CLARIN Annual Conference Proceedings

2021

Edited by

Monica Monachini, Maria Eskevich

27 – 29 September 2021
Virtual Edition

Please cite as:
CLARIN Annual Conference Proceedings, 2021. ISSN 2773-2177 (online).
Eds. M. Monachini and M. Eskevich.
Virtual Edition, 2021.

Programme Committee

Chair:

- Monica Monachini, Institute of Computational Linguistics “A. Zampolli" (IT)

Members:

- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Eva Hajičová, Charles University Prague (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Marinos Ioannides, Cyprus University of Technology (CY)
- Langa Khumalo, North West University (ZA)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Eirikur Rögnvaldsson, University of Iceland (IS)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadič , University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Eva Hajičová, CZ
- Martin Hennelly, ZA
- Erhard Hinrichs, DE
- Marinos Ioannides, CY
- Nicolas Larrousse, FR
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Jan Odijk, NL
- Stelios Piperidis, GR
- Eirikur Rögnvaldsson, IS
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičenonienė, LT
- Tamás Váradi, HU
- Kadri Vider, EE
- Martin Wynne, UK

Subreviewers:

- Federico Boschetti, IT
- Christophe Parisse, FR
- Thorsten Trippel, DE
- Valeria Quochi, IT
- Zijian Győző Yang, HU
- Efstathia Soroli, FR
- Enikő Héja, HU
- Bence Nyéki, HU
- Angelo Mario Del Grosso, IT
- Olivier Baude, FR
- Kinga Jelencsik-Mátyus, HU

CLARIN 2021 submissions, review process and acceptance

- Call for abstracts: 19 January 2021, 1 March 2021
- Submission deadline: 28 April 2021
- In total 40 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 16-17 June 2021
- Notifications to authors: 22 June 2021
- 35 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2021/clarin-annual-conference-2021-virtual-event.

How to Perform Linguistic Analysis of Emotions in a Corpus of Vernacular Semiliterate Speech with the Help of CLARIN Tools

Rosalba Nodari
Università di Siena
`rosalba.nodari@unisi.it`

Luisa Corona
Università dell'Aquila
`luisa.corona@univaq.it`

Abstract

Research has shown that words are constitutive of emotions and that language contributes to shape feelings. However, less is known about how people with basic literacies can use language to maintain, create and recreate affective bonds, and how they express themselves through the language of emotions. In this respect, digital humanities tools can help shed some light on linguistic encoding of emotions. This proposal aims to show the potential of the CLARIN infrastructure tools for carrying out such analysis on a particular corpus of letters written in the 60s' by Michela Margiotta, a semiliterate Italian woman affected by tarantism, to the anthropologist Annabella Rossi. The research will show how corpora of semiliterate letters can pose several problems when conducting research using digital humanities tools. In this respect, different methodologies will be compared in order to verify how CLARIN tools can help in the detection of encoded emotion in written documents.

1 Introduction

Research in linguistics has shown how speakers can manipulate language in order to manifest feelings and evoke emotions in their listeners (Bednarek 2008). Affect in particular is not confined to the private domain of the subjectivity of the individual, but it is usually expressed and manifested through interaction. Emotion itself can be said to be one among several types of stances (Ochs 1996, Dubois 2009, Du Bois and Kärkkäinen 2012), and speakers can contribute to the co-construction of feelings through the use of specific linguistic cues that can index specific values. Lexemes, syntactic structure and encoding strategies can be perceived, processed and linked to peculiar moods. Therefore, linguistic studies and discourse analysis have focused on how emotions and scopes can be expressed though text structure and specific rhetorical moves that may be considered appropriate for discourse types as related to specific textual genre.

Studies devoted to the language of emotion have been conducted using different strategies developed under the umbrella of the digital humanities tools. In particular, lexicon based, rule based or learning based models have explored the role of textual resources such as neighbour words, word frequency, or terms in contexts, as cues for the expression of emotions (Lindquist, Gendron and Satpute 2016).

Research on corpora of written letters can be considered a fruitful field of inquiry. Historically, letters allowed people to maintain personal connections: relationships were thus constructed in the absence of the subjects themselves, with letters allowing the negotiation of affective bonds (Cancian 2010, 2012; Lyons 2013). Crucially, this communicative need was experienced even by people with basic literacy and little familiarity of written practices, and corpora of vernacular written speech can be associated with members of all classes (for Italy see, for example, Caffarena 2005). Letters written by speakers with basic literacies all share a series of linguistic features that have been analysed by linguists (De Mauro 2015 [19701]; Spitzer 1976). However, little is known about how people with basic literacies can use language to maintain, create and recreate affective bonds, and how they express themselves through the language of emotions. In addition, analysis undertaken using corpus linguistics and digital humanities tools is still somewhat underrepresented (but see Vitali 2020). Corpora of semiliterate letters pose several problems for conducting research with digital tools. The transcription of semiliterate letters usually reflects the written habits of the writers, thus maintaining, for example, misspelling, orthographical errors, malapropism, vernacular forms. Furthermore, investigations like sentiment analysis usually require adaptation and first-pass cleaning of the data coming from written text, and this can sometimes alter the very nature of written vernacular forms that typically follow oral communication patterns.

The scope of this paper is, then, to propose to the CLARIN community an interesting case study that can be used to test and implement CLARIN resources. In particular, we want to underline how digital tools can help in tagging and analysing Italian written corpora of semiliterate writers, and how a quantitative approach can help improving knowledge regarding how emotion is encoded in written texts. However, given the nature of the specific corpus that we will use as a case study, we aim at highlighting some of the main problems encountered when conducting this specific type of analysis. CLARIN infrastructure offers, in fact, a wide variety of applications to discover, explore, exploit, annotate, analyse or combine language data. Nevertheless, many of these resources are not available for Italian and, particularly, are not even suitable for the analysis of vernacular forms. When dealing with corpora of semiliterate writers, these problems are especially amplified. We believe that this kind of research suggestion, despite all the difficulties that it poses, can be of interest to the CLARIN infrastructure. If CLARIN can help in offering tools that can be modified in order to be suitable for the analysis of semiliterate speech, researchers will benefit them for future applications.

2 The data under scrutiny: the description of Michela Margiotta corpus

In 1959, the anthropologist Annabella Rossi (1933 - 1984) is invited by Ernesto De Martino to join him and his équipe on some ethnographic fieldworks in Salento (Apulia, Italy), to document and research the phenomenon of tarantism. In this occasion Annabella Rossi meets Michela Margiotta (1898 – 1983) during the feast of San Paul in the sanctuary of Galatina (LE). Margiotta was a woman from the nearby village of Ruffano (LE) who was affected by tarantism and the “male di San Donato”, the latter being typically used in the Salento area to refer to epileptic seizures or even pseudo-seizures of psychogenic nature. Rossi shows deep empathy for Margiotta’s suffering and a curiosity about her story. From this moment onwards, Margiotta spontaneously decides to start a correspondence with Rossi, with the aim to help her in documenting tarantism in Salento, possessions, and other relevant ethnographic facts.

The correspondence between Michela Margiotta and Annabella Rossi spans around 6 years, from 1959 to 1965. In 1970 Rossi decided to collect and publish the letters received from Margiotta in a book entitled *Lettere da una tarantata*, where the woman is anonymised using the pseudonym of Anna.

Following its publication, this volume has been regarded as an important documentation of women's writing in what De Mauro (2015 [1970[1]]) defined as *‘italiano popolare unitario*’ “unitary popular Italian”, a substandard variety of Italian language used by illiterate and semiliterate people. It is also considered a key text in anthropological literature, as it allowed to reflect on the relationship between the observer – the anthropologist – and the observed – the informant, and to highlight the asymmetry of power emerging in field research (see Apolito 2006, 2015).

The epistolary of Michela Margiotta comprises a total of 65 letters: 20 letters were dictated to a more literate person (according to a widespread and well-documented practice among uneducated people), whereas the other 45 were written directly by Margiotta and are listed in the corpus in chronological order. Crucially, despite previous research conducted on several front found no trace of Rossi’s replies to any of these letters. Looking at the epistolary it can be assumed that the anthropologist has, at times, replied to Margiotta. Nevertheless, Rossi's role in this epistolary relationship can only be reconstructed from the references to letters, cards and gifts that are contained in Margiotta's letters. A personal investigation undertaken by the legal heirs of Michela Margiotta has confirmed that none of the letters received by Margiotta were ever found in her house after her death.

In addition to the epistolary corpus, we have been able to identify Margiotta in a long interview contained in *Archivio Sonoro della Puglia* (Apulia Sound Archive), recorded and conducted by Annabella Rossi in 1965. In this occasion, Margiotta talks about various topics related to tarantism and about her own personal experience of “sickness” and possession.

Overall, this archival material allows to have a significant amount of documentation about one single person, that can be of interests for many scholars. In particular, this is one peculiar instance of having access to both the written production and the speech of a semiliterate woman. The anthropological material can guide in the interpretation of her peculiar use of the language. The study of Michela Margiotta can therefore help in laying the foundations to develop a protocol for the analysis of emotional speech.
In her preface to *Lettere da una tarantata*, Annabella Rossi identified several recurring themes in Michela Margiotta's letters and grouped them into six main thematic nuclei. Overall, it can be said that, in her letters, Margiotta talked about different topics, but the whole corpus is characterised by a tension between Margiotta and the anthropologist. It can be argued that for Margiotta the letters were a possibility to construct a bond with a woman that, at least in appearance, was sincerely interested in her personal story, such as family tension, her refusal of the marital status, her phobic relationship with the males. Through these letters, Margiotta expresses herself and tries to change an apparent interviewer-interviewee relationship into something more personal, in order to be recognised as an 'individual' rather than just an informant. For this reason, the letters are characterised by an emotional tension. During these 6 years of writing, Margiotta appears to feel more and more that the relationship is a one-way one. Rossi’s expectation and Margiotta’s goal are deeply different, and Margiotta has to continuously negotiate the absence of Annabella, while at the same time, re-constructing her relationship with Rossi as friends, lovers, or mother and daughter. Margiotta uses the writing practice not only to describe or express her feelings, but also to reach in her letters and through her letters, a shared ground of emotion with Annabella Rossi.

2.1 CLARIN resources and semiliterate speech

The linguistic analysis of Michela Margiotta shall have, as a scope, the detection of linguistic patterns that can be related to the expression of emotion. To do so, researchers have to rely on corpus processing and annotation. However, a number of *caveats* are needed.
Margiotta is, in fact, semiliterate and her written Italian is characterised by misspelling, vernacular forms, and other phenomena commonly attributable to the *italiano popolare*. Since it is precisely in this struggle with a code that Margiotta does not master well that she can express her emotions with, researchers who are interested in her stylistic practice should not normalise her written speech. However, given the above, Margiotta’s written practice doesn’t appear to be suitable for a series of linguistic analysis.

As a first example, automatic basic sentiment analysis tools, such as the one that can be found in Voyant, are not suitable for this corpus because Margiotta vocabulary cannot be compared to the generic database that typically used when conducting sentiment analysis. In this respect, Margiotta semiliterate corpus exacerbates the problems that are faced when dealing with vernacular forms.

Another problem that must be faced regards the different written solutions that can be found in Margiotta letters. As for other semiliterate writers, Margiotta is not consistent with her lexical solutions. Sometimes she tends to write complex utterances as a monorhematic item (as, for example, *teneriandare* < *te ne eri andare* ‘you had to go’), to merge forms (especially articles and prepositions (*laratiolino* < *la ratiolino* ‘transistor radio’, *couna voce* < *co una voce* ‘with a voice’) or to split forms (*sono a rivata* < *sono arrivata* ‘I arrived’). In this regard, the analysis of frequencies and keywords in contexts is particularly problematic. The use of stop words is, indeed, not sufficient because it tends to eliminate parts of the lexemes. Additionally, the statistical analysis of frequencies cannot deal with different written forms, that can be counted for several times.

Finally, the style of Margiotta's letters is characterised by the use of constructions and strategies which are typical of the oral domain. Part-of-speech tagging and dependency parsing tools often encounter and show some difficulties and inconsistencies when performing NLP on speech transcriptions. Again, since the written style of Margiotta is entirely discursive, we expect to encounter similar issues.

3 Conclusion

Over the past few years, CLARIN infrastructure has helped many researchers in dealing with non-conventional objects that posed unexpected issues in linguistic analysis. For example, corpora of vernacular speech such as the Gra.Fo project (Calamai and Bertinetto 2014) have been a useful testing grounds for automatic speech recognition and segmentation, or for metadata annotation (Calamai and Frontini 2016). Therefore, with the Michela Margiotta case study, we intend to present to the CLARIN community with an example of something that includes issues of both written and vernacular speech corpora. Currently, the usage of digital humanities tools can present limits that often outweigh the pros. In this respect, we offer the Margiotta written digitised corpus as a case study that can help in developing such tools. Michela Margiotta corpus can be seen as an almost perfect opportunity to develop tailored tools for linguistic analysis. Given that Margiotta corpus comprises several documents from a single writer collected over time, it avoids additional complications such as regional variation or speakers' idiosyncrasies, which is then likely to simplify the tools developing process. This case study will therefore allow to develop some basic guidelines for tools tailoring which can then be used to analyse other illiterate corpora of vernacular speakers.

References

Apolito, P. 2006. *Con la voce di un altro. Storia di possessione, di parole e di violenza.* L'Ancora, Napoli.

Apolito, P. 2015. E sono rimasta come lisolo a mezzo a mare. In Rossi, A., *Lettere da una tarantata*. Squilibri, Roma: 13-62.

Bednarek, M. 2008. Analyzing language and emotion. In Bednarek, M. (ed.), *Emotion talk across corpora*. Palgrave Macmillan, Basingstoke: 1-26.

Caffarena, F. 2005. *Lettere dalla Grande Guerra. Scritture del quotidiano, monumenti della memoria, fonti per la storia. Il caso italiano.* Unicopli, Milano.

Calamai, S. and Bertinetto, P. M. 2014. *Le Soffitte Della Voce. Il Progetto Grammo-Foni*. Vecchiarelli Editore, Manziana.

Calamai, S. and Frontini, F. 2016. Not quite your usual kind of resource. Gra.fo and the documentation of Oral Achives in CLARIN. In *Proceedings of the CLARIN Annual Conference 2016*, Aix-en-Provence. https://hal.archives-ouvertes.fr/hal-01395027/document

Cancian, S. 2010. *Families, lovers, and their letters: Italian postwar migration to Canada*. University of Mannitoba Press, Winnipeg.

Cancian, S. 2012. The language of gender in lovers' correspondence, 1946-1949. *Gender and History*, 24: 755-765.

De Mauro, T. 2015 [1970^1]. Per lo studio dell'italiano popolare unitario. In Rossi, A., *Lettere da una tarantata*. Squilibri, Roma: 63-82.

Du Bois, J. W. 2009. Interior dialogues: The co-voicing of ritual in solitude. In Senft G. and Basso E. B. (eds.), *Ritual communication*. Berg, Oxford: 317-340.

Du Bois, J. W., and Kärkkäinen, E. (2012). Taking a stance on emotion: Affect, sequence, and intersubjectivity in dialogic interaction. *Text & Talk*, 32(4): 433-451.

Lindquist, K. A., Gendron, M. and Satpute, A. B. 2016. Language and emotion: Putting words into feelings and feelings into words. In Barrett, L. F., Lewis, M. and Haviland-Jones, J. M. (eds.), *Handbook of emotions*. 4th edition. Guilford Press, New York: 579-594.

Lyons, M. 2013. *The writing culture of ordinary people in Europe, c. 1860-1920*. Cambridge University Press, New York.

Ochs, E. 1996. Linguistic resources for socializing humanity. In Gumperz, J. and Levinson, S. (eds.), *Rethinking linguistic relativity.* Cambridge University Press, New York: 407-437.

Rossi, A. 2015 [1970^1]. *Lettere da una tarantata*. Squilibri, Roma.

Spitzer, L. 1976. *Lettere di prigionieri di guerra italiani, 1915-1918*. Bollati Boringhieri, Torino.

Vitali, G. P. 2020. What is a last letter? A linguistic/preventive analysis of prisoner letters from the two World Wars. In Marras, C., Passarotti, M., Franzini, G. and Litta, E. (a cura di), *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica. Quaderni di umanistica digitale:* 265-272.