When the field moves online: a three-step approach to collect and analyze crowdsourced dialect data

Federica Breimaier

RoLinC, 16.11.2021
Table of Contents

– Introduction:
  ➔ what do we mean with *crowdsourcing*?
  ➔ how has it been implemented in linguistic research?
– The three-step approach:
  1. Structuring the questionnaire
  2. Collecting the data
  3. Analyzing the results
     1. Inspection of the sample
     2. Exploratory statistics
     3. Inferential statistics
– Conclusion
– Bibliography
– Discussion
Introduction
What do we mean with *crowdsourcing*?

“Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call” (Howe 2006:1)

“Outsourcing to the crowd”  
(Schenk & Guittard 2011: 94)

(Hartshorne et al. 2018 ➔ 669.498 participants)
Crowdsourcing and dialectology

• Phonetics and phonology
  - Leeman et al. 2018
  - Zihlmann & Leemann 2018
  - Leemann 2020

• Syntax:
  - Microcontact Project: https://microcontact.sites.uu.nl/atlas/

• Semantics and lexicography:
  - Verba Alpina Project: https://www.verba-alpina.gwi.uni-muenchen.de/en/?page_id=162&db=211

• Pragmatics:
  - Bischetti et al. 2020

“We present a compendium of [...] projects that utilize crowdsourcing technologies for language studies, finding that the quality is comparable to controlled laboratory experiments, and in some cases superior” (Munro et al. 2010: 122)
The testing ground: the persistence of the neuter in two Central-Southern Italo-romance varieties

- “Genders are classes of nouns reflected in the behavior of associated words” (Hockett 1958: 231, quoted in Corbett 1991: 1)

- Agreement: “systematic covariance between a semantic or formal property of one element and a formal property of another” (Steele 1978: 610)

- “We should [...] differentiate controller genders, the genders into which nouns are divided, from target genders, the genders which are marked on adjectives, verbs and so on” (Corbett 1991: 151)
Macerata & Molfetta: the one-source problem

### Molfettese

<table>
<thead>
<tr>
<th>SINGULAR</th>
<th>PLURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. N. rə ttuɛskə</td>
<td>Ø</td>
</tr>
<tr>
<td>b. M. u nəpaɛtə le nəpaʊtə</td>
<td></td>
</tr>
<tr>
<td>c. A. u nɛuətə re nnóderə</td>
<td></td>
</tr>
<tr>
<td>d. F. la ssedde re ssiɛddə</td>
<td></td>
</tr>
</tbody>
</table>

Molfetta (Merlo 1917)

- Breimaier 2021: Molfetta_1
- Loporcaro, Breimaier, Manzari 2021 (on the field)

### Maceratese

<table>
<thead>
<tr>
<th>SINGULAR</th>
<th>PLURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. N. lo vorro</td>
<td>Ø</td>
</tr>
<tr>
<td>b. M. lu dènde li déndi</td>
<td></td>
</tr>
<tr>
<td>c. A. lu muru le mure</td>
<td></td>
</tr>
<tr>
<td>d. F. la mà le mà</td>
<td></td>
</tr>
</tbody>
</table>

(Paciaroni 2017: 206)

- Breimaier (in prep.): Macerata_1
- Breimaier & Zanini (in prep.): Macerata_2

---

"when there is a single source of information or documentation for a so-called ‘exceptional’, small, isolated linguistic community" (Leivada et al. 2019: 9)
The three-step approach:

1. Structuring the questionnaire
The questionnaire

- Social variables
  - age
  - gender
  - place of birth
  - how long participants have been living in the city
  - level of education
  - frequency of use of the dialect
    - daily
    - weekly
    - monthly
    - never
The questionnaire

- Social variables
  - age
  - gender
  - place of birth
  - how long participants have been living in the city
  - level of education
  - frequency of use of the dialect
    - daily
    - weekly
    - monthly
    - never
The questionnaire

- Social variables
  - age
  - gender
  - place of birth
  - how long participants have been living in the city
  - level of education
  - frequency of use of the dialect
    - daily
    - weekly
    - monthly
    - never
  - place of living
  - linguistic background of participants’ parents
  - dialect spoken
The questionnaire

– Social variables
  – age
  – gender
  – place of birth
  – how long participants have been living in the city
  – level of education
  – frequency of use of the dialect
    – daily
    – weekly
    – monthly
    – never
  – place of living
  – linguistic background
  – dialect spoken
Online tools

- Google Forms
  - [https://docs.google.com/forms](https://docs.google.com/forms)
  - Molfetta_1, Macerata_1

- Question Pro
  - [https://www.questionpro.com](https://www.questionpro.com)
  - Macerata_2

- Lime Survey
  - [https://www.limesurvey.org](https://www.limesurvey.org)

L'indagine non è attiva. Impossibile salvare le risposte.

* Come diresti in molfettese quello che vedi nella foto?
  
  !

* Da 1 a 7, quanto è stato difficile riconoscere la parola rappresentata nella foto?

Facilissimo

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Difficilissimo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Salvare i dati inseriti
The three-step approach:
1. Structuring the questionnaire
2. Collecting the data
Reaching the crowd

- Sampling strategies (Dörnyei 2007: 95-8)
  - **Probability sampling**: “it needs to be reiterated that in most […] linguistic research it is unrealistic or simply non feasible to aim for perfect representativeness in the psychometric sense”
  - **Non-probability sampling**
    - Quota and dimensional sampling
    - Snowball sampling
    - **Convenience sampling** participation depends on practical criteria like geographical proximity, availability, accessibility and willingness to volunteer.

- Crowdsourcing platforms:
  - On-site cultural associations
  - Local newspapers
  - Facebook (groups and pages)
  - Twitter
  - Instagram

---

https://www.molfettaviva.it/notizie/da-zurigo-un-questionario-per-studiare-il-dialetto-di-molfetta/

Da Zurigo un questionario per studiare il dialetto di Molfetta
Un gruppo di ricercatori svizzeri chiede l'aiuto dei molfettesi per capire lo stato del nostro dialetto. Ecco il link del questionario

MOLFETTA - GIOVEDI 28 GIUGNO 2018 ① 11.25

https://www.cronachemaceratesi.it/2021/03/03/il-dialetto-maceratese-e-una-lingua-viva-lo-parlano-tutte-le-generazioni/1501435/

«Il dialetto maceratese è una lingua viva, lo parlano tutte le generazioni»

LA LINGUISTA svizzera Federica Breimaier rende noti i risultati dello studio effettuato nel 2019 e propone un nuovo questionario ai lettori di Cronache Maceratesi

https://www.cronachemaceratesi.it/2021/03/03/il-dialetto-maceratese-e-una-lingua-viva-lo-parlano-tutte-le-generazioni/1501435/
The three-step approach:

1. Structuring the questionnaire
2. Collecting the data
3. Analyzing the results
   1. The sample
   2. Exploratory statistics
   3. Inferential statistics
The sociolinguistic profile of the participants

- Validity => internet as a fertile ground for false information?
  - Gosling et al. 2004
- Generalizability => an inherently biased sample (Hewson et al. 2003)?
  - well-educated
  - technologically
  - young
  - men

“Because of the compromised nature of non-probability sampling we need to describe in sufficient detail the limitations of such samples when we report the results” (Dörnyei 2007: 99)
Fig. 2: General demographics (Macerata_1)

2.a Age and sex

Age groups (in years)

Proportions (in %)

2.b Education

Educational level (in years)

Proportions (in %)

Sex: Men, Women

November 2019, tot. = 57%

Fig. 3: General demographics (Macerata_2)

3.a Age and sex

Age groups (in years)

Proportions (in %)

3.b Education

Educational level (in years)

Proportions (in %)

Sex: Men, Women

April 2019, tot. = 111
Frequency of use of the dialect across the three samples

Fig. 4: Frequency of use of the dialect (Molfetta)

Fig. 5: Frequency of use of the dialect (Macerata_1)

Fig. 6: Frequency of use of the dialect (Macerata_2)
The three-step approach:

1. Structuring the questionnaire
2. Collecting the data
3. Analyzing the results
   1. The sample
   2. Exploratory statistics
   3. Inferential statistics
Exploratory statistics (Molfetta)

Fig. 7: Article-Noun Agreement (Molfetta)

7.a Nouns categorized as neuter in Merlo (1917)

<table>
<thead>
<tr>
<th>Noun</th>
<th>Proportions (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Bread'</td>
<td>451</td>
</tr>
<tr>
<td>'Milk'</td>
<td>440</td>
</tr>
<tr>
<td>'Wheat'</td>
<td>295</td>
</tr>
<tr>
<td>'Pork fat'</td>
<td>362</td>
</tr>
<tr>
<td>'Rice'</td>
<td>441</td>
</tr>
</tbody>
</table>

Selected form of the article
- N. (re±, la±)
- M.SG (u)
- F.SG (la)

7.b Nouns categorized as masculine in Merlo (1917)

<table>
<thead>
<tr>
<th>Noun</th>
<th>Proportions (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Apron'</td>
<td>281</td>
</tr>
<tr>
<td>'Snot'</td>
<td>359</td>
</tr>
<tr>
<td>'Apple peel'</td>
<td>324</td>
</tr>
</tbody>
</table>

June 2018
Exploratory statistics (Molfetta)

**Fig. 7: Article-Noun Agreement (Molfetta)**

7.a Nouns categorized as neuter in Merlo (1917)

- **'Bread'**  
  - N. (ra±, la±)  
  - M.SG (u)  
  - F.SG (la)  
  - Proportions (in %)

- **'Milk'**  
  - N. (ra±, la±)  
  - M.SG (u)  
  - F.SG (la)  

- **'Wheat'**  
  - N. (ra±, la±)  
  - M.SG (u)  
  - F.SG (la)  

- **'Pork fat'**  
  - N. (ra±, la±)  
  - M.SG (u)  
  - F.SG (la)  

- **'Rice'**  
  - N. (ra±, la±)  
  - M.SG (u)  
  - F.SG (la)  

- Selected forms of the article  
  - N. (ra±, la±)  
  - M.SG (u)  
  - F.SG (la)

7.b Nouns categorized as masculine in Merlo (1917)

- **'Apron'**  
  - N. (ra±, la±)  
  - M.SG (u)  
  - F.SG (la)  

- **'Snot'**  
  - N. (ra±, la±)  
  - M.SG (u)  
  - F.SG (la)  

- **'Apple peel'**  
  - N. (ra±, la±)  
  - M.SG (u)  
  - F.SG (la)

Selected forms of the article  
- N. (ra±, la±)  
- M.SG (u)  
- F.SG (la)

June 2018
Exploratory statistics (Molfetta)

Fig. 7: Article-Noun Agreement (Molfetta)

7.a Nouns categorized as neuter in Merlo (1917)

- 'Bread': n=451
- 'Milk': n=440
- 'Wheat': n=295
- 'Pork fat': n=362
- 'Rice': n=441

Proportions (in %)

Selected form of the article
- N. (re±, la±)
- M. SG (u)
- F. SG (la)

7.b Nouns categorized as masculine in Merlo (1917)

- 'Apron': n=281
- 'Snot': n=359
- 'Apple peel': n=324

Proportions (in %)
Exploratory statistics (Molfetta)

Fig. 7: Article-Noun Agreement (Molfetta)

7.a Nouns categorized as neuter in Merlo (1917)

7.b Nouns categorized as masculine in Merlo (1917)

Selected form of the article
- N. (ra±, la±)
- M.SG (u)
- F.SG (la)

Proportions (in %)

<table>
<thead>
<tr>
<th>Noun</th>
<th>Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Bread'</td>
<td>N=451</td>
</tr>
<tr>
<td>'Milk'</td>
<td>N=440</td>
</tr>
<tr>
<td>'Wheat'</td>
<td>N=285</td>
</tr>
<tr>
<td>'Pork fat'</td>
<td>N=362</td>
</tr>
<tr>
<td>'Rice'</td>
<td>N=441</td>
</tr>
</tbody>
</table>

(3) 'Rice'
- [ɾə̃ gɡrenɔˈraisə] (a)
- [u ˈraisə] (b)

June 2018
Exploratory statistics (Molfetta)

Fig. 7: Article-Noun Agreement (Molfetta)

7.a Nouns categorized as neuter in Merlo (1917)

- 'Bread' n=451
- 'Milk' n=440
- 'Wheat' n=295
- 'Pork fat' n=362
- 'Rice' n=441

7.b Nouns categorized as masculine in Merlo (1917)

- 'Apron' n=281
- 'Snot' n=359
- 'Apple peel' n=324

June 2018
Exploratory statistics (Molfetta)

Fig. 7: Article-Noun Agreement (Molfetta)

7.a Nouns categorized as neuter in Merlo (1917)

- 'Bread': n=451
- 'Milk': n=440
- 'Wheat': n=295
- 'Pork fat': n=362
- 'Rice': n=441

Selected form of the article:
- N. (re±, la±)
- M.SG (u)
- F.SG (la)

7.b Nouns categorized as masculine in Merlo (1917)

- 'Apron': n=281
- 'Snot': n=359
- 'Apple peel': n=324

June 2018
Exploratory statistics (Molfetta)

Fig. 7: Article-Noun Agreement (Molfetta)

7.a Nouns categorized as neuter in Merlo (1917)

7.b Nouns categorized as masculine in Merlo (1917)

Selected form of the article
- N. (ra±, la±)
- M.SG (u)
- F.SG (la)

Proportions (in %)

(4) ‘Pork fat’
   a. [ɾəˈgrəsə]
   b. [uˈgrəsə]
Exploratory statistics (Molfetta)

Fig. 7: Article-Noun Agreement (Molfetta)

7.a Nouns categorized as neuter in Merlo (1917)

- 'Bread': 451
- 'Milk': 440
- 'Wheat': 295
- 'Pork fat': 362
- 'Rice': 441

7.b Nouns categorized as masculine in Merlo (1917)

- 'Apron': 281
- 'Snot': 359
- 'Apple peel': 324

Selected form of the article:
- N. (re±, la±)
- M.SG (u)
- F.SG (la)

June 2018
Exploratory statistics (Macerata_1) 1/2

Fig. 8: Article-noun agreement (Macerata_1)

8.a Neuter Nouns (according to Paciaroni 2017)

- 'Bread' (n=506)
- 'Honey' (n=542)
- 'Wine' (n=560)

8.b Masculine Nouns (according to Paciaroni 2017)

- 'Ham' (n=543)
- 'Chilli pepper' (n=510)

Selected form of the article
- N. (lo)
- M.SG (lu)

November 2019
Fig. 9 Article-noun agreement with 'chilli pepper'

<table>
<thead>
<tr>
<th>Age groups</th>
<th>Proportions (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-20</td>
<td>N. (lo)</td>
</tr>
<tr>
<td>21-40</td>
<td>N. (lo)</td>
</tr>
<tr>
<td>41-60</td>
<td>N. (lo)</td>
</tr>
<tr>
<td>61-80</td>
<td>N. (lo)</td>
</tr>
</tbody>
</table>

Selected form of the article
- N. (lo)
- M.SG (lu)
Exploratory statistics (Macerata_2)

Fig. 11: Acceptability ratings across experimental conditions (Macerata_2)

- **Exploratory statistics (Macerata_2)**
  - **Mean acceptability ratings:**
    - **44%**
    - **35%**

- **Proportions (in %):**
  - **FF** (the chair) 'la séija' [n=1110]
  - **MM** (the time) 'lu tembu' [n=2220]
  - **NN** (the butter) 'lu vurro' [n=2220]
  - **MN** (the milk) 'lo latte' [n=2220]
  - **NM** (the tooth) 'lo dênde' [n=2220]
  - **MF** (the ash) 'lu cénnera' [n=555]
  - **NF** (the snow) 'lo séija' [n=555]

- **Acceptability ratings: (1= Totally unacceptable, 2= Unacceptable, 3= Neither unacceptable nor acceptable, 4= Acceptable, 5= Perfectly acceptable)**

- **Experimental condition:** (article - noun)
The three-step approach:

1. Structuring the questionnaire
2. Collecting the data
3. Analyzing the results
   1. The sample
   2. Exploratory statistics
   3. Inferential statistics
Some basic terminology…

- dependent (outcome/ response) variable ≠ independent (explanatory) variable

- dependent variables:
  - numerical (Filipponio & Garassino 2019)
  - categorical
    - binomial ➔ Molfetta_1, Macerata_1
    - multinomial (ordered, e.g. Likert-type ratings) ➔ Macerata_2
      (Endresen & Janda 2016)

… on Random forests and Inference trees
(Tagliamonte & Baayen 2012, Levshina 2015)

- non-parametric modeling
- (Klavan et al. 2015, Hundt et al. 2018, Hundt 2020, Giudici & Zanini 2021)
- unbalanced sample
- all kinds of outcome
- relatively small number of observations ~ numerous explanatory variables
Random forest & Inference tree for Molfetta_1

Fig. 12: Conditional importance of each explanatory variable
Fig. 13: Inference tree modeling the distribution of the two forms of the definite article for four neuter nouns.
Fig. 13: Inference tree modeling the distribution of the two forms of the definite article for four neuter nouns.
Random forest & Inference tree for Molfetta_1

Fig. 13: Inference tree modeling the distribution of the two forms of the definite article for four neuter nouns
Fig. 14: Conditional importance of each explanatory variable

- Lexeme
- Age
- Place_of_Living
- gen_imput
- freq_use
- Time_spent_Macerata
- sesso
- scolarità
- place_birth
Fig. 15: Inference tree modeling the distribution of the two forms of the definite article
Inference tree & random forests (Macerata_1)

Fig. 15: Inference tree modeling the distribution of the two forms of the definite article
Inference tree & random forests (Macerata_1)

Fig. 15: Inference tree modeling the distribution of the two forms of the definite article
Inference tree & random forests (Macerata_1)

Fig. 15: Inference tree modeling the distribution of the two forms of the definite article
Fig. 15: Inference tree modeling the distribution of the two forms of the definite article
Random forest (Macerata_2)

Fig. 16: Conditional importance of each explanatory variable

- Article
- Noun_gender
- Age
- Dialect_freq.
- Parents_background
- Time_in_Macerata
- Gender_b
- Noun_numerability
- Noun_transparency

Conditional importance of variables
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

'lař séjja' (the chair)
'lař cénnera' (the ash)
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

'la₃ séjja₃' (the chair)
'la₃ cénnera₃' (the ash)
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

‘la séjja’ (the chair)
‘la cénnera’ (the ash)
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

'la séjja\-' (the chair)
'la cénnera\-' (the ash)

'lou séjja\-' (the chair)
'lou cénnera\-' (the ash)
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

'lař séjja' (the chair)
'lař cëñnera' (the ash)

'loř N séjja' (the chair)
'luµ M cëñnera' (the ash)
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

'la séjja' (the chair)
'la cénnera' (the ash)
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

- ‘la-r séjj–ja’ (the chair)
- ‘la-r cénner–a’ (the ash)

- ‘lo-n séjj–ja’ (the chair)
- ‘lu-n cénner–a’ (the ash)

- ‘lu–w tembu–w’ (the time)
- ‘lu–w dènde–w’ (the tooth)
Inference tree (Macerata_2)

Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

'la₉ séjjja₂' (the chair)
'la₉ cénnera₂' (the ash)

'lo₉ séjjja₉' (the chair)
'lu₉ cénnera₉' (the ash)

'lu₉ tembu₉' (the time)
'lu₉ dènde₉' (the tooth)

'lo₉ vurro₉' (the butter)
'lo₉ latte₉' (the milk)
Inference tree (Macerata_2)

Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

'vela séjja' (the chair)
'vela cénnera' (the ash)

'lo N séjja' (the chair)
'luN cénnera' (the ash)
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

'laₜ séjjaₜ' (the chair)
'laₜ cénneraₜ' (the ash)

'loₜ séjjaₜ' (the chair)
'luₜ cénneraₜ' (the ash)
Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

- 'la[ē] séjja-' (the chair)
- 'la[ē] cénnera-' (the ash)

- 'lo[ē] séjja-' (the chair)
- 'lu[ē] cénnera-' (the ash)

- 'lu[ē] tembu-' (the time)
- 'lu[ē] vurro-' (the butter)
- 'lo[ē] vurro-' (the butter)
- 'lu[ē] dènde-' (the tooth)
- 'lu[ē] latte-' (the milk)
- 'lo[ē] latte-' (the milk)
Inference tree (Macerata_2)

Fig. 17: Inference tree modeling the distribution of the acceptability ratings across conditions

'la-F-séjja-F' (the chair)
'la-F-cénnera-F' (the ash)

'lo-N-séjja-N' (the chair)
'lu-M-cénnera-N' (the ash)

'lu-M-tembu-N' (the time)
'lu-M-dènde-N' (the tooth)

'lo-N-vurro-N' (the butter)
'lo-N-latte-N' (the milk)

'lu-M-vurro-N' (the butter)
'lu-M-latte-N' (the milk)
To sum up…

- Crowdsourcing to study morpho-syntactic variation
  - low cost
  - fast
  - quantitative data
- exploratory vs. confirmatory studies
- importance to adapt the questionnaires to the specific profile of the dialect
- before jumping to statistical modeling:
  - inspection of the sample
  - exploratory statistics
- Random forests and inference trees


Hundt, M. (2020). It is important that mandatives (should) be studied across different World Englishes and from a construction grammar perspective. In P. P. Núñez, C. M. J. López, N. B. Méndez, & M. I. Palacios (Eds.), Crossing linguistic boundaries: systemic, synchronic and diachronic variation in English (pp. 211–238). Bloomsbury Academic. http://dx.doi.org/10.5040/9781350053885.ch-010

References (2/2)


Munro, R., Bethard, S., Kuperman, V., Lai, V.T., Melnick, R., Potts, C., Schnoebelen, T., & Tily, H.J. (2010). Crowdsourcing and language studies: the new generation of linguistic data. Mturk@HLT-NAACL.


The VerbaAlpina-Project. Retrieved November 15, 2021, from https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133&db=181

Thank you so much for your attention! Comments, questions, observations are more than welcome