Funded by the Horizon 2020
Framework Programme of the
European Union

Challenges of crowdsourcing with respect to enetCollect

Katerina Zdravkova
University Ss. Cyril and Methodius, Skopje, Macedonia

Outline of the presentation

- Goals of enetCollect
- Challenges related to all working groups
  - Collaborative creation of lesson content
  - Observing how different samples of users perform
  - Producing language learning material and language-related datasets
  - Generating exercise content by cross matching the learners' answers
  - Strategies fostering user-orientation of ensuring its capacity to attract and retain a crowd
  - Creating the technical specifications to support the functional demands
  - Devising the application-oriented part of the theoretical framework
- Conclusion

Intentions of enetCollect (1)

- WG1: Developing or adapting explicit crowdsourcing approaches for producing language learning material.
  - Example: Researching the most effective ways to collaboratively devise lesson content and to assess how effective it is by observing how different samples of users confronted with the created content perform subsequently.
- WG2: Developing or adapting implicit crowdsourcing approaches for producing language learning material and language-related datasets.
  - Example: Research ways to generate exercise content from language resources (e.g. lexica) and to crowdsource manual validation of automatically generated new entries (e.g neologisms) by cross matching the learners' answers regarding exercise content generated from such new entries.

Intentions of enetCollect (2)

- WG3: Design strategies fostering the user-orientation of an online language learning solution and ensuring its capacity to attract and retain a crowd.
  - Example: study the relevance and attractiveness of learner profiling for vocabulary training and devise related design strategies
- WG4: Creating the technical specifications to support the functional demands of WG1, WG2, and WG3.
  - Example: study technical solutions for the scalability of a crowdsourcing
- WG5: Devising the application-oriented part of the theoretical framework that relates to:
  - (1) ethical questions regarding the involvement of users and the collection of data
  - (2) legal regulations
  - (3) opportunities and models for commercialization

Collaborative creation of lesson content (WG1)

1. Content creators
   - Members of WG1 are not necessarily educationalists or language experts
   - In many occasions, lessons might be language specific (26 languages belonging to 8 language families)
   - Content creators must be professionally responsible:
     - Prompt
     - Confident
     - Trustworthy
2. Copyright infringement
   - How to prevent appropriating of the content created by others?
   - Should copyright policy be created prior to content creation?

Observing how different samples of users perform (WG1)

1. Privacy issues
   - Which personal data will be collected from the real users and why?
   - Who will have an access to these data?
   - Will the identity of the users be visible to other users (anonymity is recommended)?
   - Should privacy policy be created prior to content creation?
2. Problems with users contribution
   - In open question answers, people usually reveal private matters (own, and of other people)
   - Incompetence of some users might be a source of harassment and bullying
3. Freedom of speech
   - There is a risk that some users might deliberately promote hate speech, or users who might use the crowdsourcing to spread various dogmas or ideologies
   - Should terms of use be created prior to user enrolment to a particular lesson?

Producing language learning material and language-related datasets (WG2)

1. Institutional challenges
   - How much does enetCollect agree the national educational systems (34 countries)?
   - Is there a need for a permission to implement the content with real users?
   - Do we need a national accreditation of the content?
   - Real users recruitment heavily depends on WG3 activities
2. Copyright issues
   - Copyright rights are territorial rights!!!
   - Who will be a copyright owner (COST, the action, some publisher, …)?

Generating exercise content by cross matching the learners' answers (WG2)

1. How reliable are learners' answers?
   - In many languages and cultures, there are common wrong answers, that are more frequent than the correct ones:
     - Generally accepted wrong spelling, beliefs, viewpoints, …
     - Ambiguities due to the lack of a unique translation equivalent of some new terms, particularly in the languages spoken by few people (NB. Translate crowdsourcing into your own language)
   - Some exercises might have no correct answers (due to users' ignorance or incompetence), resulting in a poor content quality
   - All the mistakes should be carefully corrected. Otherwise, they will lead to wrong exercise content
2. Does enetCollect have experts for all the languages to rely on, or external (language, topic, course) experts should be engaged?

Strategies fostering user-orientation of ensuring its capacity to attract and retain a crowd (WG3)

1. Institutional challenges (once again)
   - National Ministry of Education should endorse and support enetCollect activities
   - National legal restrictions should be carefully examined by WG5
2. Target samples of users (age, previous experience with similar tasks, education, …) determine the appearance and functionalities of the crowdsourcing system
3. How will the users be enrolled in the system?
4. What is the minimum amount of users per language / country?
5. Should National Bureau of Education be directly involved in the aand how?

Creating the technical specifications to support the functional demands (WG4)

- Where will the technical specifications be hosted?
- Will the content of the system password protected to reduce privacy risks and copyright infringement?
- Serious security measures should be taken into consideration to prevent any form of hacking and information security threats?
- Additional measures should be implemented for the mobile version.
- How much is the host server reliable?
- How frequently will the content be backed up?
- Should a mirror host be created to minimize the risk of losing the data?

Ethical questions regarding the involvement of users and the collection of data (WG5)

- Trustworthiness and professional responsibility of content creators
- Language expertize for all the languages
- Intellectual property
- Privacy, confidentiality, and anonymity of all the users *
- Freedom of speech
- Quality of crowdsourced material
- Security and reliability of the technical solution

* So far, the identities of EnetCollect participants are visible to all

Legal regulations (WG5)

- Legal regulations should be carefully examined
- Although most laws (education, protection of private data, intellectual property) are in concordance with the European Commission recommendations, there are many country specific issues
- Should legal solicitors be involved in the action and how?

Opportunities and models for commercialization (WG5)

- Impact of enetCollect
- Is the commercialization of the final outcome accepted by COST?
- Should the final product be nationally accredited?
- In the case of further commercialization, who will sell the product (and consequently, make a financial profit)?
- How will be personal contribution in the creation of the final product be recognized?
- Will further commercialization make the system self sustainable?

Conclusion

- Various documents must be created prior to launching enetCollect:
  - Strict copyright policy
  - Privacy policy
  - Terms of use
- (Crowd) review of all the learning content must be made
- External expert reviews are inevitable, at least one per country (34 reviews?) and language (26 languages?)
- With prior definition of the strategies and carefully defined policies, most of the challenges mentioned in this presentation will be avoided