What are the Main Features of the Italian Written by Italian University Students?

What are the main features of the Italian written by Italian University students?

Silvia Dallari[1⋆], Nicola Grandi[2⋆⋆], and Angela Montanari[3]

[1] Department of Statistical Sciences, University of Bologna, Italy
silvia.dallari2@unibo.it
[2] Department of Classical Philology and Italian Studies, University of Bologna, Italy
nicola.grandi@unibo.it
[3] Department of Statistical Sciences, University of Bologna, Italy
angela.montanari@unibo.it

Abstract. This paper aims at contributing to the recent debate on the 'health state' of Italian language, with a specific focus on the Italian written by Italian university students. The topic has been addressed by the recently closed Univers-ITA project (funded in the PRIN 2017 call) from which the data analysed in this paper come. One of the research lines involved asking a sample of university students, from different Italian universities and different learning subjects, to write a formal text composed of at most 500 words. The texts have been annotated by experts and the number of occurrences of specific features related to orthography, lexicon, syntax, coherence and others have been recorded together with socio-demographic characteristics of the respondents. In order to identify groups of writers sharing the same writing features and to simultaneously allow for the correlation between the features themselves, we propose a new mixture model that models the counts as Poisson random variables whose parameters are generated according to a common factor latent variable model. Identifiability conditions are discussed. The model is applied to the project data.

Keywords: Poisson mixture model, latent variable model, EM algorithm

1 Introduction

The 'state of health' of the Italian language has been recently subject to a lively debate, especially concerning the language used by young people. Indeed, language change is usually activated in younger generations and in informal and spoken situations. Formal and written uses are more conservative. The discussion focused on university students, because they represent the more educated and learned layer of the younger generation, so they are expected to be the most conservative component of the most 'innovative' generation. This means

⋆ Corresponding author.
⋆⋆ Univers-ITA project PRIN 2017 PI.

that the linguistic changes attested in the formal language of university students are the ones closest to entering the system permanently. For this reason, possible systematic deviations from the standard observed in the language used by university students are of major importance and deserve a thorough analysis. Univers-ITA [8], a recently closed project funded by the Italian Ministry of Education, has addressed the problem from different perspectives involving questionnaires, analysis of social network data, study of the language used by students in the draft version of their thesis and analysis of purposely written formal texts. The goal was to draw an exhaustive picture of the Italian written by university students, including both formal and informal varieties, integrating the sociolinguistic and the typological perspectives. In particular, in this paper we will analyse the data obtained after annotating the texts, at most 500 word long, written by a sample of 2137 university students from different universities and from different learning subjects. The research objectives are to assess whether and in which ways the occurrence of features specific to the Italian of university students displays significant correlations; which and how many features are connected by an implicational relation, that allows the formulation of predictions on their co-occurrence; if it is possible to single out a deep tendency motivating such co-occurrence. After annotation, seven main features have been identified. In this paper we propose to analyse the data related to the number of times different features have been annotated in each text through a novel mixture model which assumes that inside each mixture component the counts are modelled by Poisson random variables whose parameters are dependent on a set of latent random variables. The data will be briefly described in Section 2, while Section 3 is devoted to model description and identifiability. In Section 4 some empirical results will be presented and discussed.

2 The data

A sample of 2160 university students from different universities has been involved in the study. After cleaning the data, the final sample we have taken into account includes 2137 individuals. The text written by each student has been annotated. Annotations include both actual errors, that is, forms that would be considered incorrect in any type of text, and forms that are not (or not entirely, or no longer) acceptable in a very formal text as the one requested. Specifically, the following features, with the corresponding labels in brackets, have been annotated and the number of occurrences of each of them has been recorded: orthography (nORT), linguistic register (nREG), marked sentences (nMRC), lexicon (nLES), morphosyntax (nMFS), coherence (nCOE) and syntax (nSIN). Other variables related to the socio-economic condition of the respondents have been registered (e.g. gender, socio-economic class, reading habits) together with other information such as the degree program level and subject, the geographical area of the university and the diploma they obtained in high school.

3 Mixtures of GLLVM for multivariate count data

We consider a mixture of generalized linear latent variable models (GLLVM) [2] for count data, where count variables are modelled as Poisson distributed random variables (see [4] for a similar model referred to binary data). This model relates, for each individual, the p-dimensional vector $\mathbf{y}$ of observed count variables with a q-dimensional latent vector $\mathbf{z}$ distributed according to a finite mixture of multivariate Gaussians as follows:

Latent layers

$$f(\mathbf{z}) = \sum_{i=1}^{k} \omega_i \phi_i^{(q)}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

$$f(\mathbf{s}) = \prod_{i=1}^{k} \omega_i^{s_i} \tag{2}$$

Observation layer

$$f(\mathbf{y}|\mathbf{z}) = \prod_{j=1}^{p} f(y_j|\mathbf{z}) = \prod_{j=1}^{p} \frac{\pi_j(\mathbf{z})^{y_j} e^{-\pi_j(\mathbf{z})}}{y_j!} \tag{3}$$

where $\omega_i > 0$ and $\sum_{i=1}^{k} \omega_i = 1$, $\mathbf{s}$ is the allocation variable, $\phi_i^{(q)}$ is the q-variate normal density with $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ as mean and covariance matrix, and the relation between $\mathbf{y}$ and $\mathbf{z}$ can be modelled through the link function, that is:

$$log(\boldsymbol{\pi}(\mathbf{z})) = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\mathbf{z}, \tag{4}$$

where $\boldsymbol{\lambda}_0$ is the vector of dimension $p \times 1$ containing the intercepts, $\boldsymbol{\Lambda}$ is the $p \times q$ factor loading matrix and $\boldsymbol{\pi}(\mathbf{z}) = (\pi_1(\mathbf{z}), ..., \pi_p(\mathbf{z}))$ is the vector of the p Poisson random variable parameters. In order to get unique and consistent estimates of the parameters the model needs to be identified. For this purpose the following restrictions are imposed:

1. $E(\mathbf{z}) = \sum_{i=1}^{k} \omega_i \boldsymbol{\mu}_i = \mathbf{0}$; $\quad Var(\mathbf{z}) = \sum_{i=1}^{k} \omega_i (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T) = \mathbf{I}_q$
2. The upper right triangle of $\boldsymbol{\Lambda}$ is set equal to 0 (following the proposal of [6] of fixing $q(q-1)/2$ loadings equal to zero).
3. $\lambda_{10} = 0$

Model estimation can be performed via an EM algorithm [5]; the integrals involved in the E-step that cannot be directly evaluated are numerically approximated.

4 Real Data Analysis and Conclusions

We applied the model introduced in Section 3 to the data of Section 2, examining different values of q and k. We assume that the maximum number of q can be set at 3 according to the Ledermann's condition [7] referred to the continuous approximation of counts. However, the improvement in the model goodness of fit for $q = 3$ is negligible with respect to the one obtained with $q = 2$, and so we decided to take into account all the possible combinations given by $q = 1, 2$ and $k = 1, 2, 3, 4, 5$. Based on the results in [4], Akaike Information Criterion [1] is used for model selection and the best combination has turned out to be the one with two latent factors and one group. We performed an oblique rotation [2] using the `oblimin` function in the `R` package `GPArotation` [3] to allow for possible correlations among the latent factors. The matrix of factor loadings obtained after performing the rotation is presented in Table 1, while Table 2 shows the correlation matrix of the rotated factors.

Table 1. Loading estimates after oblimin rotation. Only loadings whose absolute value is greater than 0.2 are shown.

	$\hat{\boldsymbol{\lambda}}_1$	$\hat{\boldsymbol{\lambda}}_2$
nCOE		0.53
nLES		0.71
nMFS		0.56
nMRC	0.71	
nORT	0.66	0.41
nREG	1.34	
nSIN		0.45

Table 2. Correlation matrix of the rotated factors.

	$\hat{\boldsymbol{\lambda}}_1$	$\hat{\boldsymbol{\lambda}}_2$
$\hat{\boldsymbol{\lambda}}_1$	1	0.37
$\hat{\boldsymbol{\lambda}}_2$	0.37	1

The high loadings for the variables regarding marked sentences and linguistic register in the first factor suggest that this factor discriminates students who make inappropriate choices with respect to the context. Instead, the second factor is common to lexicon, morphosyntax, coherence and syntax. Thus, this

factor represents difficulties in structuring a complex text as the one requested. Orthography contributes to both factors, since it can be correlated with stylistic aspects summarized by both. Moreover, as can be seen in Table 2, the two factors are correlated.
In summary, these preliminary results suggest the presence of two types of annotation, one linked to inadequate choices and the other one connected to weaknesses in organizing a formal text, but these two are not uncorrelated.
As a further development, some of the covariates mentioned in Section 2 could be included in the mixture weights to allow a better characterization of the individuals and eventually highlight clusters.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723 (1974)
2. Bartholomew, D., Knott, M., Moustaki, I.: Latent Variable Models and Factor Analysis: A Unified Approach, 3rd edn. Wiley, Chichester, UK (2011)
3. Bernaards, C.A., Jennrich, R.I.: Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. Educational and Psychological Measurement **65**(5), 676–696 (2005)
4. Cagnone, S., Viroli, C.: A factor mixture analysis model for multivariate binary data. Statistical Modelling **12**(3), 257–277 (2012)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**(1), 1–38 (1977)
6. Jöreskog, K.G.: A general approach to confirmatory maximum likelihood factor analysis. Psychometrika **34**(2), 183–202 (1969)
7. Ledermann, W.: On the rank of the reduced correlational matrix in multiple-factor analysis. Psychometrika **2**(2), 85–93 (1937)
8. Univers-ITA: https://site.unibo.it/univers-ita/it