# A Crowdsourcing Platform for Italian Linguistic Field Research

François Bry
Institute for Informatics
Ludwig-Maximilian University
Munich, Germany
bry@lmu.de

Fabian Kneissl
Institute for Informatics
Ludwig-Maximilian University
Munich, Germany
fabian.kneissl@ifi.lmu.de

Thomas Krefeld
Institute for Romanic Philology
Ludwig-Maximilian University
Munich, Germany
thomas.krefeld@lmu.de

Stephan Lücke
IT Group Humanities
Ludwig-Maximilian University
Munich, Germany
luecke@lmu.de

Christoph Wieser
Institute for Informatics
Ludwig-Maximilian University
Munich, Germany
christoph.wieser@ifi.lmu.de

## ABSTRACT

Linguistic field research depends on collecting phrases and sentences as well as their geographical and social characteristics. Collecting such data is usually done by sending researchers in the field to ask questions and fill forms. This traditional field research is time-consuming, costly, and not free of biases. This demonstration paper presents *metropolitalia*, a Web-based crowdsourcing platform for linguistic field research aiming at overcoming some of the drawbacks of traditional linguistic field research. metropolitalia is built upon *Agora*, a market for trading with phrases and speculating on their characteristics (such as geographical spread and gender, age, and level of education of speakers) in a playful manner. *Borsa Parole*, a first game built upon Agora and presented here, incites players to express their own knowledge or, rather, beliefs and aims at gathering data for language studies. This paper describes Agora and Borsa Parole itself, reports on first evaluations of the data gathered so far, and shows a demonstration of Borsa Parole's use.

## Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; J.5 [**Arts and Humanities**]: Linguistics

## General Terms

Languages; Algorithms

## Keywords

Crowdsourcing, Games With A Purpose (GWAP), Linguistic Field Research



## 1. INTRODUCTION

Linguistic field research is concerned with gathering and analyzing speech data from speakers of some language(s) under observation. The data gathered comprise the speech data itself as well as characteristics of the speakers such as their geographical location and social characteristics, like age, gender, or level of education. Traditionally, such multidimensional data are collected by sending scientists, typically doctoral students or other low paid researchers, to the speakers' locations, usually in certain geographical regions, where they interview speakers, record and/or transliterate the interview, and report on these interviews by filling forms. This process is time-consuming because each researcher can only interview a limited number of speakers, costly because the researchers or students involved have to be paid, and furthermore can be biased because of (conscious or unconscious) preconceptions an interviewer might have [1]. As a consequence, only relatively limited areas can be covered by traditional linguistic field research.

The crowdsourcing platform *metropolitalia* –accessible at `http://www.metropolitalia.org` since August 2012– is conceived as a Web-based platform for linguistic field research. It encourages people to participate in the process of gathering a large linguistic dataset from a wide geographical area with low costs for the linguists. Such a participation of many users to reach certain goals –that are not necessarily known to the users– is called crowdsourcing, a current trend on the Web which provides a cost- and time-efficient way of gathering data [2]. One way to gather data using crowdsourcing is by employing games known as "games with a purpose" ("GWAP") [9], which is the approach we describe in this paper.

We designed the market-based system *Agora* (Greek for "market") for data gathering. On games based on Agora symbolic goods can be traded and speculated with. People can submit symbolic goods –like dialect phrases– together with their own assessment of characteristics of that symbolic good –where or within which social group the dialect phrase is used– and compare their own assessments with those of the community. Thus, one can speculate in both senses of forming conjectures and investing money with a symbolic good and its characteristics. One then receives a payment

in form of points, which can be seen as play-money or as tokens of expertise, when the community agrees. Agora is used as operating system of a first game called *Borsa Parole*, Italian for "word stock exchange". On Borsa Parole, the better phrases and their characteristics are recognized by the user community, the more successful is a user expressing the same belief. A second game aiming at gathering complementary data, also based on Agora, is under development.

The Italian language is especially interesting for linguistic field research, making Borsa Parole excellent means for investigating how to perform linguistic field research via crowdsourcing. Indeed, the Italian language spoken today everywhere, in cities and countryside alike, and within all social groups is currently undergoing a divergence that originates in the big cities and spreads from there [5]. This makes today's Italian different from languages such as German, English, or French. The manifold vernaculars –that is, unstandardized language varieties– differ from dialects – that is, languages socially or geographically subordinate to a (national or regional) standard language– and from each other in vocabulary, grammar, and/or pronunciation. In all cultures there is a considerable interest in language issues and in reflecting on one's own language variations. People interested in their own language are likely to participate in Borsa Parole just for seeing what others disclose on the platform, both phrases they do not know and assessments they are not aware of.

### *Related Work.*

Crowdsourcing [2] is applied in many different contexts, like the collaborative web platform Wikipedia or games solving image labeling tasks. Similar to crowdsourcing, human computation refers to applications in which humans consciously or unconsciously collaborate to solve problems that so far can not be solved purely algorithmically [6]. If a game is designed such that users solve this problem while playing the game, the application is called a GWAP [9]. Several GWAP have been designed that solve different problems, the first one being image labeling [9]. Also in linguistics, crowdsourcing has already been applied successfully, mainly in theoretical linguistics. Often Amazon Mechanical Turk is employed for human computation, where users are paid for completing small tasks [7]. An important conclusion of [7] is that the linguistic quality achieved using human computation is comparable to that of controlled laboratory studies. Further articles report on using GWAP for gathering corpora annotations [3, 8].

Prediction markets are employed for estimating what the results of unknown future events are. Here, users trade contracts whose payoff depends on such events [10]. In an efficient market, the price of such a contract directly correlates with the probability of the future event. Prediction markets are supposed to be efficient markets [10] and therefore can quite closely predict future events.

To the best authors' knowledge, no other crowdsourcing using games than that built upon Agora have been proposed so far that rely on a market for gathering data for linguistic field research.

### *Contributions.*

This paper demonstrates how linguistic field research can be performed by Web-based crowdsourcing. Agora accounts for this need by providing the exploitation systems for a first game for gathering quantitative and manifold data.

The contributions of this paper and of the associated demonstration are as follows:

- Presentation of the market-like operating system Agora;
- Presentation of the game Borsa Parole, run by Agora, aiming at gathering linguistic data and meta-data;
- First evaluation of data gathered so far with Borsa Parole.

## 2. AGORA: A MARKET FOR GATHERING DATA

Agora is a generic software for running Web-based play-markets in which a community of users can share symbolic goods as well as assessments of characteristics of these symbolic goods. A symbolic good can be a text (as in metropoli-italia), an image, an audio file, or any other immaterial good (or combination thereof) that needs to be characterized by users. The good is symbolic in the sense that it can occur on Agora multiple times, be possessed by multiple users, and – technically– be transferable over the Internet. Agora makes it possible for a user to:

- add her own symbolic goods to the market,
- propose assessments for her own symbolic goods as well as for symbolic goods proposed by others,
- review and adapt her own assessments based on assessments from other users, and
- trade assessments with other users.

An assessment consists of a user assessing one or more characteristics of a symbolic good and additionally estimating which proportion of users are likely to assign the same characteristics as she does. All assessments for a symbolic good together represent the market's view for the symbolic good and if a user agrees with the aggregated view of the market, she gains (play-)money. The closer her estimation is to the proportion of users assigning the same characteristics, the more money she gains. Assessments can be offered for sale for a user-defined price and bought by other users. Thus users can create their own portfolio of assessments and gather assessments they deem to be important or valuable.

If over time the agreement of an assessment diverges from the user's estimation, the user looses part of the money the assessment was worth before. If it converges to her estimation, she gains money. When a user reconsiders her assessments, for each one a summary of the other users' assessments is displayed. Based on this feedback she can adjust her assessments to fit the market. Here, the market regulates itself and users are rewarded for visiting the platform again. As in real markets, rules can be defined to limit the amount or frequency of changes of an estimation, e.g., through imposing a transaction cost for each change.

In order to effectively gather data with social media operated by Agora, users are encouraged to suggest symbolic goods themselves. This is important to enliven the media run on Agora so that they can grow both in the number of symbolic goods gathered and in the number of their users.


**Figure 1: Borsa Parole during the choice of a region for the displayed sentence. The currently selected region (northern Italy) is highlighted in blue.**

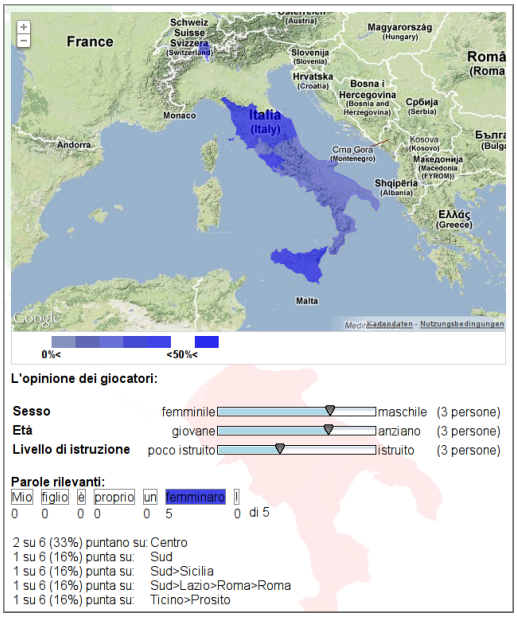


**Figure 2: metropolitalia platform displaying the data gathered for the sentence "Mio figlio è proprio un femminaro!" (in English: "My son really is a womanizer!")**

## 3. BORSA PAROLE: TRADING WITH ONE'S OWN BELIEFS

Specifically on the platform metropolitalia, Agora is used for running the game Borsa Parole, where Italian dialect or vernacular phrases –that is, sentences or parts of sentences– are traded with.[1] In other possible applications of Agora, completely different symbolic goods could be traded with.

The goals of Borsa Parole are to gather new phrases and to encourage users to share their assessments on new or existing phrases. Specifically, the user is asked to indicate in which geographical region a phrase is spoken (see Figure 1), how many people recognize the phrase as being from that location, which word(s) of the phrase are linguistically distinct, and who the speakers are in terms of age, gender, and level of education. Each user action is optional, i.e., can be skipped, to give users freedom in their choice and to prevent false data if users do not know what to choose. The trade of assessments is excluded in this first version of Borsa Parole for the sake of simplicity and will be added at a later stage.

For being successful on Borsa Parole, one has to submit phrases with characteristics that many other users of Borsa Parole are likely to agree with, because there it is easier for others to determine the characteristics. As a consequence, success on Borsa Parole depends on how one is skilled at forecasting others' conceptions. This is a typical case of a "beauty contest", as Keynes described the effect in a speculative market where participants reflect on each others' behaviour and adapt their behaviour accordingly [4]. However, while the beauty contest analogy was meant by Keynes as a criticism of speculation on financial markets, a beauty contest-like speculation contributes to the aim of Borsa Parole. Indeed, in linguistic field research the true opinion of a single speaker is much less relevant than her perception of the community's opinion. In other words, for linguistic field research, speculating speakers are welcome!

[1]So far, the game provides written sentences but an extension with spoken sentences is foreseen. This extension does not require any change in the media logic but only additional user interfaces for collecting and rendering spoken language.

## 4. FIRST EVALUATION

For this evaluation, data was gathered on the platform metropolitalia with the game Borsa Parole during the first six months of its public availability.

During that time 569 users played 3345 rounds of Borsa Parole in total. 40% of all rounds were skipped, probably because the user did not know the phrase well enough to estimate a geographical region. This is natural and was foreseen, giving users the option to skip rounds. For 92% of all geographical assessments the agreement proportion was estimated. This indicates that users are confident in giving such estimations, a finding that encourages to employ the market-based approach Agora in further games. The decreasing numbers of word selections (48% of all rounds) and social characterizations (28% of all rounds) bear evidence that completing these steps is optional. Furthermore for many phrases a social characterization can not be given because it does not exist from a linguistic point of view.

The possibility for users to add new phrases or sentences led to 104 new phrases that were contributed to metropolitalia by 11% of all users who played Borsa Parole. This indicates that the incentives for adding phrases to Borsa Parole are good enough.

Besides the quantity, also the quality of the data gathered is convincing. In Figure 2, the results for a phrase as displayed on the platform are shown. The phrase is assessed to be spoken more in the south of Italy (see the coloured map), the speaker is characterized as male, older, and less educated (see the three sliders), and the selected relevant word is "femminaro", a vernacular word for a womanizer. Though only six users assessed the phrase so far, a clear tendency to the use in the center and south of Italy can be seen. And according to a native Italian speaker knowing this

word, it is well known in Sicily (island in the south of Italy).

Six months are a short period of time because such a platform needed at least one or two years for being sufficiently known, especially if –as in this case– no monetary rewards are provided. Also, GWAP for linguistic field research are uncommon and novel. Therefore, the data gathered so far provide a positive signal on metropolitalia's approach.

## 5. DEMONSTRATION

The demonstration (at `http://www.vimeo.com/59723042` a screencast is available) shows the three most important aspects of Agora with the game Borsa Parole: (1) users assessing characteristics of existing phrases, (2) users adding phrases to the platform, (3) users browsing the market's phrases and adjusting own assessments.

The user starts the game whereupon the first out of three game rounds is started. A phrase in an Italian vernacular or dialect –which we translate and explain for the user– is displayed and the user can choose the geographical region where she guesses that the phrase is spoken (see Figure 1). Furthermore, the user can estimate how many other users she thinks would choose the same region. Next, she can select individual words of the phrase that are specific for this vernacular or dialect and specify the social attributes age, gender, and level of education of speakers of the phrase. After two more rounds –which can be skipped– a summary with the user's score and other users' assessments are shown. This first part of the demo shows that the gameplay is intuitive, gives users freedom in their actions, and provides enough incentives for being accepted by many users. Also the scoring system for earning play-money is demonstrated.

As a second aspect of Borsa Parole, the user is invited to add a new phrase to the platform. This highlights the gathering of new symbolic goods in addition to the assessment of symbolic goods shown before. The user contributes the phrase, its geographical origin, and the user's estimation of how many other users assign the same region.

Finally, browsing the phrases contributes to the attractiveness of the platform and shows the market mechanisms of Agora. Users can look up the assessments of every phrase on the platform and also compare own assessments to other users' assessments and adjust their own ones to the market.

## 6. OUTLOOK AND CONCLUSION

Agora is designed as a generic and modular system and therefore its deployment (1) in complementary games and (2) in other application areas than Italian linguistics is possible. As a complementary game to Borsa Parole, the game *Poker Parole*, Italian for "word poker", is under development and will be available in the next few months. Poker Parole is also based on Agora and shares many properties of its gameplay with Borsa Parole, with one exception: While success on Borsa Parole comes from submitting commonly recognized phrases, on Poker Parole it comes from submitting phrases that most users are not likely to recognize. Such phrases are equally important for linguistic research and therefore need to be gathered as well. The two games therefore will complement each other in the data they gather. We envisage a similar game built upon Agora for the area of art history on the Artigo platform[2] where the symbolic goods traded with would be artworks and the characteristics assessed could be the artist, style, and epoch.

For linguistic field research, crowdsourcing has the potential to gather a huge amount of data from many people in a cost-effective way. The approach furthermore lowers the risk of biased data and offers the possibility to conduct long-term studies over several years, as it is comparably inexpensive to run a Web-based platform for several years. The market-based game design provides new incentives for users which in the evaluation are indicated to be accepted by users. Its analogy to speculation on real markets furthermore yields richer meta-data for evaluation than traditional questionnaire-based field research.

[2] `http://www.artigo.org`

## 7. ACKNOWLEDGMENTS

We thank Hubertus Kohle from the Institute for Art History at the Ludwig-Maximilian University of Munich and all other Play4Science project members for useful suggestions.

This research has been funded in part by the German Foundation of Research (DFG) within the project Play4Science number 578416 (cf. `http://www.play4science.org`).

## 8. REFERENCES

[1] K. A. Davis. Qualitative Theory and Methods in Applied Linguistics Research. *TESOL Quarterly*, 29(3):427–453, 1995.

[2] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing Systems on the World-Wide Web. *Communications of the ACM*, 54(4):86–96, 2011.

[3] B. Hladká, J. Mírovský, and P. Schlesinger. Designing a Language Game for Collecting Coreference Annotation. In *Proceedings of the 3rd Linguistic Annotation Workshop (ACL-IJCNLP)*, 2009.

[4] J. M. Keynes. *The General Theory of Employment, Interest, and Money*. Macmillan Cambridge University Press, 1936.

[5] T. Krefeld. Italienische Varietätenlinguistik. *Italienisch. Zeitschrift für italienische Sprache und Literatur*, 63:56–62, 2010. (In German).

[6] E. Law and L. von Ahn. Human Computation. In R. J. Brachman, W. W. Cohen, and T. Dietterich, editors, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, pages 1–121. Morgan & Claypool Publishers, 2011.

[7] R. Munro, S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily. Crowdsourcing and Language Studies: The new Generation of Linguistic Data. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk at NAACL-HLT*, 2010.

[8] M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. *ACM Transactions on Interactive Intelligent Systems*, 2012.

[9] L. von Ahn and L. Dabbish. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2004.

[10] J. Wolfers and E. Zitzewitz. Prediction Markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.