METROPOLITALIA: A CROWDSOURCING PLATFORM FOR LINGUISTIC FIELD RESEARCH

Fabian Kneissl and François Bry
Ludwig-Maximilian University of Munich, Institute for Informatics
Oettingenstr. 67, 80804 Munich, Germany

ABSTRACT

Traditional linguistic field research employs the cost- and time-consuming and possibly biased approach of gathering data by sending out researchers and filling in questionnaires. We improve linguistic field research using crowdsourcing techniques to remedy these issues and open research to a wider audience. On the online platform MetropolItalia, going public in August 2012, Italian native speakers are united to gather and assess linguistic data themselves. We focus on the Italian language because it is currently undergoing a divergence and in this aspect it differs from many other languages. In this article, Borsa Parole, the first game made public on this platform, is described and a short evaluation of its exploitation in a beta phase is presented. Two important requirements of game-based crowdsourcing – motivating enough users and gathering the data wanted – necessary for such a platform are discussed. The game Borsa Parole is shown to fulfil these requirements.

KEYWORDS

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

1. INTRODUCTION

Linguistic field research is concerned with gathering and analysing speech data from speakers of some language(s) under observation. Traditionally, such multi-dimensional data are collected by sending researchers or students to the speakers' locations, usually in certain geographic regions, where the speakers are interviewed, questionnaires are filled in, and/or speech data are recorded. 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 (Davis 1995, Lazaraton 1995). As a consequence, only relatively limited areas can be covered by traditional linguistic field research.

The online platform MetropolItalia (available at http://www.metropolitalia.org from August 2012) is conceived as a platform for linguistic field research which encourages people to participate in the process of gathering a big linguistic dataset from a wide geographic area with low effort. Such a participation of many users to reach certain goals – that are not necessarily known to the users – is called crowdsourcing and it is a cost- and time-efficient way of gathering data (Doan 2011). The approach furthermore lowers the risk of biased data as data is directly entered into the platform by the speakers themselves without interpretation or other processing by interviewers.

Crowdsourcing can only be successful if firstly many users can be motivated to contribute on the platform and secondly the data gathered comply with the goals of the project. In MetropolItalia, these two requirements are accounted for with carefully designed "games with a purpose". The term "Games With A Purpose" (GWAP) has been introduced by von Ahn and Dabbish (2004) for games where users play and at the same time contribute with data for some goal. As a first GWAP on the MetropolItalia platform, Borsa Parole was launched as beta version in March and is to go public in August 2012. Borsa Parole focuses on gathering phrases in Italian vernaculars – that is, unstandardised language varieties – and dialects together with peoples' assessments of the regions where the phrases are spoken and of social characteristics – age, sex, education, etc. – of the speakers uttering the phrases. Users can submit their own phrases to the platform and guess on the regional origin of phrases already gathered by the platform. Additionally, users can place a bet

on a phrase expressing a freely chosen percentage of users expected to assign the phrase to a specific region. The user can trace the performance, that is correctness, of her own bets and adjust them. For the linguistic field research which is the raison d'être of the platform, such percentages provide valuable information about the regional anchoring of phrases. Borsa Parole is presented in this article focusing on motivating players (or "fun factor") and gathering the data sought for.

The contributions of this article are as follows:

- Presentation of the platform MetropolItalia and its first game Borsa Parole
- Discussion about two goals of successful crowdsourcing in GWAP like Borsa Parole
- Presentation of preliminary results gathered by Borsa Parole.

2. GATHERING DATA FOR ITALIAN LINGUISTIC FIELD RESEARCH

The Italian language is particularly interesting for linguistic field research because the language spoken today everywhere and within all social groups is currently undergoing a divergence which originates in the big cities and spreads from there (Melchior 2009, Krefeld 2010). Vernaculars and dialects from various regions influence each other, generating new vernaculars. This makes Italian different from most other languages like German, English, or French. The Italian vernaculars differ both in their written and spoken forms and can therefore be captured in both forms. Using a crowdsourcing approach where all speakers with an Internet connection are able to contribute can reveal a far larger amount of multi-dimensional linguistic data than is possible with traditional field research.

Several attributes of vernaculars and dialects are of interest for linguistics. A comprehensive description of the field research for which this crowdsourcing approach is presented is given in (Krefeld 2011). Note that a phrase's region as well as a speaker's age, gender, and level of education are worth collecting. Whether these characteristics are enough to provide sufficient linguistic insights on the use and spread of Italian vernaculars and dialects remains to be investigated in further linguistic research. They do however provide a good basis for the main research questions in this field. They are therefore the characteristics gathered by the game Borsa Parole.

3. BORSA PAROLE AS A MARKET FOR LINGUISTIC SPECULATION

Borsa Parole acts as a market where users submit phrases, assess, bet on, and characterise other users' phrases and their speakers, select distinctive words of the phrase, and receive points for their doing when they are successful. A phrase consists of one or more words and it should be assigned a vernacular or dialect (though this is not necessary). It is accompanied by a phrase in standard Italian so that other users understand the phrase's meaning even if they are not proficient in its vernacular or dialect. An assessment is composed of the user creating it, a phrase, and a selected region. A bet in addition contains the chosen proportion of speakers likely to agree on a phrase's region. A characterisation is the rating of age, gender, and level of education of the speakers of the phrase together with the user and the phrase. Finally, a selection consists of the user, the phrase, and the words by which the user recognised her assessment and characterisation.

Two modes exist in Borsa Parole: In a first "input mode", users enter a phrase together with its standard Italian equivalent and assess it. In a second "game mode" consisting in a number of "rounds" – we experiment with three rounds – one phrase per round is presented to a user. First, she has to name the region where the phrase is spoken. Second, she has to enter a proportion of speakers likely to agree on this region. She can also opt to skip this step if she does not feel confident about making a choice. Next, she has to provide a characterisation and a selection, what can also be skipped. After the user completed the required number of rounds, a summary of the user's actions and a comparison to the actions of the other users is shown so that the user can explore what others think of the phrases she has just "processed". Points are awarded if the user's choice matches the average choice of other users: The better they match, the more points a user gets.

Two goals have to be fulfilled for this game to be successful: Enough users have to be motivated and the desired data has to be collected. Borsa Parole provides the following incentives: First and foremost, language itself inspires people and playing with one's own language, guessing and getting new insights into it turns out

to be enough of an incentive in itself. Second, the platform being a gaming platform, points, and competition between users provide a "fun factor". Third, a user can access the list of her bets, revise them and doing so try to increase her gains. Thus, several incentives are provided. An evaluation of the public version will reveal whether they are sufficient.

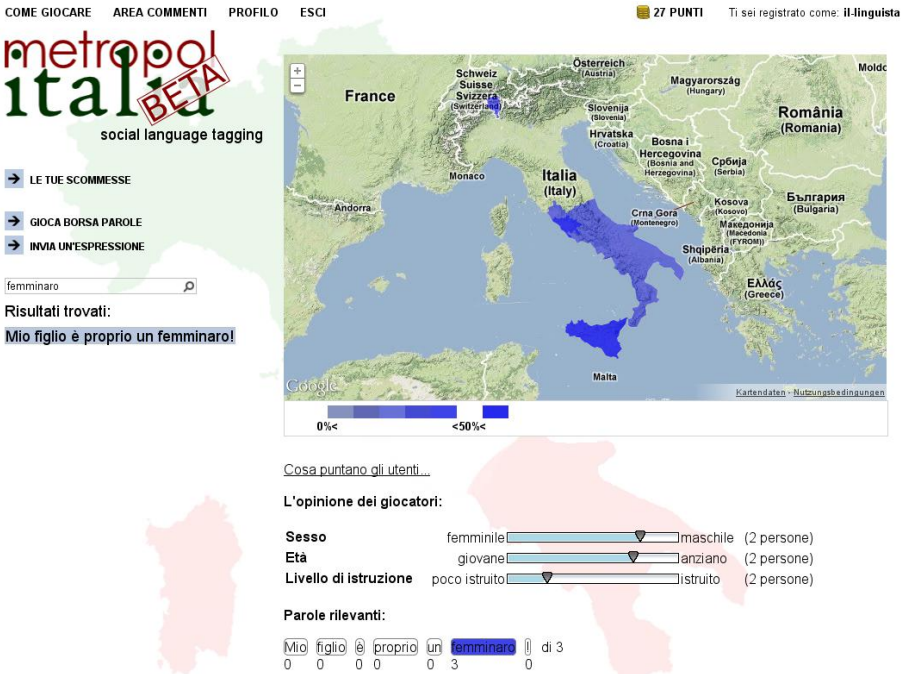
The second goal, gathering of the data sought for, is fulfilled as well. First, the data generated by the users is stored in a database to allow its analysis at any time. This is especially important as users can change their bets what can provide useful insight into a user's language knowledge. Furthermore, as it is done in many other GWAP, the resource shown to the user, that is the phrase, is chosen randomly to prevent purposeful data manipulation and cheating. As more and more users contribute with their assessments of a phrase, the quality of the results increases, converging to an assessment the majority of users agree on.

4. PRELIMINARY RESULTS OF THE BETA-VERSION

A non-advertised beta phase of Borsa Parole with a limited group of people was initiated to get feedback about the game logic, usability, and to test the stability of the platform. During the beta phase 30 assessments, 78 bets, 69 selections, and 46 characterisations have been made by 41 (mostly anonymous) users in 186 completed game rounds. These data show that bets are accepted well (72% created a bet and not just an assessment), a bit less than half of the game rounds (42%) are skipped (e.g., because of an unknown phrase), and users are less likely to enter characterisations than selections.

The relatively low amount of data collected only permits conclusions of limited significance. Below, in Figure 1, the results as displayed on the platform are shown. The highlighted phrase (in English "My son really is a womaniser!") is assessed to be spoken more in the south of Italy (see the coloured map), the speaker characterised as male, older, and less educated (see the three sliders), and the selected relevant word is "femminaro", a vernacular word for a womaniser. Though only four users assessed the statement, a clear tendency to the south 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).

Figure 1: MetropolItalia platform displaying the data gathered for "femminaro" (English: "womaniser")



Besides data gathered by the platform, users gave informal feedback. This feedback lead us to display better responses to the user's actions, display more results on the platform, adjust the scoring system to a more user-friendly "fuzzy region approach" which takes neighbouring and higher- and lower-level regions

into account, reduce the number of game rounds from five to three, improve the graphical user interface and wording, and enhance the performance and stability of the platform.

5. RELATED WORK

The term GWAP has been introduced by von Ahn and Dabbish (2004). Since, several GWAP have been developed – for a comprehension see (Law and von Ahn 2011). Related to GWAP, though broader, are the terms "human computation" and crowdsourcing which both describe the act of relying on people to solve problems not solvable for computers (Law and von Ahn 2011, Doan 2011). Several crowdsourcing studies for linguistics have been conducted relying on low paid online workers, or "turkers", using Amazon's Mechanical Turk (Munro et al 2010). To the authors' best knowledge, no crowdsourcing approach so far has focused on gathering and assessing vernaculars or dialects on an online gaming platform.

6. OUTLOOK AND CONCLUSION

This article describes the platform MetropolItalia together with its first game Borsa Parole, shows two requirements especially important for GWAP, and presents preliminary results from the beta phase of the game. The forthcoming launch of MetropolItalia will give further results and more insights into its ideal of an enlivened crowdsourcing platform for Italian linguistic field research. Also further games are in development which will complement Borsa Parole with other ways to gathering more and different linguistic data, e.g., also audio material.

ACKNOWLEDGEMENT

We thank Thomas Krefeld from the Institute for Roman Studies, Stephan Lücke from the IT Support Group for the Humanities and Christoph Wieser from the Institute for Informatics, all three with the Ludwig-Maximilian University of Munich, for useful suggestions. This research has been funded in part by the German Foundation of Research (DFG) within the project Play4Science number 578416.

REFERENCES

Davis, K., 1995. Qualitative Theory and Methods in Applied Linguistics Research. In *TESOL Quarterly*, Vol. 29, No. 3, pp. 427-453.

Doan, A., Ramakrishnan, R., Halevy, A., 2011. Crowdsourcing systems on the World-Wide Web. In *Communications of the ACM*, Vol. 54, No. 4, pp. 86-96.

Krefeld, T., 2010. Italienische Varietätenlinguistik. In *Italienisch. Zeitschrift für italienische Sprache und Literatur*, Vol. 63, pp. 56-62. In German.

Krefeld, T., 2011. Alter Standard - Neue Medien. Zur Erfassung von Restandardisierungsprozessen im Italienischen. In Schmid, S. et al (eds.), *Koineisierung und Standardisierung in der Romania*. Universitätsverlag Winter. In German.

Lazaraton, A., 1995. Qualitative Research: in Applied Linguistics: A Progress Report. In *TESOL Quarterly*, Vol. 29, No. 3, pp. 455-472.

Law, E. and von Ahn, L., 2011. Human Computation. In Brachman, R. et al (eds.), *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool.

Melchior, L., 2009. «Frocio», «checca», «morosa» e… un problema lessicografico. In *Italienisch. Zeitschrift für italienische Sprache und Literatur*, Vol. 62, pp. 67-88. In Italian.

Munro, R. et al, 2010, Crowdsourcing and Language Studies: The new Generation of Linguistic Data. *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk at NAACL-HLT*. Los Angeles, CA, USA, pp. 122-130.

von Ahn, L. and Dabbish, L., 2004, Labeling Images with a Computer Game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Vienna, Austria, pp. 319-326.