafting-light per single feature (faster, but less precise), then the following features are selected using grafting/grafting-light with continuously increasing step sizes (so the number of features which are selected at the same time increases, and this is again faster and less precise). We thus asked Daniël de Kok (of the University of Tübingen) whether he would be interested in collaborating (i.e. programming this) with us for a joint work on this topic. The answer was positive and we currently starting to plan the work to be done in this direction.

TASK 5: DISCOVERING PATTERNS OF LEXICAL VARIATION IN TUSCANY AND UNDERLYING LINGUISTIC FEATURES

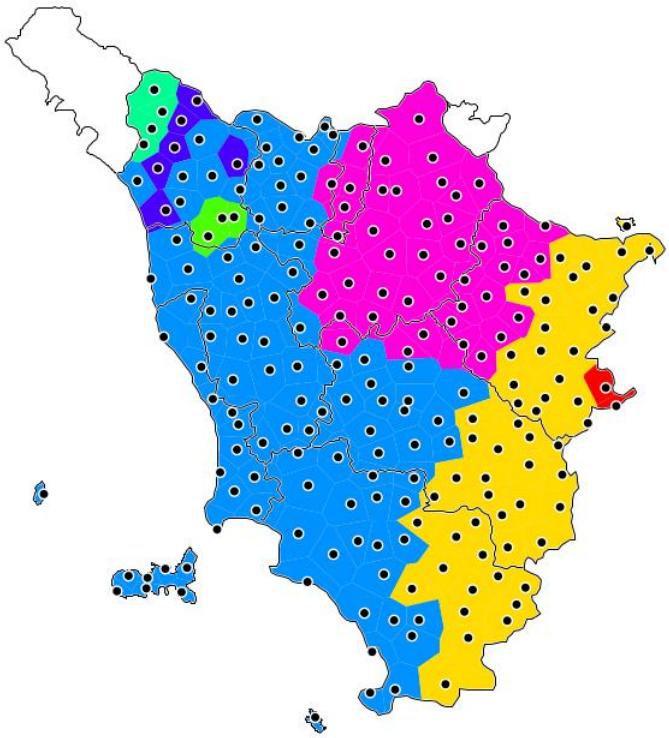
Experiment 1

This line of activity can be seen as a continuation of previous research which started in previous STM research visits by Simonetta Montemagni (Groningen, June 2010) and Martijn Wieling (Pisa, May 2011). The aim consists in the identification of Tuscan dialect areas together with their characteristic lexical variants. The same technique was previously used to identify phonetic areas.

The study focuses on Tuscan dialects only, spoken in 213 out of the 224 investigated locations (Gallo-Italian dialects spoken in Lunigiana and in small areas of the Apennines were excluded). We used the normalized lexical answers to a subset of the ALT onomasiological questions (i.e. those looking for the attested lexicalizations of a given concept). In particular, out of 460 onomasiological questions, we selected only those which prompted 50 or fewer distinct normalized lexical answers (nouns only), for a total of 170 concepts. For each concept, we selected the lexicalizations attested more than 10 times, for a total of 1996 concept-lexicalization types (on average, for each concept we considered the top 12 lexical variants) and 57576 concept-lexicalization tokens.

One novelty of this experiment with respect to previous work is that the importance measure was adapted by multiplying distinctiveness and representativeness instead of averaging: in this way features which are very low in one and high in the other are penalized, which is desirable (such a measure is reminiscent of tf.idf).

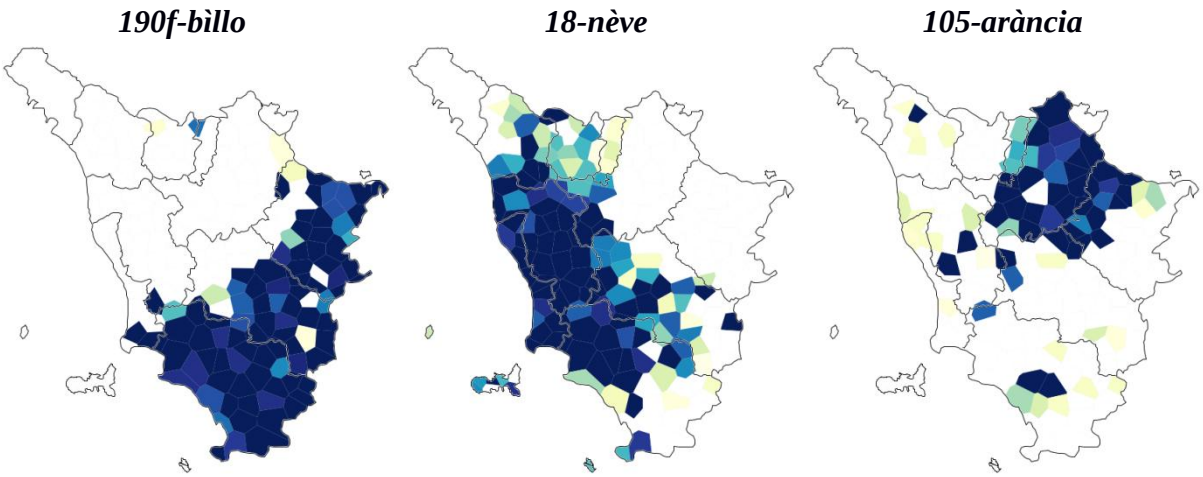
The map below shows the geographic visualization of the clustering of Tuscan varieties into seven groups:



The linguistic features underlying the clusters identified have also been identified and ranked on the basis of the Importance value. We report below the 5 topmost lexical features for the three major clusters:

CLUSTER	CONCEPT-LEXICALIZATION	IMPORTANCE	REPRESENTATIVENESS	DISTINCTIVENESS
Yellow cluster	190f-bìllo	0.354538	0.876520	0.404483
	74-vitàlbia	0.333418	0.461815	0.721973
	167-zìro	0.323925	0.900928	0.359545
	149-rìcci	0.312519	0.448381	0.696994
	211-fringuéllo	0.304078	0.599528	0.507195
Light blue cluster	18-nève	0.428565	0.708102	0.605230
	405-pùppa	0.359588	0.509311	0.706028
	272-pugnétte	0.348224	0.441112	0.789423
	192-pucìno	0.321867	0.448931	0.716964
	228-pùce	0.309313	0.469309	0.659082
Pink cluster	105-arància	0.526111	0.778889	0.675463
	167-órcio	0.407841	0.679380	0.600313
	273P-romaiòlo	0.405753	0.775104	0.523481
	190f-tàcco	0.390295	0.390295	1.000000
	323-capofréddo	0.389089	0.432224	0.900201

In the following maps the distribution of the topmost lexical feature for each of the three major lexical areas is reported, where darker shades of blue indicate higher frequency.



Conclusion

The STM visit of Martijn Wieling was very fruitful, given that the obtained results are interesting and shed light on unexplored areas of readability assessment, genre classification as well as Tuscan lexical variation. In particular, results with respect to readability assessment and genre classification represent innovative contributions with respect to the current literature on the topic; the same holds for the Tuscan results.

We thus planned to write three or four different papers:

1) one paper on sentence-based readability assessment to be submitted to a Workshop or Conference focusing on educational applications or other contexts which may benefit from this type of task;
2) one journal paper about feature ranking and selection for genre-classification and readability;
3) one journal paper on Tuscan dialect clustering with lexical features;
4) depending on the collaboration with Daniël de Kok, a fourth paper could be written about native language identification.