الجمهورية الجزائرية الديمقراطية الشعبية

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

وزارة التعليم العالي و البحث العلمي

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

جامعة عمّار ثليجي بالأغواط

UNIVERSITY OF AMAR TELIDJI LAGHOUAT

كلية العلوم

FACULTY OF SCIENCES

قسم الإعلام الآلي

DEPARTMENT OF COMPUTER SCIENCES

Master Thesis

Domain Mathematics and Computer Science

Field Computer Science

Option Networks and Distributed Systems

By:

Bougrine Fatna

Djellikh Soumia

Topic

Free Crowdsourcing-Based Corpus Annotation

Defended Publicly on 01-06-2017 in Front of the Jury Composed of

M_r.	A. NEHAR	M.C.A	PRESIDENT	UZAD
M_s.	A. CHORANA	M.A.A	REVIEWER	UATL
M_s.	H. CHERROUN	PROF.	SUPERVISOR	UATL

Academic Year 2016/2017

Dedication

This thesis is dedicated to:

- The sake of Allah, my Creator and my Master, My great teacher and messenger, Mohammed (May Allah bless and grant him), who taught us the purpose of life.

- Our family, the reason of what we become today, thanks for your support and continuous care.

- Our friends, who encourage and support us.

- All our Prophs who gave us knowledge.

- Laghouat University prophs and students.

- Finally, to every one who contributed from near or far in this thesis.

Bougrine and Djellikh ...

Acknowledgements

Firstly, I must acknowledge my limitless thanks to Allah, the Ever-Magnificent; the Ever-Thankful, for His help and bless. I am totally sure that this work would have never become truth, without His guidance.

I owe a deep debt of gratitude to our university for giving us an opportunity to complete this work.

I would like then to express my deepest gratitude to my beloved family: Parents, my lonely brother and my two sisters, thank you for your support and limitless love and care, you were always there for me, without you i could not reach at this level.

- Special, thanks to my lovely sister Soumia, you were always beside me when no one is, you teach me how to be patient in my hardest days. Honestly, without your support and instructions i could not make it. I am so grateful that you are my sister, God bless you.

I thank Ms.Bouzouad for accepting supervising me despite their big busyness and for their encouragements, guidelines, and kindness. Big thanks go to the jury for accepting to examine this modest manuscript.

I would I would like to take this opportunity to say warm thanks to all

my beloved friends, who have been so supportive along the way of doing my thesis.

Last but not least, deepest thanks go to all people who have had a significant impact on my graduate career by their support and encouragement. Thank for making this thesis real.

Bougrine Fatna ...

Abstract

Large corpora are very useful to develop and validate Natural Language Processing (NLP) systems. However, these corpora are generally collected and annotated automatically. To validate such annotation, two solutions are possible. We can use skills of expert, which can be costly and time consuming, or use crowdsourcing technique. Crowdsourcing can be defined as the act of attracting many non experts to complete a certain task by using paid/unpaid dedicated platform. In this work, we are interested to validate a semi-automatic dialect annotation of Kalam'DZ corpus. Our approach relies on free crowdsourcing using Crowdcrafting platform. The validation is performed on 10% (11 hours) of the total size of Kalam'DZ. A quality control of this validation is ensured through a confrontation with expert annotation, which shows that more than 80% of annotations are similar. Our results confirm that free crowdsourcing is effective for speech dialect annotation.

Keywords: Algerian dialects, ANLP, Corpus Annotation, Crowdcrafting, Crowdsourcing.

مُلخّص

تستعمل المدونات في مجال معالجة اللغات الطبيعية بشكل كبير، ولكن هذه المدونات يتم على الأغلب جمعها و تصنيفها بشكل شبه آلي. في سبيل التحقق من صحة هاته التصنيفات يمكن ان نستعمل طريقتان: اما الاستعانة بخبراء في المجال، مما قد يستغرق وقتا كثيرا و يكلف ماديا، أو استعمال تقنية التعهيد الجماعي. يمكن تعريف هذه الأخيرة على أنها جذب الجمهور لأداء مهمة معينة على مستوى منصة مجانية أو مدفوعة الأجر. في هذا العمل نتطلع إلى التحقق من صحة تصنيف لهجات مدونة KALAM'DZ أين تم الحصول عليها من خلال موارد الشبكة العنكبوتية. الطريقة المتبعة تعتمد على التعهيد الجماعي المجاني باستعمال منصة .Crowdcrafting اعتمدنا في التقيم على نسبة 10% (10 ساع) من مجموع المدونة. من خلال مقارنة جودة هاته التعليقات التوضيحية مع أخصائيين في هذا المجال. حيث تطابقت الاجوبة بنسبة 80% . أظهرت النتائج أن التعهيد الجماعي المجاني فعال في تصنيف اللهجات الصوتية.

الكلمات المفتاحيّة:

اللهجات الجزائرية، معالجة اللغات الطبيعية العربية، التعليقات التوضيحية للمدونة، التعهيد الجماعي، كراودكرافتينق.

Résumé

Les grands corpus sont très utiles pour développer et valider les systèmes de traitement du langage naturel. Cependant, ces grands corpus sont généralement collectés et annotés automatiquement. Afin de valider cette annotation, deux solutions sont possibles. Nous pouvons faire appel aux compétences des experts, qui peuvent être coûteuses et longues, ou utiliser la technique de crowdsourcing. Crowdsourcing peut être défini comme l'act d'attirer un nombre important de non experts pour accomplir une certaine tâche en utilisant une plate-forme web dédiée payée/non payée. Dans ce travail, nous nous sommes intéressés de valider l'annotation du dialecte pour le corpus Kalam'DZ. Ce dernier est un corpus vocal des dialectes algériens. Notre approche repose sur le crowdsourcing (gratuit) utilisant la plate-forme crowdcrafting. La validation est effectuée sur un échantillon de 10% de la taille total du corpus (11h). Un contrôle de qualité de cette annotation a été assuré via une confrontation avec une annotation faite par des experts, ce qui montre que plus de 80% de réponses similaires. Nos résultats confirment que le crowdsourcing gratuit est efficace pour l'annotation du dialecte vocal.

Mots-clés: Dialectes Algériens, Traitement automatique du langage naturel arabe, Annotation de corpus, Crowdsourcing, Crowdcrafting.

Table of Contents

List of Figures

List of Tables

Abbreviations

NLP	Natural Language Processing
MSA	Modern Standard Arabic
DA	Dialectal Arabic
Mturk	Amazon Mechanical Turk
CF	CrowdFlower
GWAP	Games With A Purpose
WotC	Wisdom of the Crowds
POS	Part Of Speech
CL	Computational Linguistics
ANLP	Arabic NLP
WER	Word Error Rate
CSS	Cascading Style Sheets

Introduction

The field of study that focuses on the interactions between human language and computers is called Natural Language Processing, or NLP for short [1]. NLP is a way that uses computers to analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks [2].

The goals of NLP are to design and build applications [3] that facilitate human interaction with machines and other devices through the use of natural language such as *Machine Translation* used by Google Translate, Twitter, and Facebook posts, *Sentiment Analysis* which is oriented to identify subjective information in texts. It can be a judgment, an opinion or an emotional state, *Spam filters* such as Gmail use NLP to determine which emails are good and which are spam and *Language Identification* is to identify the language being spoken by some unknown speaker using a given speech sample.

A speech corpus is a large collection of audio recordings of spoken language. It is crucial for both developing and evaluating NLP systems. Moreover, such corpora have to be large to achieve NLP communities expectations. Geographically, Arabic is one of the most widespread languages of the world [4]. It is spoken by more than 420 million people in 60 countries of the world [5].

Actually, it has two major variants: Modern Standard Arabic (MSA), and Dialectal Arabic (DA). MSA is the official language of all Arab countries. It is used in administrations, schools, official radios, and press. However, DA is the language of informal daily communication. Recently, it became also the main of communication on the Web, in chat rooms, social media etc. This fact, amplifies the need for language resources and language related NLP systems for dialects.

For many languages, the state of the art of designing and developing speech corpora has achieved a mature situation. On the other extreme, there are few corpora for Arabic [6]. Moreover, very few attempts have considered Algerian Arabic dialect.

Recently, KALAM'DZ corpus is developed by Bougrine et al. [7] which covers eight major Arabic sub-dialects of Algeria. KALAM'DZ corpus is collected by using web-based method. The size of the corpus is about 104 hours with 4881 speakers. Despite its important size, most of the dialect annotations are provided from the related metadata of the Web-source when they exist. The results of these annotations need a validation process by experts to taken into account. This process can be consuming time and costly.

Crowdsourcing is a way of solving problems, obtaining services, ideas, and producing things by connecting on-line with large group of people.

Crowdsourcing technique offers the promise of dramatically lowering the cost of collecting and annotating speech data. In addition, Zaidan and Callison-Burch claim that one of the advantages of crowdsourcing is "access to foreign markets with native speakers of many rare languages" [8] . This feature is particularly useful for those that work on less-resourced languages

such as Arabic [9].

In this work, we are interested to investigate the validity Kalam'DZ annotations by exploring and harnessing free crowdsourcing.

This manuscript contains three chapters arranged as follows:

1. The first Chapter of this manuscript is dedicated to describe the generalities of crowdsourcing, its origins, some popular platforms, and we view the crowdsourcing corpus annotation process.
2. In the second Chapter, we review some related work about Arabic natural language processing: annotation, transcription, translation, and game with purpose.
3. In the third Chapter we describe and explain the principle of our crowdsourcing approach. Followed by implantation, experimentation, and the results of participation.
4. In the conclusion, we summarize the main results and point out some perspectives.

Chapter 1

Crowdsourcing: Generalities

In this chapter, we describe generalities of crowdsourcing. Some basic requirements for Natural Language Processing crowdsourcing are shown. Speech crowdsourcing is more detailed. Then, we show some platforms dedicated to crowdsourcing. Finally, we show the link between crowdsourcing and Social Media.

Most of the content of this chapter came from two main specialised references dues to [10], [11] and [12].

1 Crowdsourcing Definition

In literature, many definitions of the crowdsourcing concept exist, we have chosen two among them: an operational definition and a functional one.

1.1 Operational Definition of Crowdsourcing

The operational basis of crowdsourcing rests on the idea that a task is to be done, there is means to attract many nonexperts to complete this task, and that some open call has gone out to advertise the task to the nonexperts. The presence of the Internet and cellphones facilitates the open call for nonexperts, the presentation of the task, its accomplishment, and the aggregation of the nonexperts' opinions, that they have by being native speakers of a given language.

Eskenazi et al. believed that the aggregation of opinion of many non experts will approach the quality of the opinion of an expert [10]. Where the use of non experts in this manner will be less onerous and more rapid than the use of experts.

1.2 Functional Definition of Crowdsourcing

James Surowiecki [1] defines four characteristics of the helpful crowd [13]. First, the members of any crowd have a *diversity of opinions*. Where it can be only a little different from one another, and some may be correct while others are wrong. Second, each member of the crowd has an opinion that is *independent* of all of the other members of the crowd, to avoid that member's opinion can be influenced by any other member. Third, information that the crowd may have is *decentralized*. Everyone has some local information, but no one in the crowd has access to all of the information that may be related to the task. Finally, the opinions of the members of the crowd can be *combined* to form and aggregate, one collaborative solution [10].

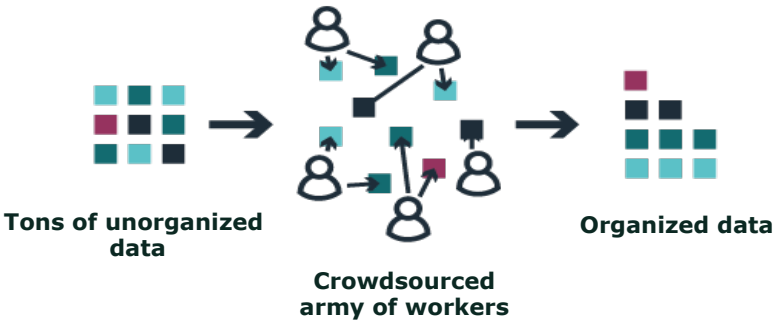

Figure 1.1: Schematic of Crowdsourcing Principe.

2 Origins of Crowdsourcing

In Wired Magazine, Jeff Howe created a buzz word crowdsourcing in 2006, but the process of crowdsourcing was invented as early

[1] James Michael Surowiecki is an American journalist. He is a staff writer at The New Yorker. In 2004, he published The Wisdom of Crowds.

as 1714. Since then, crowdsourcing has helped create some of the world's greatest inventions and biggest brands. Figure 1.2 represents the history of crowdsourcing [5].

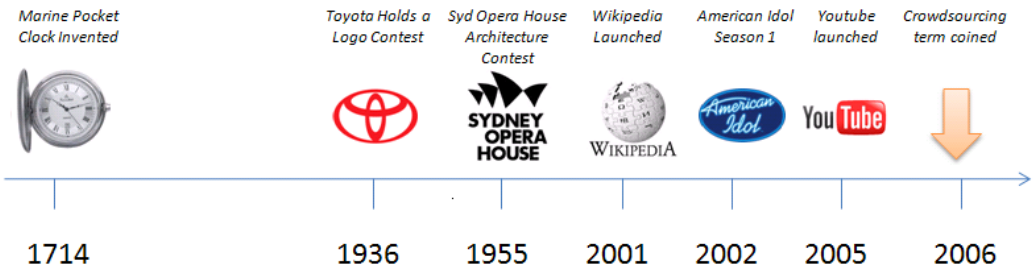

Figure 1.2: The History/Genesis of Crowdsourcing.

2.1 The Longitude Prize

In 1714, the first ever example of crowdsourcing was according to the British Government. They searched a solution to what they called "The Longitude Problem" which made sailing difficult and risky (killing 1,000s of seamen every year). They offered 20,000 £ for people to invent a solution. This is possibly a great example due to its ability to highlight one of the crowdsourcing principles which is "innovation and creativity can come from anywhere ".

2.2 Toyota Logo Contest

In 1936, Toyota held a logo contest to redesign its logo. The winning logo was the three Japanese katakana letters for *Toyoda* in a circle, which was later modified to *Toyota* by Risaburo.

2.3 The Sydney Opera House

In 1955 the Premier of New South Wales state of Australia, Joseph Cahill, ran a contest offering £5,000 to design a building for part of Sydney's Harbour. The contest received 233 entries from 32 countries around the world. The winning design is one of the most innovative landmarks. Architectural contests continues to be a popular model for getting buildings designed.

2.4 2000 to 2006: YouTube, Wikipedia

In this period of time, Youtube[1] and Wikipedia[2] was exhibit to the public which are in those days the most popular and useful sources.

2.5 2002 to 2006: American Idol

In 2002, American Idol Season kicked off Kelly Clarkson's career as well as a plethora of talent contests So You think You Can Dance, Next Top Model, Masterchef. These contests, often described as reality TV, at their core, public crowdsourcing contests that aim to produce an album, a cook book or a superstar (along with entertainment for plus than 1 billion people) [5].

3 Why You Should Use Crowdsourcing?

Many reasons lead to crowdsource usage, it is inherent to human being to share work:

[1] Crowdsourced entertainment/TV
[2] Crowdsourced knowledge

- *Need More Talent*: It is difficult to find a top talent, take a time to train, and expensive to retain. Use crowdsourcing to access the right talent when you need them.
- *Need Deliver Fast*: Today, digital world happens in days and weeks, not months and years. For that, crowdsourcing can help you move faster to get more done.
- *Need to Innovate*: The use of crowdsourcing can also uncover innovative solutions, you can execute with confidence.
- *Need to Reduce Cost*: Presenting your project to a group of non-experts will be less cost compared to experts [14].
- *Need to Create Buzz*: If you have an idea or project and you need to make a publicity, definitely crowdsourcing is an efficient way to present your work for a wide range of people.

4 Who Can Use Crowdsourcing?

Figure 1.3 summarizes who can profit from crowdsourcing. In fact, both individuals and groups, private and governments institution can use crowdsourcing concept. In general, any task that can be described and explained can be done by crowd.

5 Some Terminology

In order to be familiarised with the concept of crowdsourcing, let us define those terms: crowd, requester, worker, task, set of tasks, unit task, throughput, and submission.

- A crowd is a group of non experts who have answered an open call to perform a given task.

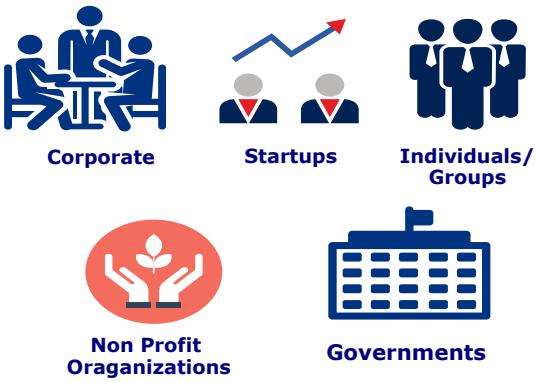

Figure 1.3: Who Can Profit from Crowdsourcing.

- The person who is *creating the task* and *who submits* it is called *"the requester"* (such as in mturk). This person may be called *the client* at other crowdsourcing platforms.
- The person in the crowd who does the work is accordingly called *the worker*. However, according to the used platforms we can meet some other terms such as Turker, a *freelancer*, *contributor* respectively for mturk, MiniFreelance, and CrowdFlower (CF) platforms.
- The individual task itself is called a *Human Intelligence Task* (HIT) at mturk, or a *mission* at AgentAnything.com, or a *microjob* at MicroWorkers, and a *task* at cf platform.
- Set of tasks is the complete set of items that the requester wants to have done.
- Unit task, for example, transcribing one utterance out of the 1000 hours of speech in the set of tasks.
- *Throughput* referring to the number of unit tasks completed per hour.
- *Submission* referring to when the requester makes a set of tasks available to the workers.

Those definitions are more detailed in [10].

6 Common Types of Crowdsourcing

There are many types of crowdsourcing, we will mention some of them: crowd voting, crowd funding, crowd searching, micro work, macro work, and inducement prize contests.

6.1 Crowd Voting

Crowd voting rests of idea that people are called to vote or Voice their opinion on a particular thing. This can be either in garment industry, food products, or even when selecting a wonder of the world.

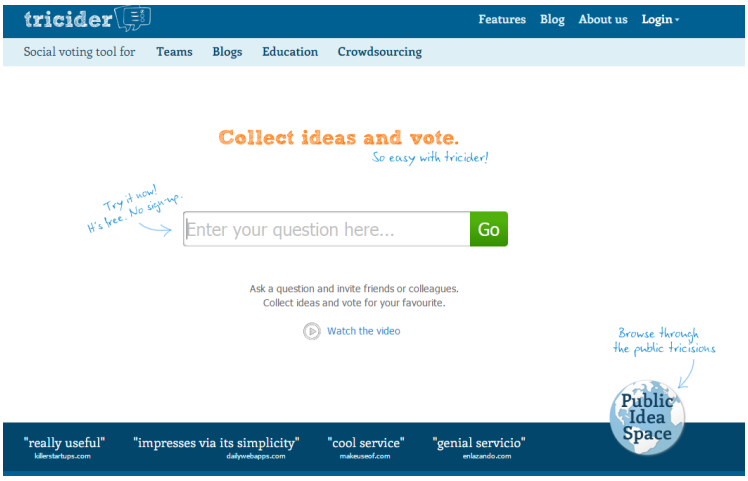

Figure 1.4: Crowd Voting: Tricider Website.

Example: Tricider[1] is a virtual collaborative tool which supports the creation of ideas, writing and commenting of

[1]Tricider, https://www.tricider.com/

opinions and even voting for them. Figure 1.4 illustrates Tricider Website.

6.2 Crowd Funding

Crowdfunding is a way of raising finance by asking a large number of people each for a small amount of money for a global cause or monetary.

Example: GoFundMe[1] website is the world's largest social fundraising platform, with over $3 billion raised so far, which is launched in 2010. Figure 1.5 represents GoFundMe website.

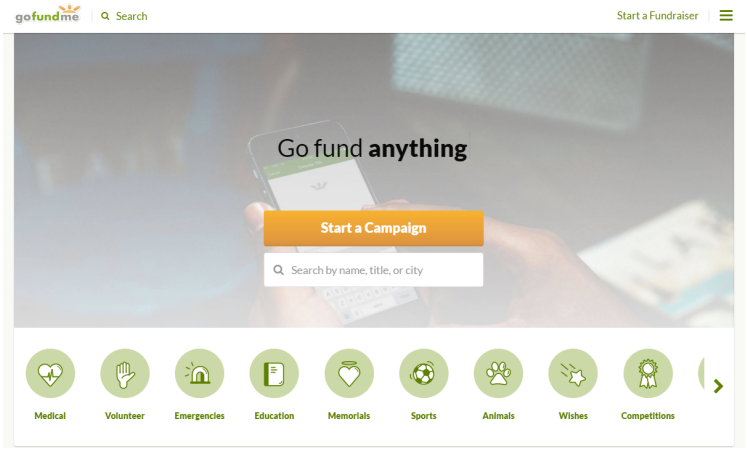

Figure 1.5: Crowd Funding: GoFundMe Website.

6.3 Crowd Searching

Crowd searching is similar to a lost and found items. When a valuable thing has lost like a pet or item or property. A virtual

[1]GoFundMe, https://www.gofundme.com

search can be passed by Internet to many people could searching at the same time.

Example: Yelp[1] was founded in 2004 to help people find great local businesses like dentists, hair stylists and mechanics. Yelp allows customers to share their experiences and businesses to track reviews. Figure 1.6 illustrates Yelp website.

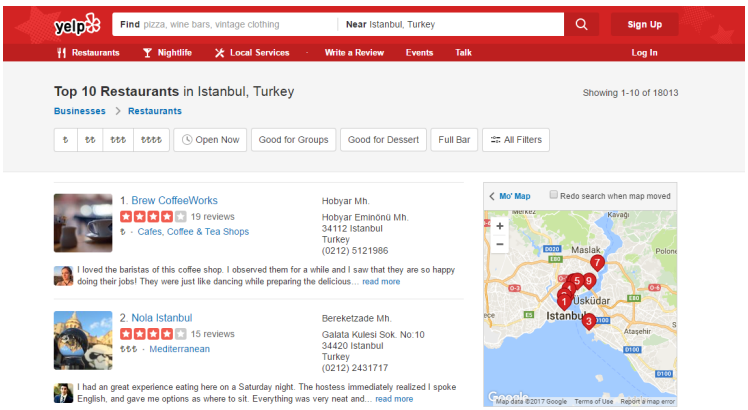

Figure 1.6: Crowd Searching: Yelp Website.

6.4 Micro Work

This is a type of crowdsourcing where a small tasks or jobs are accomplished by many people over the Internet with a small amount of payment.

Example: AskforTask[2] website is a Canadian Company that is focused on one very simple mission, "to be the world's Largest Local marketplace for daily missions". Figure 1.7 reports the most popular services in AskforTask website.

[1]Yelp, https://www.yelp.com/sf
[2]AskforTask, www.askfortask.com

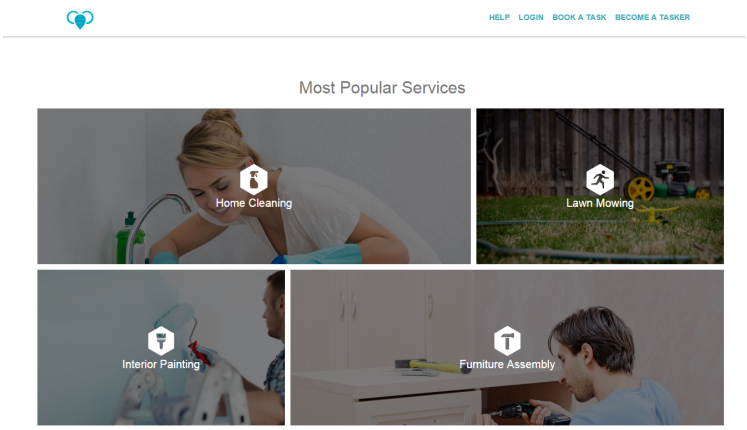

Figure 1.7: Most Popular Services in AskforTask Website.

6.5 Macro Work

Macro work is a type of crowdsourcing where a specialized skills are involved to perform specific projects. This is ideal for Web design, collateral development, content writing and application development.

Example: Topcode[1] Community includes more than one million of the world's top designers, developers, data scientists, and algorithmists. Figure 1.8 illustrates Topcoder algorithms and analytics.

6.6 Inducement Prize Contests

This type of crowdsourcing demand people to come up with their own ideas for a grand prize which may be any type of ideas depending on the company. This type of sourcing is ideal for

[1]Topcode, https://www.topcoder.com/

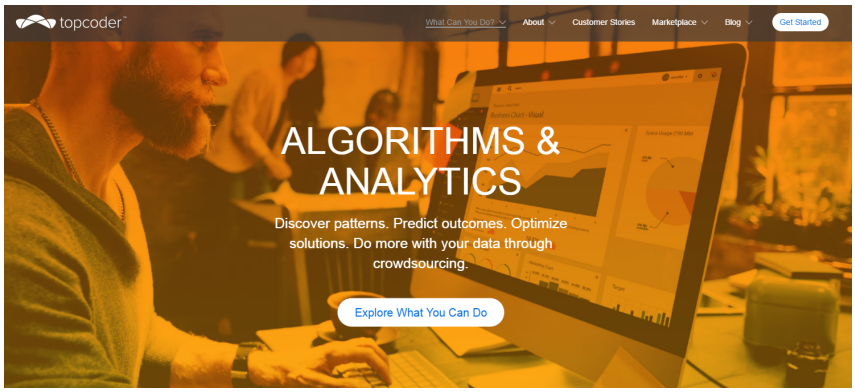

Figure 1.8: Macrotasking: Topcoder Algorithms and Analytics.

tasks including graphic design, software testing, analytics and other creative projects [15].

Example: DesignCrowd[1] is an online marketplace providing logo, website, print and graphic design services by providing access to freelance graphic designers and design studios around the world. Figure 1.9 represents DesignCrowd Website.

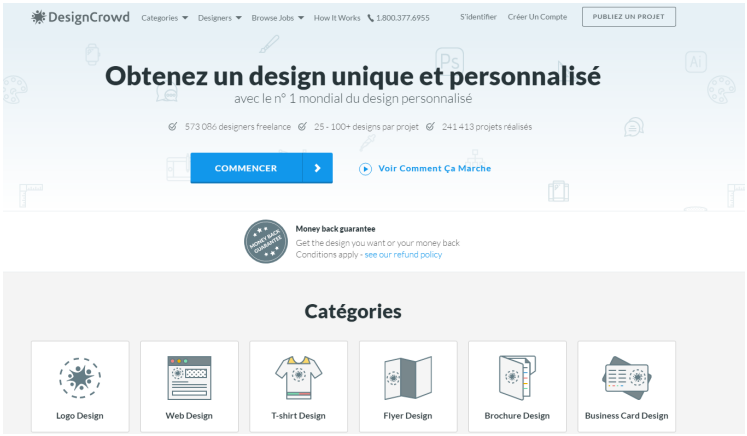

Figure 1.9: Inducement Prize Contests: DesignCrowd Website.

[1]DesignCrowd, https://www.designcrowd.fr/

7 Basic Requirements for Crowdsourcing

Crowdsourcing seems to be a remarkable solution for many fields, especially for some Natural Language Processing problems. But, it must be approached with care since deceptive or incorrect results can also easily be obtained from crowdsourcing. Some issues should be kept in mind to prevent this.

- *Giving the crowd too much information* that can influence their decisions.
- *A crowd that is too homogeneous* will not give a superior result. Oinas-Kukkonen [16] has found that the best decisions come when there is disagreement and competition within the crowd.
- *Having too much communication with the crowd* give rise to the possibility of imitation. If participants are given access to the opinions of other workers, they may be influenced by them.
- *Requesting prerequisites from the crowd* which are supposed to have some local knowledge. It is not evident that everyone who responds to a call has that knowledge.
- *Maintaining crowd motivation*, Why should someone participate in a task? Are they learning something, playing a game, being rewarded? There should be some reason for an individual to not only sign up to work on a task but also want to continue to work on it [10].
- *Presenting a reasonable expectation of workload.* If members of the crowd are lead to believe that there is less work than what is actually expected, especially in the case of remunerated work, they will quit the task and recommend

to others.

- *Conducting quality control.* It is meant to weed out the work of *poor workers* (who have good intentions, but their work is not of good quality) and *malicious workers* (those who randomly enter answers or automated bots).

8 Speech Crowdsourcing

Speech crowdsourcing is a special task of crowdsourcing, where tasks deal with speeches via audio or video data. Thus, when approaching and designing a speech crowdsourcing task. The requester has to assure that the audio works and also take consideration to other points as: payment, choice of platform, prequalification, native language of workers, and task complexity.

In this section, we point out some of those considerations:

8.1 Hearing and Being Heard over the Web

It's important to ensure that the worker can hear the speech signal correctly and/or that the speech signal can be correctly recorded.

There are still several possible reasons that audio might not play or be acquired. Some causes of audio problems are:

- Worker not wearing the headset.
- Headset not plugged in.
- Sound levels too high or too low.
- High levels of ambient noise.
- Failure to correctly follow instructions.

Thus the design of task have to prevent such problems.

8.2 Prequalification

Prequalification is important phase to detect the non qualified workers such as: Bots, malicious individuals and non native speakers. It can not be assumed that every native speaker of a language is capable of performing any type of linguistic task in their language.

There are two types of prequalifications:

- One that is general and is designed to remove malicious workers, using some sort of past approval rating (that will only work, of course, for those who use a platform that has this type of service).
- Another, is to give the worker a small sample of the task to perform.

8.3 Native Language of the Workers

Some platform such as mturk and cf offer a geo-location services to finding out the country of the worker, but just knowing the location does not help filter for the worker's native language.

Lane et al. [17] ask workers to select their native language at the beginning. Kunath and Weinberger [18] asked not only for the native language of the worker but also how well the worker knew other languages.

8.4 The Complexity of the Task

A task designer should take an impartial look while the creation of task in order to determine its complexity:

- What should the worker do/follow to accomplish this task?
- Is it possible to divide the work into separate tasks?
- What is specific instructions to accomplish this task?

8.5 Quality Control

Quality control often compares the performance of the crowd to that of the experts. The literature in speech processing has shown thus far that the quality of the work of crowd is starting to approach that of the experts.

Quality control can be performed at several different stages in the crowdsourcing process. First, **before** the worker starts on the task, where prequalification requirements implemented such as work history, native language, and/or success on a prequalification task. Then, Online filtering can be used **during** the task to assess the quality of the workers' production. Finally, quality control can be carried out **after** a worker has submitted all of their work.

9 Crowdsourcing platforms

A crowdsourcing platform is a soft system used by requesters to publish tasks and by crowd workers to complete tasks. It is commonly a Web application that provides functionalities for task and crowd management. For requesters, the choice of a

crowdsourcing platform may depend on the nature of the projects they want to crowdsource and the incentive they are willing to provide to the workers [12].

General purpose platforms usually specializes in handling simple tasks or microtasks. As seen above, microtasks are tasks that require minimal time and cognitive effort but when combined can result in major accomplishments.

Crowdsourcing platforms are classified as either paid or unpaid. In paid platforms, requesters can deal with large population of workers around the world to accomplish tasks in a fraction of the time and money of more traditional methods. On the other hand, requesters rely on volunteers or other crowd gathering techniques in unpaid platforms. In the following subsections, we will discuss examples of paid and unpaid platforms.

9.1 Paid Platforms

Amazon Mechanical Turk (Mturk)[1] is one of the sites of Amazon Web Services. It was launched publicly in 2005. Mturk enabling multiple applications for its users such as: missing persons searches, social science experiments, and artistic/educational research. Its requesters must provide a billing address in the USA, Australia, Canada or the UK in order to submit a request for tasks to be completed through the Mturk platform. Figure 1.10 illustrates interface of Mturk platform.

CrowdFlower[2] is a crowdsourcing company based in San Francisco, United States. It was founded in 2007. Typical

[1]Mturk, https://www.mturk.com/mturk/welcome
[2]CrowdFlower, https://www.crowdflower.com/

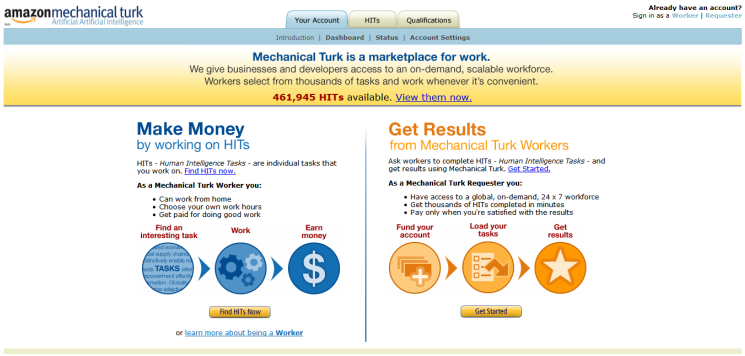

Figure 1.10: Mturk Platform Interface.

users of CF are data scientists who use the software to create training data to build models and train machine learning algorithms. CF enabling many applications for its users such as: sentiment analysis, categorization, and data collection/enhancement. Figure 1.11 represents Interface of CrowdFlower platform.

Figure 1.11: CrowdFlower Platform Interface.

9.2 Unpaid Platforms

PyBossa[1] is a free, open-source framework for crowdsourcing, developed by Sci-Fabric [2]. It is equipped with features that help in the management of tasks and analysis of results. Requesters have the option to deploy their instance of PyBossa or publish tasks in Crowdcrafting. Figure 1.12 reports Pybossa platform.

Figure 1.12: Pybossa Platform Interface.

Crowdcrafting[3] is a Web-based service that invites volunteers to contribute to scientific projects developed by citizens, professionals or institutions that need help to solve problems, analyze data or complete challenging tasks that cant be done by machines alone, but require human intelligence. The platform is 100% open source – that is its software is developed and distributed freely – and 100% open-science, making scientific research accessible to everyone. Figure 1.13 represents Crowdcrafting platform.

Zooniverse[4] platform of crowdsourcing describes itself as the world's largest and most popular platform for people powered

[1] Pybossa, http://pybossa.com/

[2] Sci-Fabric, a company that develops open source software for crowdsourcing research. https://scifabric.com/

[3] Crowdcrafting, https://crowdcrafting.org/

[4] Zooniverse, https://www.zooniverse.org/

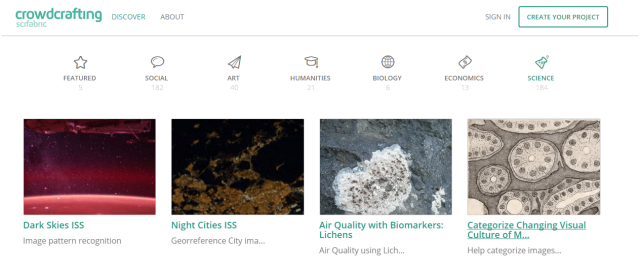

Figure 1.13: crowdcrafting Platform.

research, Zooniverse hosts dozens of projects, where anyone can participate in crowdsourced scientific research. Additionally, anyone can create a project in Zooniverse and tap on its community of volunteers. Figure 1.14 illustrates Zooniverse platform.

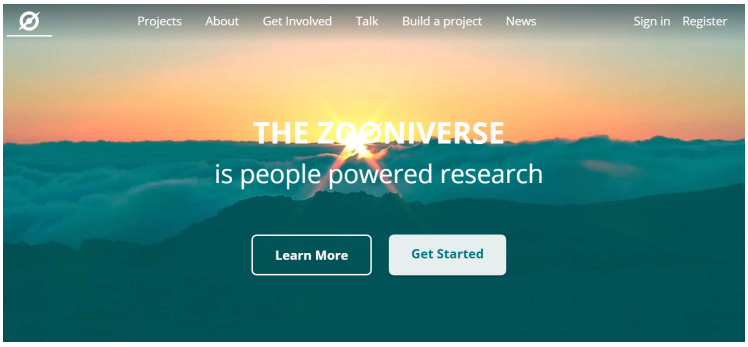

Figure 1.14: Zooniverse Platform.

Table 1.1 represents the functional differences between platforms.

10 Crowdsourcing and Social Media

Social Media are a powerful Networks where people can: exchange information, express ideas, share opinions, create content, and

Platform	Mturk	CF	Pybossa	Crowdcrafting	Zooniverse
Location	owned by Amazon, USA	San Francisco, USA	Madrid, Spain	Unknown	Chicago, USA and Oxford, England
Created	2005	2007	2011	2011	2009
Allowed Data Formats	Visual, Text, Audio	Visual, Text, Audio	Visual, Text, Audio	Visual, Text, Audio	Visual, Audio
Allows Anonymous Users	No	No	Yes	Yes	Yes
Data Hosting	Project's own server		Project's own server	Crowdcrafting Server	Project's own server

Table 1.1: Paid/Unpaid Crowdsourcing Platforms.

seek out new knowledge [19]. Figure 1.15 report some social media Networks.

Figure 1.15: Some popular Applications of Social Media.

Internet played a role in the evolution of the word, while social media is transforming the way we think of crowdsourcing and

will continue to do so as the benefits of using social media to crowdsource become more well-known.

Social media is becoming an essential component to crowdsourcing as it allows organizations to reach a wider audience faster, cheaper and more efficiently than ever before [20].

Current crowdsourcing campaigns almost always use social media to obtain a higher number of contributions, in theory leading to a better quality idea, service or whatever the desired end-product might be.

11 Crowdsoucing Corpus Annotation Process

NLP can benefit from the power of crowdsourcing to annotate large corpora. In order to make annotation scalable and of high quality, while ensuring sufficient annotator variety. Sabou et al. [11] described the annotation process for crowdsourced corpus, where can be applied in all crowdsourcing genres.

The corpus annotation process of crowdsourcing can be broken down into four main stages: *Project Definition*, *Data Preparation*, *Project Execution*, and *Data Evaluation and Aggregation*. Figure 1.16 illustrates the process of crowdsourcing annotation.

11.1 Project Definition

The first step is to select NLP problem and crowdsourcing genre. There are three major genres of crowdsourcing. In *Games With A Purpose* (GWAP) where the main motivator is fun. In Mturk where the main motivator is profit. *Wisdom of the Crowds* (WotC) is another major genre for crowdsourcing. In the case

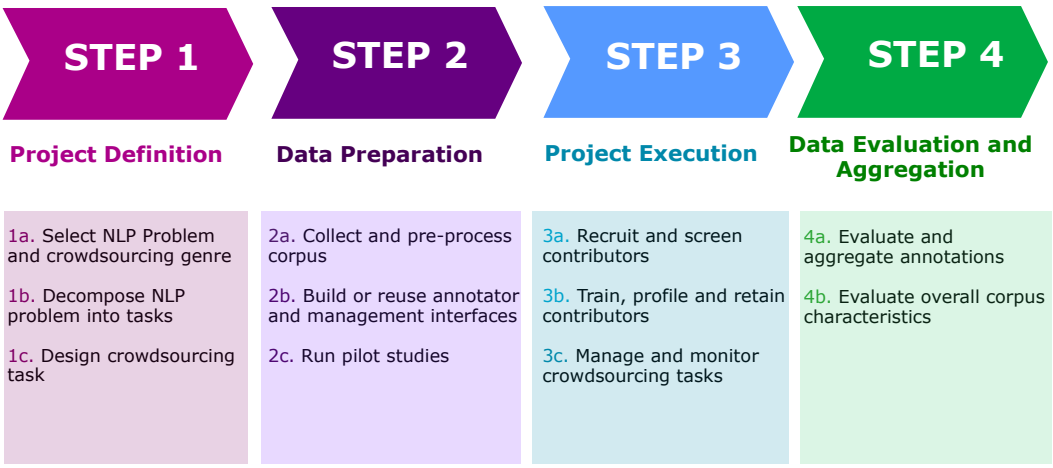

Figure 1.16: Crowdsourcing corpus annotation process.

of WotC applications, such as review sites and Wikipedia, where the community benefits as a whole as more users contribute [21].

Second step, the chosen NLP problem needs to be decomposed into a set of simple crowdsourcing tasks, which can be understood and carried out by non-experts with minimal training and compact guidelines.

Final step is to design crowdsourcing task that must be simple and intuitive, where a simpler design without too much variance tends to lead to better results.

11.2 Data Preparation

In this stage, user interfaces need to be designed, the data collected and prepared, and run pilot studies.

Interface design can be a major task. There are two kinds of user interface tool. First, *Acquisition interfaces* are designed for and used by the non-expert human contributors to perform crowdsourcing tasks. Second, *Management interfaces* are required by the person running the crowdsourcing project, in order to allow them to monitor progress, assess quality, and manage contributors.

Data processing may involve preliminary annotation with an automated tool or filtering objectionable content. The corrected annotations generated by the crowd can then be used to improve the application.

There are proven benefits to performing a small scale pilot for testing the task. It requires that the complete application is in place, and therefore it is performed in the "Preparation" rather than "Project definition" step. If a pilot is not successful however, the project definition step would need to be revisited.

11.3 Project Execution

This is the main phase of each crowdsourcing project. It consists of three kinds: recruit contributors, train/retain contributors, and manage/monitor crowdsourcing tasks.

Contributor recruitment consists in a set of primarily advertising activities to attract contributors to the crowdsourcing project. Most NLP projects recruit their contributors from marketplaces that offer a large and varied worker base.

Attracting and retaining a large number of contributors is key to the success of any crowdsourcing system. Therefore, a core challenge of all crowdsourcing approaches is how to motivate

contributors to participate.

Manage crowdsourcing tasks by filtering workers prior to the task (based on e.g. prior performance, geographic origin, and initial training) to improve quality. Extensive screening can however lead to slower task completion, so filtering through task design is preferred to filtering through crowd characteristics.

11.4 Data Evaluation and Aggregation

This stage is required in order to make acquisition tasks reproducible and therefore scalable, and to ensure good corpus quality.

Contributor aggregation primarily relies on majority voting or average computation based algorithms, while the evaluation of the resulting corpus is usually performed by computing inter annotator agreement (IAA) within crowd-workers and/or with a baseline resource provided by an expert; by task-centric evaluation as well as by Precision, Recall and F-measure metrics.

Conclusion

In this chapter, we have seen the generalities about crowdsourcing, described the crowdsourcing paradigm by showing its process, common types of crowdsourcing, and point out the genre of crowdsourcing with some platform examples. In the next Chapter, we will present the crowdsourcing for Arabic Natural Language Processing state of the art.

Chapter 2

Crowdsourcing for Arabic Natural Language Processing

1 Natural Language Processing

Natural Language Processing (NLP) can be defined as the ability of a machine to analyze, understand, and generate human speech [2]. Dataset for Natural Language Processing are referred to as corpus. They play an essential role in NLP research as well as a wide range of linguistic investigations. They provide a material basis and a test bed for building and validating NLP systems.

1.1 What is a Corpus?

A corpus[1] is a large collection of data, such as: texts, audios, and videos, nowadays usually electronically stored and processed. They are used to do statistical analysis and hypothesis testing, checking, validating linguistic rules within a specific language territory.

A corpus can be composed of written language, spoken language or both [22]. Written texts in corpora might be drawn from books, newspapers, or magazines that have been scanned or downloaded electronically. People build corpora of different sizes for specific reasons. For example, a very large corpus would be required to help in the preparation of a dictionary. Spoken corpus is a large collection of audio recordings of spoken language. Most speech corpora also have additional text files containing transcriptions of the words spoken and the time each word occurred in the recording [23].

[1] The plural form of corpus is corpora.

1.2 Corpus Annotation

Corpus annotation, sometimes called "tagging", is the practice of adding interpretative linguistic information to a corpus [24]. Part Of Speech (POS) tagging, also called grammatical tagging, is the commonest form of corpus annotation, in which information about each word is part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of tags.

In addition of POS tagging, there are other types of annotation, corresponding to different levels of linguistic analysis of a corpus or text such as:

- Phonetic Annotation: Adding information about how a word in a spoken corpus was pronounced.
- Semantic Annotation: Adding information about the semantic category of words.
- Pragmatic Annotation: Adding information about the types of speech act (or dialogue act) that occur in a spoken dialogue.

2 Arabic Natural Language Processing

Arabic is the largest member of the Semitic language family and it is spoken by nearly 500 million people worldwide. It is one of the six official languages. Despite its cultural, religious, and political significance, Arabic has received comparatively little attention in modern computational linguistics [25].

Recently, the Arabic language has become the focus of an increasing number of projects in NLP and computational linguistics (CL). Arabic NLP (ANLP) deals with developing tools

and techniques that deliver state-of-the-art performance in a variety of Arabic language processing tasks [26].

3 Crowdsourcing in Arabic Natural Language Processing

For many worldwide researchers and institutions, crowdsourcing has become a standard method for accessing large number of participants who have a number of skill sets that can be utilized for collection and annotation of data in various speech and language processing studies.

In what follows, we review some works on crowdsourcing usage for Arabic NLP such as corpus annotation, transcription, and translation.

3.1 Crowdsourcing for Arabic Annotation

Crowdsourcing has been used effectively for text [27] and speech [28] Arabic annotation.

Zaghouani and Dukes [27] compare two Arabic annotations which are POS tagging and grammatical case-ending by using Mturk platform. They used the gold standard analysis of the Quran, annotated by Arabic experts, in order to measure the quality of the annotation in the experiment. In POS tagging, annotators were asked to choose a correct word-category for each Arabic word, such as noun, verb, adjective, pronoun, particle and other/unknown, from the first 200 words from the chapter 23 of the Quran. In the second annotation task, annotators were asked to identify grammatical case endings, such as: nominative

case, genitive case, or accusative case, where the first 100 words from chapter 23 of the Quran are selected. The total number of interested Turkers in both experiments combined was 137. They are paid 0.01$ for each task. The whole experiments was carried over a period of one month. As results, they claim that annotating Arabic grammatical case is harder than POS tagging. In order to ensure quality control by non-native speakers of Arabic, the Turkers have to pass a simple Arabic qualification test, which requires a basic understanding of the Arabic language. They excluded any results of participants who failed in the qualification test from the reports.

Wray and Ali [28] used CF platorm to create a labeled multi-dialectal speech of Al Jazeera corpus with 47,696 segments of unknown dialects representing 404 hours of speech. Their task was restricted to users in the Arab world. All directions for the task were written in Modern Standard Arabic. Contributors were directed to listen to the short speech segments and determine which dialect they thought the speaker was speaking. Dialect judgment was answered by a seven-way forced choice between MSA, Levantine Arabic, Egyptian Arabic, North African Arabic, Gulf Arabic, non-Arabic speech, and nonspeech. Compensation for this task was USD 0.03 per page of 10 items.

The whole experiments was carried over three weeks, with 2,053 users contributed to the labeling task, costing a total of USD 971. They follow two way to ensure quality control. First, Quiz Mode offered by CF optional where contributors have to answer five gold standard items before entering the main portion of the task. Second, for every five items, contributors were presented with a gold standard item. The participation of contributors can be ended if their task accuracy less than 65%.

3.2 Crowdsourcing for Arabic Transcription

Samantha et al. [29] investigated different approaches using crowdsourcing for transcriptions of Dialectal Arabic speech with automatic quality control using CF. They submitted several tasks of Egyptian dialect audio totaling approximately 10 min of speech and selected the High Speed option, which allows every user in the selected country to participate. They collected five transcripts for each item. In average, the task was completed after 3 hours from its launch and the total cost was 7 USD.

Audio for the transcription task was taken from debate and news programs uploaded to Al Jazeera's website between June 2014 and January 2015.

A total of 149 users participated in the transcription tasks of Egyptien audios. The average Word Error Rate (WER) for each user was calculated based on comparing each transcript to the four other user-provided transcripts for each item.

Concerning the validation, Samantha et al. [29] modify the gold standard mechanism where they compared the transcriptions of the workers with automatic speech recognition system results.

3.3 Crowdsourcing for Arabic-English Translation

In order to support research in speech translation, Kumar et al. [30] introduce the Callhome Egyptian Arabic-English Speech Translation Corpus by using crowdsourcing techniques. Callhome corpus consists of 160 telephone speech conversations between 5-30 minutes of Egyptian native speakers. In addition, for each of the conversations have a transcript file that covers 5-10 minutes segment. The total number of Arabic transcripts file

is 35 842. They used the crowdsourcing platform, Mturk to obtain translations. For each Arabic transcripts, there are four English translations. As result, 838 translators participated in this process, producing 143,568 translations in English.

Kumar et al. [30] have used the following quality control mechanisms to ensure that the quality of the translations was acceptable:

- They obtained for each utterance a translations from Google Translate. The translation will be rejected if it had a small edit distance from the translation obtained via Google Translate.
- To prevent users from using online translation services to cut and paste translations. They present the utterance as an image rather than as text.
- Manually translated gold standard segments were inserted into their dataset. The translation of the worker were rejected if it was not similar to the gold standard translations.
- Gathered self-reported geographical and language information for each of their contributors on Mturk. Higher preference was given to trusted Arabic speakers that they have worked with on other translation tasks.

3.4 Arabic Game With Purpose

Hakouz et al. [31] present Lahajet, GWAP for crowdsourcing classifications of different varieties of Dialectal Arabic in multi-dialectal audio. The data was obtained from Al-Jazeera Arabic corpus that we mentioned it above in [28]. Lahajet consist

Work	NLP Gender	Platform	Corpus	# Workers	Cost	Duration
Zaghouani and Dukes [27]	Text Annotation	Mturk	Quran	137	0.01$ (per task)	one month
Wray and Ali [28]	Speech Annotation	CF	Al Jazeera	2,053	0.03 USD (per task)	Three weeks
Samantha et al. [29]	Speech Transcription	CF	Al Jazeera	149	0.05 USD per 5 tasks	-
Kumar et al. [30]	Text Translation	Mturk	Callhome	838	-	-
Hakouz et al. [31]	Speech Annotation	GWAP	Al Jazeera	-	-	-

Table 2.1: Crowdsourcing for Arabic Natural Language Processing Tasks.

of principle that players listen to short audio clips and select a character representing the dialect they have heard by moving avatars of characters who represent the four major regional DA groups (Egyptian, North African, Gulf, Levantine). Character movement, point rewards, and obstacles are implemented to ensure player engagement and interest.

3.5 Discussion

Table 2.1 summarizes the main works that have used crowdsourcing for Arabic NLP. We observed that the most reviewed crowdsourcing applications used for Arabic NLP tasks

are done with fees by using the Mturk or CF platforms. In order to ensure the quality control, each researcher followed a specific mechanism, such as "test question", or unique proceedings for their work.

Conclusion

In this chapter, we have reviewed some related work about ANLP and crowdcourcing usage. These applications have touched Arabic dialects annotation, transcription, translation, and GWAP.

Chapter 3

Our Crowdsourcing based Annotation

In this chapter, first, we give a brief overview of the targeted corpus, which is KALAM'DZ speech corpus. Then, we describe in details our approach to use free crowdsourcing to validate annotation of this corpus. Our implementation and deployed tools are described and their use is justified. Finally, we show the results and analyze the outcomes.

1 Targeted Corpus

KALAM'DZ corpus is developed by Bougrine et al. [7] which covers eight major Arabic dialects of Algeria. This corpus is collected from the Web sources namely YouTube, Online Radio and TVs. The size of the corpus is about 104 hours with 4881 speakers. The most of dialect annotations are provided from the related metadata of the Web sources when they exist. These metadata are namely the title, category, the location from where the source is posted, and the identity of the publisher. Table 3.1 reports the distribution of KALAM'DZ corpus by source.

In what follows, we give a brief overview on Algerian dialects features. Algeria is a large country, administratively divided into 48 departments. Its first official language is the Modern Standard Arabic (MSA). However, Algerian dialects are widely the predominant means of communication. Algerian Arabic dialects are the major sub-dialects as they are spoken by 75% to 80% of the population. The Algerian dialect is known as Daridjah (الدارجة) to its speakers.

Algerian Arabic dialects are resulted from two Arabization processes due to the expansion of Islam in the 7^{th} and 11^{th} centuries, which lead to the appropriation of Arabic language by

Sub-Dialect	# Speakers	Web sources (h)			
		Algerian Tv	Local Radios	On YouTube	Total (h)
Hilālī-Saharan	1338	12.7	16.5	-	29.4
Hilālī-Tellian	605	3.6	-	-	3.6
High-plains	297	2.0	-	-	2.0
Ma'qilian	421	4.1	-	20.7	24.8
Sulaymite	914	6.7	-	6.9	13.6
Algiers Blanks	723	5.1	-	16.1	21.2
Sahel-Tell	447	3.1	6.0	-	9.1
Pre-Hilālī	136	0.7	-	-	0.7
Global	4881	38.2	22.5	43.7	104.4

Table 3.1: KALAM'DZ Corpus: Distribution by Source.

the Berber population.

According to these both Arabization processes, it is showed that Algerian Arabic dialects can be divided into two major groups [7]:

1. Pre-Hilālī (ما قبل الهلالين) dialect is spoken in cities: Tlemcen, Constantine and their rural surroundings.
2. Bedouin dialect that is divided into four distinct dialects:
   (a) Sulaymite (الشرق الجزائري) dialect is connected with Tunisian Bedouin dialects.
   (b) Ma'qilian (الغرب الجزائري) dialect is connected with Moroccan Bedouin dialects.

(c) Hilālī dialect contains three nomadic sub-dialects.

i. Hilālī-Saharan (الصحراء) dialect: It covers the totality of the Sahara of Algeria.

ii. Hilālī-Tellian (التل الهلالي) dialect: Its speakers occupy a large part of the Tell of Algeria,

iii. High plains of Constantine (الهضاب العليا) covers the north of Hodna region to Seybouse river.

(d) Completely Bedouin dialect that covers Algiers' Blanks, and some of its near sea coast cities. It can be divided into two sub-dialects: Algiers-Blanks (العاصمة و ضواحيها) and Sahel-Tell (التل الساحلية).

According to these information, we have created a map of Algerian dialects in a vectorial format. It will help the contributors to validate dialect annotation of the corpus. Figure 3.1 reports the map of Algerian dialects.

2 Crowdsourcing Dialect Annotation of Kalam'DZ Corpus

Despite that Kalam'DZ corpus size is large, the most available annotations concern the spoken dialect. This knowledge is provided and extracted from the related metadata of the Web-source, gathered/collected when it exists. The results of those annotations need a validation process by experts in order to be taken into account. In fact, these annotations are assigned semi-automatically relying on the information inherent in the title of the source, its subject, and the geographic location of the source.

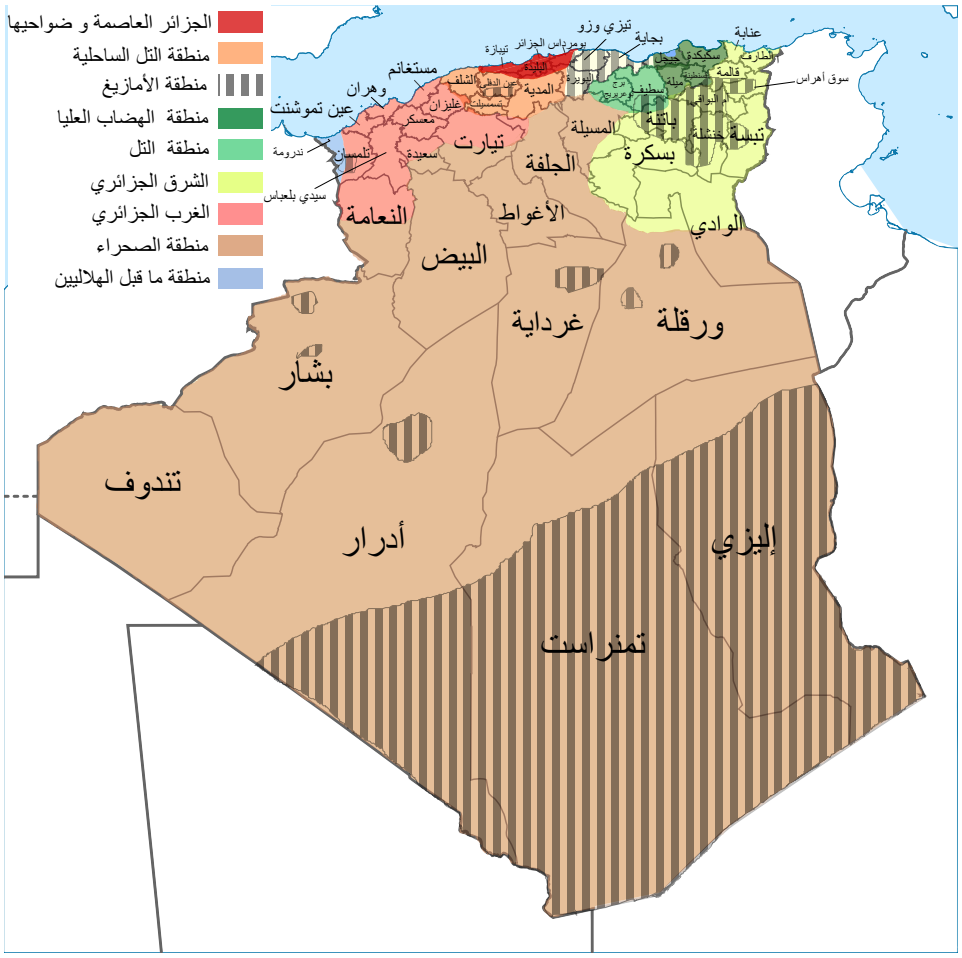

Figure 3.1: Geographic distribution of Algerian Dialects.

However, such validation process can take more time and be costly. For that crowdsourcing is the best way to validate those annotations. In what follows, we will describe our investigation aiming to build a crowdsourcing-based system.

2.1 General Overview of the Process

In order to make annotation scalable and of high quality, we have followed the state of the art of crowdsourcing engineering process

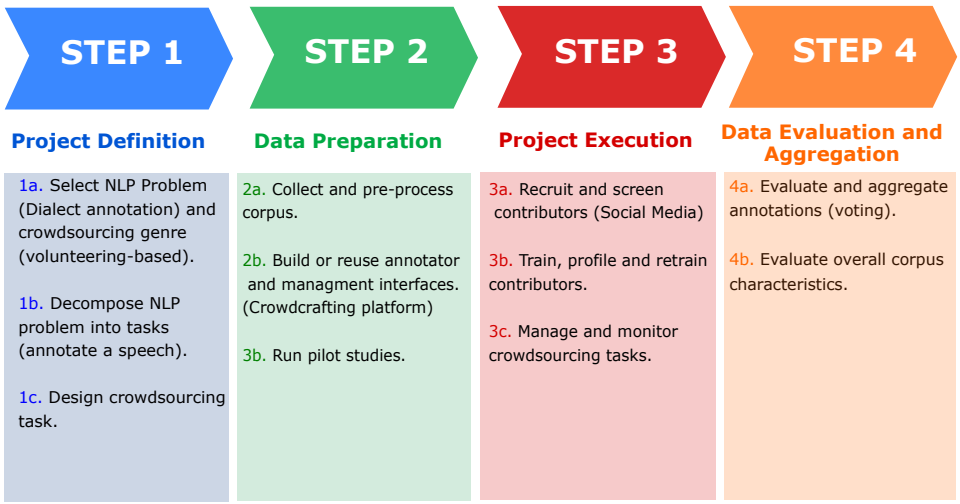

Figure 3.2: Crowdsourcing Corpus Annotation Process.

defined by Sabou et al. [11] described previously in Chapter 1.

This crowdsourcing corpus annotation process suggests designing the system in four stages. First, the project definition stage may be divided into three main steps: select the NLP problem, crowdsourcing genre, decomposed into tasks, and design crowdsourcing task. Second, in data preparation stage, the requester must build or reuse project and manage interfaces. Then, the project execution step is the main phase of each crowdsourcing project, where the contributors must be recruited and retrained to participate in the project with some motivation. Then, the project owner should manage/monitor crowdsourcing tasks. Finally, requester must evaluate and aggregate annotations. Figure 3.2 recalls the main step of this process.

To implement this process, some tools are selected and data processing are performed. Let us start by explaining the chosen crowdsourcing platform.

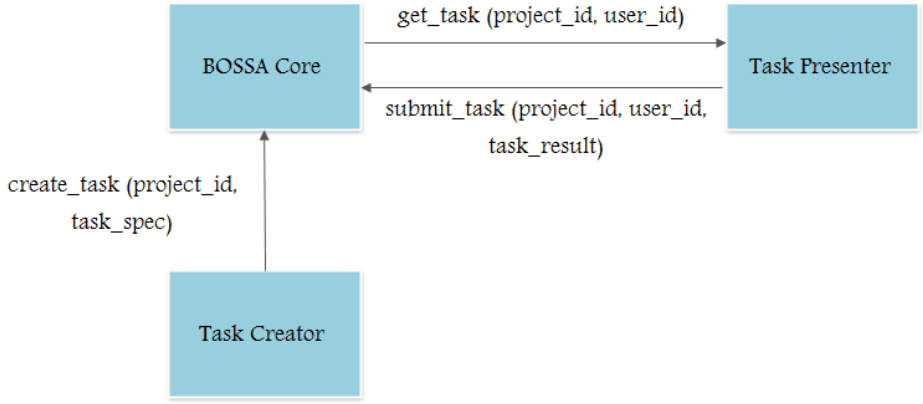

Figure 3.3: PyBossa System Modules and Functions.

2.2 Used Tools

In fact, the corpus annotation was conducted through a crowdsourcing application implemented on Crowdcrafting platform [1] under the name " Dz كـلام" project.

2.2.1 Crowdcrafting Platform

CrowdCrafting is a free and open-source crowdsourcing platform based on PyBossa[2]. This latter is a Web service with a database with simple architecture. The database is a PostgreSQL one which is an object-relational database. The back-end is written in Python while the front-end features deployed HTML5, CSS3 and JavaScript for modern Web browsers.

CrowdCrafting platform enables people to create/run projects that utilize on-line assistance in performing tasks that require

[1] Crowdcrafting is a community platform where NLP-based applications can be deployed, https://crowdcrafting.org/

[2] PyBossa, http://pybossa.com/

human cognition such as transcription, image classification, and more.

In Crowdcrafting, a project has a set of tasks. Each task is a problem in itself, that volunteers will try to solve. Crowdcrafting distributes the tasks among the volunteers being sure that each participant can only save one answer per task. By default, each task is reviewed by 30 different persons, but this threshold can be modified.

Figure 3.3 gives an overview of how a PyBossa system functions. A Crowdcrafting project, as in PyBossa, has two main components that allow requester to customize his own interfaces:

- *Task Presenter*: An html document where the javascript will load the task data. It is responsible for presenting tasks to user in a convenient user interface [32].
- *Task Creator*: Usually, it is a script that will upload the tasks for the project into the Crowdcrafting server.

2.3 Project Definition

Our problem is to annotate a part of KALAM'DZ corpus by using unpaid (volunteering-based) crowdsourcing.

Two alternatives of tasks can be defined: *i)* The contributor will be asked to detect the spoken dialect using a proposed multi-choice. *ii)* or he will be asked to validate the available semi-automatic dialect annotation. For the sake of facility and to avoid contributor workload, we have chosen the second alternative.

We have designed a form containing a question about speech segment, buttons for response, an Algerian dialects map for

facilitating the contributors work, and links to the social networks accounts (Facebook, Twitter) to keep track of project. Our task was restricted to users in the Algerian regions for that the form was written in Modern Standard Arabic.

2.4 Data Preparation

In Data Preparation stage, we have designed and implemented both user and management interfaces. Then, we have prepared the tasks, and run pilot studies. The duration of our preparation stage has taken almost one month and half to be completed.

In order to allow sharing our set of tasks with the contributor, we have used another tool called SoundCloud[1]. In fact, as a good practice, PyBossa propose to use a SoundCloud based template[2] which shows how to solve crowdsourcing sound recognition problem. By using SoundCloud, we have the possibility of using SoundCloud API.

2.4.1 User Interface Design

Our user interface is done using HTML, JavaScript and Cascading Style Sheets (CSS) libraries. In order to design and customize a pretty interface, we have used BootStrap [3] design framework. Figure 3.4 illustrates the user interface during annotation.

The user interface contains an Algerian dialects map that helps the contributors working, a form containing short speech segment, a question with buttons for response, and a progress bar for indicating the progress/completion of tasks by the user.

[1]SoundCloud, https://soundcloud.com/
[2]SoundCloud App, http://crowdcrafting.org/project/soundcloud
[3]Bootstrap, http://getbootstrap.com/

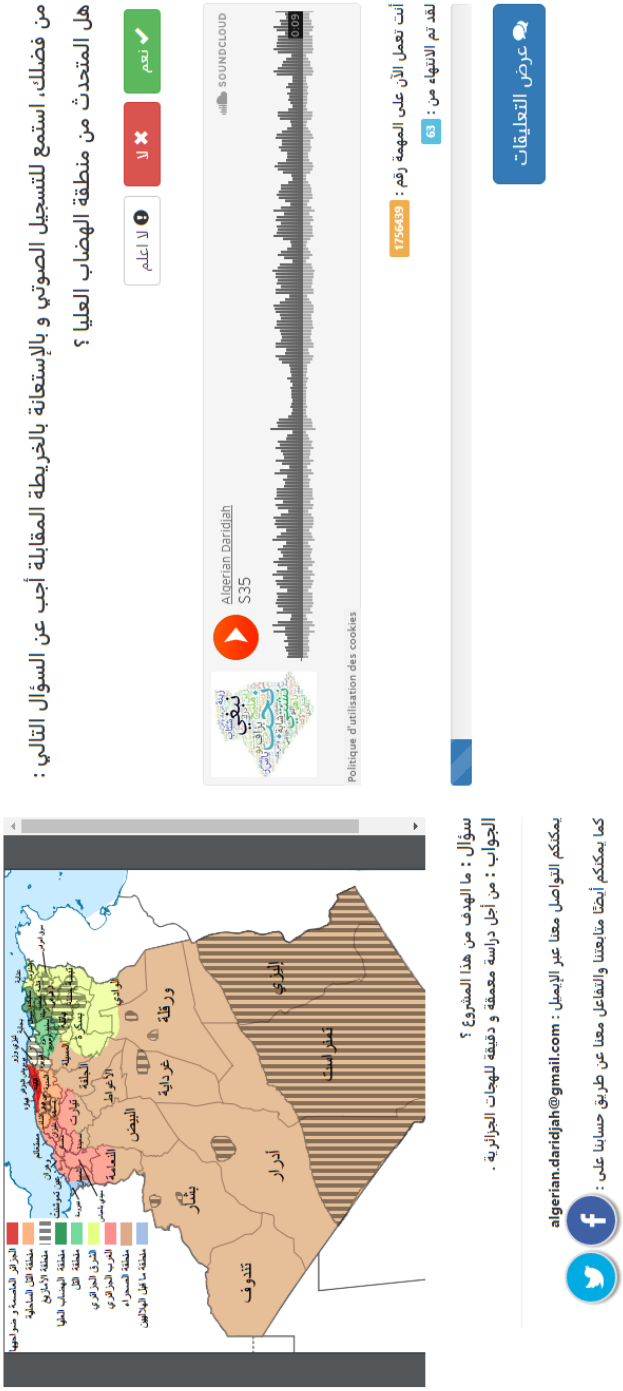

Figure 3.4: User Interface of our Project.

The Algerian dialects map was created by using free Inkscape [1] tool where the formated image is a vectorial form and allowed exportation format. The map is added to interface using an external application (embed tag) which provided zooming and panning capabilities to the image. In fact, the image could be zoomed in or out by clicking the respective buttons on the toolbar, by using the '+' or '-' keystrokes.

A form containing a short speech segments where contributors are directed to listen and validate the spoken dialect.

The identification of the current task is placed at the bottom of the short speech segment. An indicator area provides feedback for the user regarding status of his/her completed task using a progress bar.

2.4.2 Management Interface

For the management interface, crowdcrafting server is the responsible of monitoring progress and managing contributors. When the contributor clicks to one of the response buttons, a new task will be load automatically. In addition, users could leave the annotation at any point.

2.5 Tasks Preparation

Now that we have designed the user interface, we add the tasks to our project. Let us mention that our Crowdcrafting project delivers the tasks for the contributors, without conditions on their

[1] Inkscape is professional quality vector graphics software which runs on Windows and GNU/Linux.

status profile of connexion (authenticated or anonymous), and store the submitted answers in their database.

From KALAM'DZ corpus, we selected a set of a sample. In fact, we haven't dealt with the whole corpus for timing sake. It represents each dialect by at least two hours, totalling eleven hours. These utterances are used to create our 1011 micro-tasks. Table 3.2 shows general information about these tasks.

Dialect	#Utterances	Total Time	Average Time (s)
Sulaymite (الشرق الجزائري)	222	2h 12 mn	36
Ma'qilian (الغرب الجزائري)	126	1h 20 mn	38
Hilālī-Saharan (الصحراء)	126	1h 18 mn	38
Hilali-Tellian (التل)	112	53 mn	28
High plains of Constantine (الهضاب العليا)	126	1h 20 mn	38
Algiers-Blanks (العاصمة و ضواحيها)	126	1h 41mn	50
Sahel-Tell (التل الساحلية)	97	59 mn	36
Pre-Hilālī (ما قبل الهلالين)	76	41 mn	32
Total	1011	10h 24 mn	-

Table 3.2: Information About Project Tasks.

Let us mention that "Sahel-Tell" dialect is less represented in the task set as it is spoken by a small population compared with the other dialects.

In order to share those utterances with Crowdcrafting, we have created a Google Drive Spreadsheet file that contain information which need to be processed by the volunteers. It holds in 1011 lines (number of tasks) and three columns: question, embed (soundcloud[1] clip), and public link to a media file uploaded to SoundCloud website that needs to be processed. Figure 3.5 reports a part of our Google Drive Spreadsheet file.

Knowing that SoundCloud Website limited each user to three hours, as storage space, to upload their Sounds. We have used four accounts to upload the whole part of KALAM'DZ corpus with almost eleven hours.

question	embed	audio_url
هل المتحدث من منطقة الصحراء ؟	<iframe width="100%" height="20%" scrolling="no"	https://soundcloud.com/us
هل المتحدث من منطقة الهضاب العليا ؟	<iframe width="100%" height="20%" scrolling="no"	https://soundcloud.com/alg
هل المتحدث من منطقة الصحراء ؟	<iframe width="100%" height="20%" scrolling="no"	https://soundcloud.com/us
هل المتحدث من منطقة الهضاب العليا ؟	<iframe width="100%" height="20%" scrolling="no"	https://soundcloud.com/alg
هل المتحدث من منطقة التل الساحلية ؟	<iframe width="100%" height="20%" scrolling="no"	https://soundcloud.com/da

Figure 3.5: A part of our Google Drive Spreadsheet file.

2.6 Project Execution

Contributor recruitment consists in a set of primarily advertising activities to attract contributors to the crowdsourcing project. In our case, we have relied on two mechanisms. Fist, we have used

[1]SoundCloud, https://soundcloud.com, is an online audio distribution platform based in Berlin, Germany, that enables its users to upload, record, promote, and share their originally-created sounds.

Social Media power by creating our Facebook page and Twitter account. In addition, we have done an open call by posting in all Facebook groups of Algerian universities by using our Facebook page. As second mechanisms, we have directly invited people to participate by using mailing lists that contain students, some Algerian NLP experts, and academic researchers. As filtering way, we advertise the project only in Algerian communities.

Attracting and retaining a large number of contributors is the key to the success of any crowdsourcing system, especially when it is unpaid crowdsourcing one. The contributor find all instructions that should be followed to do correctly the task in email messages and Facebook posts. As an unpaid crowdsourcing frame, the main motivation is considered as *community help* to survey the Algerian dialects.

2.7 Data Evaluation and Aggregation

Each task is presented for annotation to five users, where it will be marked completed if it is annotated by five unique users. This multi request is performed to gather data such that a consensus could be drawn between the responses of the users. Contributors are asked to listen only as long as necessary to determine the dialect being spoken. In data aggregation, we have selected the majority voting technique.

3 Experiments and Results

Citizen scientists were requested for voluntary participation via email and social networks posts. The application was hosted for

a period of 18 days during which more than 200 users registered for participation and volunteered for the tasks assigned.

The "Dz كـلام" project was launched on 17 April by inviting contributors to participate in the project. The results start appearing from the first day of launching KALAM'DZ. Crowdcrafting system allows multiple statistics for following the progress of the project. Table 3.3 gives some statistics on the project execution.

Launching date	17 April
Termination date	04 May
Total duration	17 Days
Number of crafter	218
Number of answered task	5187

Table 3.3: Statistics about the Project Execution.

In order to analyse the results of the Crowdcrafting annotator project, we try to reply to the following questions and assertions. The aim is to give some good practices to respect when designing a free crowdsourcing framework for dialect annotation:

1. Is free crowdsourcing efficient to annotate corpora?
   - How about needed time compared with paid annotation?
   - How is the annotation quality?
2. What are the mechanisms that increase the free contributor motivation to actively participate?
3. What can conclude about the validation of Algerian dialect annotation ?

3.1 How about Needed Time ?

The total completed tasks per day are illustrated in Figure 3.6. These results show that the number of completed tasks augment gradually in fourteen days after the start of "Dz كـلام" project, the number of completed tasks is 268 tasks which is about 1.340 answers. Big part of work was finished on the second day of May. The total period of project was taken 18 days to be completed.

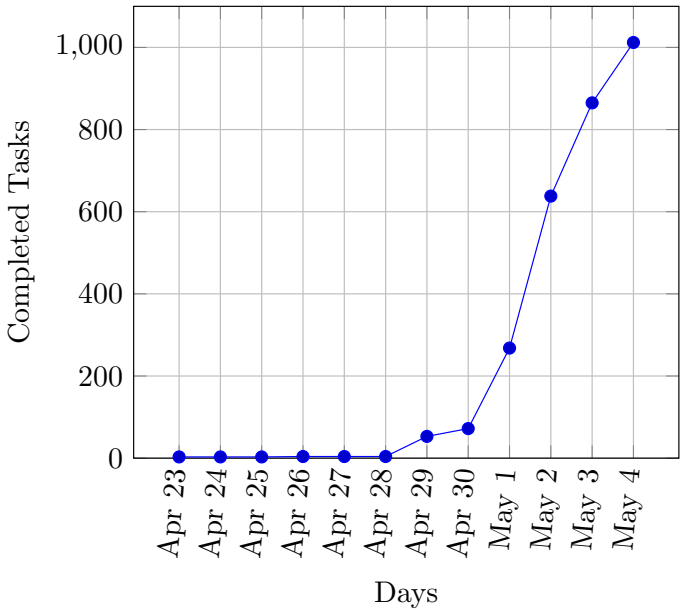

Figure 3.6: The Total Completed tasks per Day.

To get a comparison with paid crowdsourcing framework, we have compared our results with those of Wray and Ali [28]. They performed a dialect annotation using a paid crowdsourcing platform (CF) the well know platform. Figure 3.4 represents comparison between Wray and Ali [28] and KALAM'DZ annotation.

	Wray and Ali [28]	Our annotation
# of segments	47696	5187
# of contributors	2053	218
Corpus time	404 hours	10h 24 mn
Platform	CF	Crowdcrafting
Project duration	21 days	18 days
Answers per day	2271	288

Table 3.4: Comparison between Wray and Ali [28] and KALAM'DZ Annotation.

These results can be considered as comparable results due to the fact that our advising method is modest, our mailing list are also limited to students and academic communities.

3.2 Annotation Quality

We have selected two strategies to ensure the quality control. First, we have excluded all answers that have listened to the speech less than one second to eliminate malicious works. Second, In order to measure the quality of the annotation, we have selected 10% of the task, which is about 106 tasks from total to be validate it manually by expert.

Table 3.5 reports confusion matrix of expert and crowds annotation. We have observed that the accuracy of crowd annotation is about 81.1 %, where expert agree crowds in 86 tasks from total (106 tasks).

	Expert								
	Sulaymite			Ma'qilian			Hilali-Tellian		
	Yes	No	Unkn.	Yes	No	Unkn.	Yes	No	Unkn.
Yes	18	1	-	13	-	-	6	-	-
No	2	-	-	1	-	-	3	2	-
Unkn.	-	1	-	-	-	-	-	-	-
	Sahel-Tell			Algiers-Blanks			Hilālī-Saharan		
	Yes	No	Unkn.	Yes	No	Unkn.	Yes	No	Unkn.
Yes	7	-	-	11	-	-	11	-	-
No	1	-	-	1	1	-	2	1	-
Unkn.	2	-	-	-	-	-	-	-	-
	High plains			Pre-Hilālī					
	Yes	No	Unkn.	Yes	No	Unkn.			
Yes	10	-	-	5	-	-			
No	2	-	-	3	-	-			
Unkn.	1	-	-	-	-	1			

Table 3.5: Confusion Matrix of Expert and Crowds Annotation.

3.3 Toward Good Practice of Crowdsourcing

Figure 3.7 reports the distribution of contributors participation according to the connexion profile (authenticated, anonymous).

We observe that in the beginning of the execution period (from 17 April to 21 April), so for the five first days a highest participation rate is reported. It is due to it closeness from the massive open call. When the participation has decreased, we have relaunching more calls both by emails and on Social Media posting.

As a deduced rule, we advise that periodically, until all tasks will be completed, the requester must recall newest volunteers

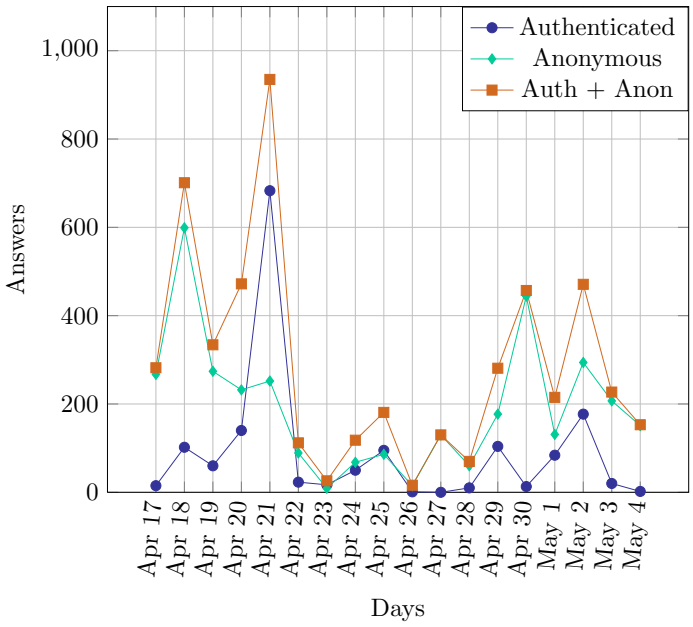

Figure 3.7: The Distribution of Answers per Day.

contributors and remind them to participate.

In addition, the participation calls must be examined. Figure 3.8 illustrates the distribution of answers per hour. We note that, at least in Algerian community, between 16 and 20 hours, the participation are higher. It leads us to advice that the participation call are efficient to be launched at this period.

Figure 3.9 shows the distribution of answers according to connexion user profile (authenticated or anonymous). Even Crowdcrafting allows authenticated connexion, this result show that crowd prefer to work in anonymous way.

Table 3.6 reports the distribution of annotatos per dialect and global, after the voting judgement of crowd. We observe that:

- Sulaymite, Ma'qilian, and Hilālī-Saharan dialects are the

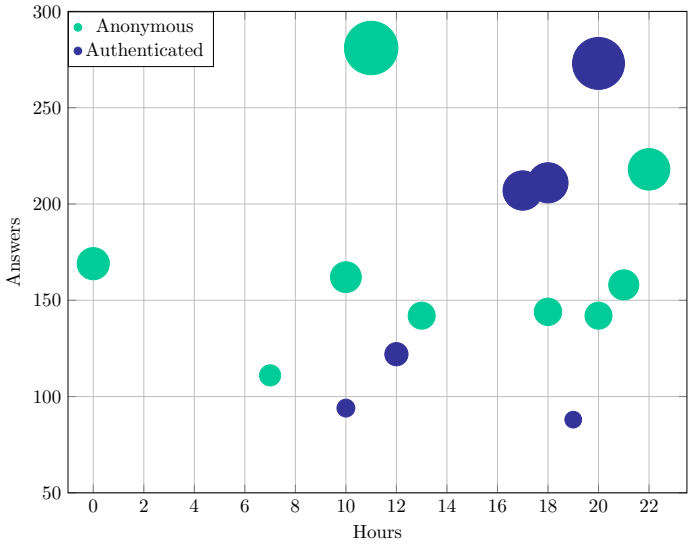

Figure 3.8: The Distribution of Answers per Hour.

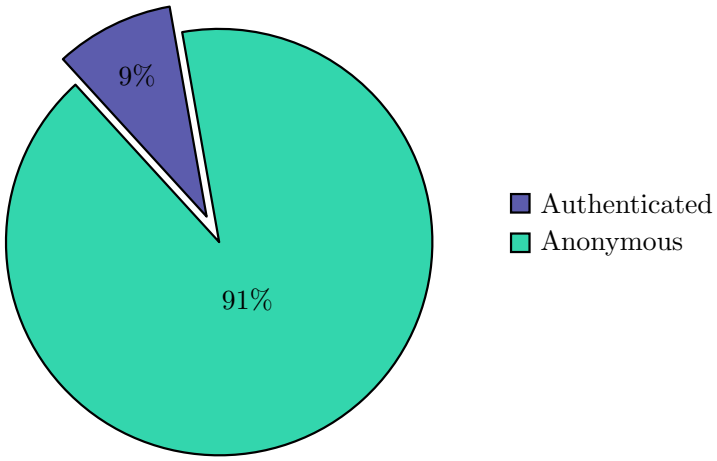

Figure 3.9: The Distribution of Answers According to User Profile (Authenticated or Anonymous).

most to detect, the contributors validate it with more than 90%.

- High plains of Constantine, Sahel-Tell, and Algiers-Blanks

are bit less known.

- In third place, Pre-Hilālī and Hilālī-Tellian dialects present as considered as known between 60-65%.
- The crowd agree with semi-automatic annotation with more than 82% percent. After these results, we can take into account the semi-automatic dialect annotation.

Dialect	Yes	No	NotKnown
Ma'qilian (الغرب الجزائري)	94.44 %	5.56 %	0 %
Hilālī-Saharan (الصحراء)	93.6 %	6.4 %	0 %
Sulaymite (الشرق الجزائري)	90.99 %	8.56 %	0.45 %
Algiers-Blanks (العاصمة و ضواحيها)	80.95 %	17.46 %	1.59 %
High plains of Constantine (الهضاب العليا)	73.81 %	10.31 %	15.87 %
Hilālī-Tellian (التل الهلالي)	65.18 %	18.75 %	16.07 %
Pre-Hilālī (ما قبل الهلالين)	62.34 %	27.27 %	10.38 %
Sahel-Tell (التل الساحلية)	79.38 %	13.40 %	7.22 %
Global	82.19 %	12.27%	5.53%

Table 3.6: Responses Statistics for Each Dialect and Global.

4 Summary

Based on the results presented, we have proposed a working list of best practices for using free crowdsourcing platform such as Crowdcrafting for validation of dialect annotation:

- Restrict tasks to users in specific area to match the required language skills needed for dialectal annotation.
- As unpaid platform, daily observation must be done to check the progress of completed tasks to recall new volunteers when it is needed.
- The crowd prefer to work as anonymous contributor. Then, do not oblige them to be authenticated.
- The time of launched calls must be considered to get a large participation.
- Quality control is not offered by unpaid platform, for that it must implemented on the project to avoid malicious work and get best results.

Conclusion and Future Work

For many researchers and institutions, crowdsourcing has become a popular method in Natural Language Processing (NLP) for lowering time and cost comparing to expert requirements. We have reviewed some related work about Arabic NLP and crowdcourcing usage. Most reviewed crowdsourcing applications are done with fees by using the paid platform.

In this work, we have experimented the crowdsourcing technique to validate the dialect annotation of Kalam'DZ corpus. Our main goal is deploying free crowdsourcing as an alternative of paid one.

Kalam'DZ corpus is collected from the Web sources namely YouTube, Online Radio and TVs. The size of corpus is about 104 hours with 4881 speakers. The most of dialect annotations are provided from the related metadata of the Web source which are namely semi-automatic. Our goal is to validate these semi-automatic annotation, we have taken a part of Kalam'DZ corpus which represent 10% of total duration with 1011 tasks.

To complete the tasks, we have used unpaid platform which is Crowdcrafting. We have called voluntary Algerian crowd for participation via email and social networks posts. The application was hosted for a period of 18 days during which, more than 218 users registered for participation and volunteered for the tasks assigned.

We have selected two strategy to ensure the quality control of responses.

First, we have call only Algerian crowd to participate to match the required dialect skills needed for Algerian dialectal annotation. Second, we have excluded all answers that have listened to the speech less than one second to eliminate malicious works.

After using some quality control mechanism, we find that the opinions of crowd are similar to semi-automatic annotation with 82% of accuracy. In order to measure the quality of the crowd annotation, we have selected 10% of the tasks, which is about 106 tasks from total, to be validated by expert. We find that the accuracy is about 81% between the expert and crowd annotation. In addition, we have proposed some guideline which can be serve as good practices when using free crowdsourcing for speech corpora.

As future work, we plan to extend the usage of crowdsourcing to transcript Arabic Algerian dialects.

References

[1] Ela Kumar. *Natural Language Processing*. IK International Pvt Ltd, 2011. 2

[2] Matt Kiser. Introduction to Natural Language Processing (NLP) 2016. http://blog.algorithmia.com/introduction-natural-language-processing-nlp/. Accessed: 25-04-2017. 2, 31

[3] Robin Sandhu. Applications of Natural Language Processing Technology. https://www.lifewire.com/applications-of-natural-language-processing-technology-2495544. Accessed: 25-04-2017. 2

[4] Peter Behnstedt and Manfred Woidich. Dialectology, 2013. 2

[5] Alec Lynch. Crowdsourcing Is Not New - The History of Crowdsourcing (1714 to 2010). http://blog.designcrowd.com/article/202/crowdsourcing-is-not-new--the-history-of-crowdsourcing-1714-to-2010. Accessed: 10-02-2017. 2, 8, 9

[6] Mohamed Abdelmageed Mansour. The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus. *International Journal of Humanities and Social Science*, 3(12):81–90, 2013. 3

[7] Soumia Bougrine, Aicha Chorana, Abdallah Lakhdari, and Hadda Cherroun. Toward a Web-based Speech Corpus for Algerian Arabic Dialectal Varieties. *WANLP 2017 (co-located with EACL 2017)*, page 138, 2017. 3, 40, 41

[8] Omar F Zaidan and Chris Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1220–1229. Association for Computational Linguistics, 2011. 3

[9] Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. Using Mechanical Turk to create a corpus of Arabic summaries. 2010. 4

[10] Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann. *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. Wiley Publishing, 1st edition, 2013. 6, 7, 12, 17

[11] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *LREC*, pages 859–866, 2014. 6, 26, 44

[12] Ria Mae Borromeo and Motomichi Toyama. An investigation of unpaid crowdsourcing. *Human-centric Computing and Information Sciences*, 6 (1):11, 2016. 6, 21

[13] James Surowiecki. The wisdom of crowds. 2004. *New York: Anchor, cop*, 2004. 7

[14] Peter Gasca. 6 Reasons to Use Crowdsourcing. `http://www.inc.com/peter-gasca/6-reasons-to-use-crowdsourcing.html`. Accessed: 20-02-2017. 10

[15] Phaniraj Kandakatla. Crowd sourcing Social Media. https://www.slideshare.net/phanirajkandakatla/crowd-sourcing-social-media. Accessed: 02-03-2017. 16

[16] Harri Oinas-Kukkonen. Network analysis and crowds of people as sources of new organisational knowledge. *Knowledge Management: Theoretical Foundation*, pages 173–189, 2008. 17

[17] Ian Lane, Alex Waibel, Matthias Eck, and Kay Rottmann. Tools for Collecting Speech Corpora via Mechanical-Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 184–187, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 19

[18] Stephen A Kunath and Steven H Weinberger. The wisdom of the crowd's ear: speech accent rating and annotation with Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 168–171. Association for Computational Linguistics, 2010. 19

[19] Crowdsourcing and Social Media: the Future of Engagement. https://fr.slideshare.net/Cat_L/crowdsourcing-and-social-media-the-future-of-engagement, . Accessed: 02-03-2017. 25

[20] Michael Marchionda. Crowdsourcing Spreading like Wildfire with Social Media. http://www.prescientdigital.com/articles/web-2.0/crowdsourcing-spreading-like-wildfire-with-social-media. Accessed: 02-03-2017. 26

[21] Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. Perspectives on

crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31, 2013. 27

[22] Robin. What is Corpus? http://language.worldofcomputing.net/linguistics/introduction/what-is-corpus.html. Accessed: 17-03-2017. 31

[23] Colleen Richey. Speech Corpora. http://web.stanford.edu/dept/linguistics/corpora/material/X_Speech_Corpora.pdf. Accessed: 17-03-2017. 31

[24] Geoffrey Leech. Developing Linguistic Corpora: a Guide to Good Practice. https://ota.ox.ac.uk/documents/creating/dlc/chapter2.htm. Accessed: 17-03-2017. 32

[25] The Stanford NLP Group. Arabic Natural Language Processing. https://nlp.stanford.edu/projects/arabic.shtml. Accessed: 10-03-2017. 32

[26] Nizar Y Habash. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187, 2010. 33

[27] Wajdi Zaghouani and Kais Dukes. Can Crowdsourcing be used for Effective Annotation of Arabic? In *LREC*, pages 224–228, 2014. 33, 37

[28] Samantha Wray and Ahmed Ali. Crowdsource a little to label a lot: Labeling a speech corpus of dialectal arabic. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. xii, 33, 34, 36, 37, 54, 55

[29] Samantha Wray, Hamdy Mubarak, and Ahmed Ali. Best Practices for Crowdsourcing Dialectal Arabic Speech Transcription. In *ANLP Workshop 2015*, page 99, 2015. 35, 37

[30] Gaurav Kumar, Yuan Cao, Ryan Cotterell, Chris Callison-Burch, Daniel Povey, and Sanjeev Khudanpur. Translations of the CALLHOME Egyptian Arabic corpus for conversational speech translation. Citeseer, 2014. 35, 36, 37

[31] Waed Hakouz, Abdurrahman Ghanem, Samantha Wray, and Ahmed Ali. LAHAJET: a Game for Classifiying Dialectal Arabic Speech. 36, 37

[32] Open Source Crowd Sourcing Platforms. http://www.it.iitb.ac.in/frg/wiki/index.php/Open_Source_Crowd_Sourcing_Platforms, . Accessed: 27-04-2017. 46