A Paradigm for Democratizing Artificial Intelligence Research

Erwan Moreau, Carl Vogel, Marguerite Barry

A Paradigm for Democratizing Artificial Intelligence Research

Erwan Moreau, Carl Vogel and Marguerite Barry

Erwan Moreau, *Computational Linguistics Group, Trinity College Dublin & the SFI ADAPT Centre*, moreaue@tcd.ie

Carl Vogel, *Computational Linguistics Group, Trinity College Dublin*, vogel@tcd.ie

Marguerite Barry, *School of Information and Communication Studies, University College Dublin*, marguerite.barry@ucd.ie

Abstract

This proposal outlines a plan for bridging the gap between technology experts and society in the domain of Artificial Intelligence (AI). The proposal focuses primarily on Natural Language Processing (NLP) technology, which is a major part of AI and offers the advantage of addressing problems that non-experts can understand. More precisely, the goal is to advance knowledge at the same time as opening new communication channels between experts and society, in a way which promotes non-expert participation in the conception of NLP technology. Such interactions can happen in the context of open-source development of languages resources, i.e. software tools and datasets; existing usages in various communities show how projects which are open to everyone can greatly benefit from the free participation of enthusiastic contributors (participation is not at all limited to software development). Because NLP research is mostly experimental and relies heavily on software tools and language datasets, this project proposes to interconnect the societal issues related to AI with the NLP research resources issue.

1 Introduction

We propose a means of bridging the gap between technology experts and society in the domain of Artificial Intelligence (AI). Here we focus primarily on Natural Language Processing (NLP) science and technology, which is a major part of AI, for practical reasons explained in §2.1.2. However, the proposal we articulate takes the form of a paradigm[1] which may be adapted to other areas of artificial intelligence, retaining the key organizational principles. "Big data" creates the promise of gaining information not available through more sampling, but at the expense of extreme processing challenges. Under this perspective, our proposal focuses on extending the possibilities for scientific exploration and technology development outside the confines of academic and commercial labs to include the general public in a kind of principled "crowd-sourcing".

One goal that we take to be shared generally with the research community, concomitant with advancing knowledge, is to have a dialogue with the society about new findings and the direction of scientific progress. We propose to open new communication channels between experts and society, in a way which promotes non-expert participation in the conception of NLP technology. Such interactions can happen in the context of open-source development of languages resources, i.e. software tools and datasets; existing uses in various communities show how projects which are open to everyone can greatly benefit from the free participation of enthuastic contributors (participation is not at all limited to software development). Because NLP research is mostly experimental and relies heavily on software tools and language datasets, this project proposes to interconnect the societal issues related to AI with the NLP research resources issue. Thus the key objectives of the proposal are:

1. Improve the reproducibility of NLP research, especially via more systematic publication of software and data resources;
2. Improve the reusability and dissemination of language resources coming from NLP research, in a way which improves openness and transparency in NLP research;
3. Reach out to a larger public, and encourage interactions between experts and the general public using open-source research resources as a support for discussion;

[1] This is one of a number of senses of "paradigm" as used by Kuhn [23, p.23]; see also, [29].

4. Integrate contributions from the general public into the research life cycle, i.e. allow the general public to become an actor of innovation[2] rather than a passive user only;
5. Credit contributions from individuals from all participating sectors in a manner that respects each individual's right to privacy as well as fair recompense.

In §2 we go over the reasons why there is a large consensus about the interest, if not the need, to include the general public in the "AI revolution" as much as possible; we also look more closely at some specific challenges related to resources in the field of Natural Language Processing (NLP). Then in §3 we explain how these two challenges can be addressed together; we propose an original and pragmatic approach based on involving the general public very early in the making of the resources used in NLP research. Finally in §4 we examine the potentially far-reaching benefits that can arise from the proposal on many levels: ethics, society, research and industry.

2 Motivations

2.1 Familiarizing the General Public with AI

2.1.1 Society and AI

AI technology is ubiquitous. The development of AI carries a number of societal and ethical questions (e.g. [30], [53]). AI is already introducing major disruptions in many areas of life (e.g. [6]), and is likely to cause more in the future. There are legitimate concerns about how the general public might react to the potential problems caused by AI [44], as well as about potential negative biases or flaws in AI systems (e.g. [41], [1]) which might prove hard to detect.

We think that lack of information about AI technology is a major source of confusion, misunderstanding and even potentially fear among the general public. As a matter of fact, lack of transparency itself in a legitimate cause for concern from a democratic perspective. Currently, the main providers of AI that the public is exposed to are large private companies, typically the "GAFA";[3] these companies lead the field in terms of breaking the barriers of AI (e.g. [10]), but the AI technology used in their everyday products is not open to scrutiny; this can lead to various kinds of suspicions, based on reality or not [50]. Most of the information that mainstream media give to the general public is understandably short and focuses on major achievements or societal questions. Overall, there are very few ways for the general public to familiarize themselves with AI technology, unless they are willing to enroll for a university degree on the topic.

Clearly, the question of *how* AI works requires a level of expertise which is worth acquiring almost only for professionals. Nevertheless, the *what* question can address many concerns: by "familiarizing" the general public with AI, we mean making it possible and accessible for non-experts to get a general sense of *what* AI can do (and what it cannot), using *what* kind of information.[4] The accessibility of knowledge about AI is currently limited mostly to experts. We advocate for opening access to non-experts, through realistic feasible means.

2.1.2 Why Natural Language Processing (NLP) Technology is a Good Starting Point

The challenge of popularizing AI is obvious: it is a very advanced technology which requires some scientific background. Even in the perspective that we propose of "familiarization" to AI (defined in §2.1.1), some applications include domain-specific complexities which require specialized expert background (e.g. in physics or biology). However some applicative domains are more intuitive; compared to other fields, NLP has several advantages with respect to popularization of scientific knowledge:

- NLP focuses on automatizing language-related tasks, and nearly every adult human knows how to speak, and possibly read and write at least one language. In other words, the object of study is something that everyone is familiar with. Thus, understanding the main NLP problems and applications does not require any advanced knowledge. For instance, everyone does not know how machine translation works, but they understand what machine translation is supposed to do. This is a great advantage compared to domains in which understanding the problems themselves requires advanced knowledge.

[2] The word *innovation* is used here in a general sense, that is, including but not restricted to commercial applications.

[3] Acronym for the four major technology companies: Google, Apple, Facebook and Amazon.

[4] As a simple analogy, how many people know how modern car engines work? Certainly very few, but the vast majority of people are comfortable with using cars nonetheless.

- NLP contributes to applied science – its outcomes have concrete applications in the form of software libraries or products. Nowadays a large proportion of the population is computer-savvy (at least in developed countries). Additionally the immaterial nature of software technology makes it an ideal product in terms of logistical costs, since it can be reproduced and distributed cost-free (as opposed to any physical-device dependent technology).
- NLP is largely an experimental science. The concept of experiment is intuitive and can be appealing to many people. Thanks to the previously mentioned advantages (little domain-specific knowledge required and software tools), there is no major obstacle to non-experts being able to understand, reproduce and do their own NLP experiments (see also §2.2.1 below).
- NLP is large. Developing and testing theories of the world's languages – past, present and future – involves an expanse of data that no one individual can ever hope to comprehend in any literal sense. This invites the benefit of contributions from many contributors across many categories of expertise, experts and citizen scientists alike.
- NLP is AI-complete, in the sense that it involves problems (such as enumeration of ambiguities inherent in natural language expressions) that are in the worst-case intractable, and which therefore require approximations in order to achieve human-level performance, in equivalence with other domains of AI research. Therefore, gaining advantage from a "big team" approach in NLP may be expected to yield progress in other domains of AI as well.

The scope of NLP is quite flexible. It intersects with various other domains or sub-domains, e.g. information retrieval, knowledge representation, data mining, multimodality and machine learning (ML). We propose to focus on NLP as a starting point for the reasons explained above, and then to progressively broaden the range of domains/applications, for instance to other kinds of concrete documents (images, audio, video). In the remainder of this document we will use NLP as the target field of interest.

2.2 Boosting NLP Research by Enhancing NLP Tools

2.2.1 NLP: Experimental Research

Much of the research carried out in NLP involves a corpus-driven view of language (i.e. in which language is defined empirically) conveniently coupled with Machine Learning (ML) methods. Naturally, this empirical approach relies heavily on experimental results: by definition, corpus-driven NLP aims at automating the extraction of some form of "language knowledge" from some "language data"; an experimental process is also used to validate, compare and sometimes combine ML methods. Experiments usually involve some kind of software prototype, or at least some form of algorithm describing the steps of the experiment (i.e. potentially convertible to a software prototype), as well as some data resources used to train and/or validate the system.

As a means of verifying or extending the findings, research papers are supposed to make it possible to replicate and/or reproduce the experiments. In particular, they usually give some details about the data and software used, sometimes by referencing previously published material for some part. Recently, major NLP conferences (e.g. ACL[5], COLING[6], EMNLP[7] or LREC[8]) have increasingly encouraged authors to favor reproducibility in different ways, especially by making their software and their data resources publicly available. Additionally, in many research projects, resources are part of the outcomes, typically meant to demonstrate the validity and feasibility of a method. Thus, there seems to be a positive trend towards publishing the corresponding software and data resources together with experimental research papers.

[5] *"Papers that are submitted with accompanying software/data may receive additional credit toward the overall evaluation score, and the potential impact of the software and data will be taken into account when making the acceptance/rejection decisions."* ACL 2018 Call For Papers, `http://acl2018.org/call-for-papers/` – last verified: February 2018.

[6] COLING invites papers in 6 categories including *"Reproduction papers"* and *"Resource papers"*. COLING Call For Papers, `http://coling2018.org/final-call-for-papers/` – last verified: February 2018.

[7] *"Each EMNLP 2018 submission can be accompanied by a single PDF appendix, one .tgz or .zip archive containing software, and one .tgz or .zip archive containing data. EMNLP 2018 encourages the submission of these supplementary materials to improve the reproducibility of results, and to enable authors to provide additional information that does not fit in the paper."* EMNLP 2018 Call For Papers, `http://emnlp2018.org/calls/papers/` – last verified: February 2018.

[8] *"Describing your LRs [Language Resources] in the LRE Map is now a normal practice in the submission procedure of LREC (introduced in 2010 and adopted by other conferences). To continue the efforts initiated at LREC 2014 about "Sharing LRs" (data, tools, web-services, etc.), authors will have the possibility, when submitting a paper, to upload LRs in a special LREC repository. This effort of sharing LRs, linked to the LRE Map for their description, may become a new "regular" feature for conferences in our field, thus contributing to creating a common repository where everyone can deposit and share data."* LREC 2018 web page, `http://lrec2018.lrec-conf.org/en/` – last verified: February 2018.

Data resources are often shared because they are costly to produce. However, probably most software prototypes are still either not published or not publicized, and consequently never re-used; once they have served their purpose as support for an experiment and the results have been published, unless their authors plan to re-use them in subsequent work, they are simply abandoned.[9] Of course, not every software prototype is worth being re-used, but out of the thousands of experimental papers published every year, very few prototypes get enough scientific and/or technological interest to be deemed worthy of devoting resources to their maintenance, development and documentation.

Hence the vast majority of software prototypes are either lost or re-used only by a small group of people, typically the team in which it originated. At the other end of the spectrum, only a handful of tools are widely used by the NLP community. We explain below in §2.2.2 and §2.2.3 why this is problematic from a scientific perspective.

2.2.2 Reproducibility

The scientific value of NLP research relies on a rigorous evaluation of experimental results. The experimental setup in which the results are obtained is a major part of this evaluation, as well as the choices related to the data used in the experiment (origin, type, size, annotations, etc.). Results can vary widely depending on the data used as input, since textual data is extremely diverse. In the typical context of supervised ML this has to be expected; but NLP methods also aim at a certain level of generality: a method which does not scale to various kinds of text data is of little use. This is a distinctive feature of NLP which often surprises ML experts who are more familiar with different types of data: the richness and diversity of language makes the task of finding generalizable patterns much harder.[10]

With ML applied to text data, it is common to reach a satisfying level of performance with a specific dataset by "cherry-picking" the features and meta-parameters of the model, intentionally or by mistake.[11] The reviewing process of experimental papers includes checking for this kind of scientific error; this is why authors have to describe the experimental setup, including the steps they take to avoid overfitting for instance (if such details are missing in a paper, experienced reviewers take it into account in their final evaluation). However, because the space is limited in most papers, and because the paper is supposed to focus on the main contribution rather than give every possible technical detail, the reviewing process cannot guarantee that all the conclusions made by the authors are correct, let alone generalizable. In other words, the paper itself is not always a sufficient proof of the validity of a method.

Replicability and/or reproducibility[12] are major evaluation tools in the methodology of experimental sciences such as NLP. A researcher or a reviewer, even doing their job diligently and honestly, is subject to many potential biases [39]. Ultimately, the scientific value of a finding lies in the extent to which it is reproducible.[13] The field of NLP has seen significant progress in this areas, for example with the well established organization of competitions[14] as a means to compare different approaches on the same grounds; participants in these competitions are encouraged to publish their software tools, so that everyone can reproduce their results and build upon them. In general, an increasing amount of research software and data is published together with traditional papers in many conferences. Despite the existence of many such initiatives, a majority of the papers published in NLP describe research without providing the means to reproduce it; it might be reproducible, but the cost involved in developing a system to actually reproduce it is prohibitive. Besides, since NLP does not deal directly with critical issues such as health or aircraft safety, there is no major need for a strict and thorough validation of its findings. As a consequence reproducibility studies are very rare in NLP, and the state of the art can be fuzzy in certain areas about which method works best for a particular application.

[9] For example, it is very common that PhD students or postdocs implement some software tool as part of their PhD project, but it is rarely picked up when the author of the software leaves at the end of their contract.

[10] One might reflect on the observations of those who try to identify linguistic universals. One might imagine it possible to generalize that nouns and verbs are categories of words attested in every human language. However, even this candidate universal is controversial [13].

[11] This is indeed a common beginner's mistake: it is easy to overlook the risk of overfitting, and to mistakenly interpret good performance results on a test set drawn from the same dataset as the training set as a sign that the approach is valid.

[12] The exact definition of replicability and reproducibility can differ by scientific field. Nevertheless the two always differ in a way which makes replicability a more general condition than reproducibility: according to Wikipedia, reproducibility is the ability to get the same research results using the exact same raw data and computer programs as the original researchers, whereas replicability is the ability to independently achieve similar conclusions when differences in sampling, research procedures and data analysis methods may exist (`https://en.wikipedia.org/wiki/Reproducibility` – last verified: November 2017).

[13] That is, it is easy to reproduce a demonstration that water flows downhill, but this is not a "finding". "Cold fusion" is a finding, but it is difficult to reproduce.

[14] Also called "challenges" or "shared tasks".

2.2.3 Today's Software Prototypes are Tomorrow's Research Instruments

The relatively low level of software outcomes from NLP research weakens reproducibility, but also impacts the pace of research itself, especially academic research. Near-future challenges in the domain will precisely rely more and more on robust, scalable and complex software components, at least for two reasons: the amount of data to process and the growing complexity of the tasks that we try to achieve (hence a growing need for a variety of modular components).

"Normal" scientific research[15] progresses iteratively by incremental improvements, a fact commonly summarized by the phrase *"standing on the shoulders of giants"*.[16] The re-use of previous findings is the bread and butter of NLP research, either to improve over existing research or to combine them in order to solve new problems. This is true about theoretical findings as well as empirical results, the latter being ubiquitous in NLP research (as explained in §2.2.1). As a consequence, research prototypes originally developed to demonstrate the validity of a method (or clones based on them) are often re-used to achieve new goals. Moreover, text processing traditionally involves a processing chain: different levels of linguistic information are extracted incrementally in order to obtain a rich linguistic representation of the structure of a sentence or document. Every level is analyzed in turn using a tool specialized for this task. The processing chain typically starts with word and sentence segmentation, followed by Part-Of-Speech (POS) tagging and lemmatization; depending on the task, other steps might consist in named entity recognition, dependency parsing or other more advanced tasks. This hierarchical process implies that the accuracy at a given stage depends on the quality of the results obtained at earlier stages. Thus, error propagation, i.e. the fact that a minor error at an early stage might cause more serious errors which accumulate down the stream, is an important issue: the more advanced the task is, the more the latest levels of the chain rely on the quality of the earlier levels. Arguably, the state of the art in NLP keeps progressing towards more and more complex tasks. As a consequence, not only this issue is becoming more and more crucial, but also the research tends to be more and more specialized: for instance, experts working at the semantic level cannot afford to spend time on the morphosyntactic analysis part, which is not their focus. Hence they rely on software tools made by others for the first components of their processing chain. The impact of their research depends on the quality of these tools and also their diversity, because a certain implementation might be more appropriate for certain usages than others.

In other words, the software prototypes become the instruments used for further research. In many experimental sciences, the importance of scientific instruments is acknowledged. Research resources are devoted to improve them, and making progress in the quality of the instruments is rewarded as a real contribution to the research: the Nobel prize awarded to Georges Charpak in 1992 for his invention and development of particle detectors is a striking example of such recognition. There is arguably not enough effort in NLP towards developing and using quality tools, despite their importance in the research process. In the next section we illustrate this observation with an example.

2.2.4 An Example: TreeTagger

Part of speech (POS) tagging is a task which is included in the standard text processing chain, and it is an essential step in many tasks in NLP. It consists in labeling every word in a sentence with its morpho-syntactic category, called the POS tag. TreeTagger is the name of one of the first supervised Part-Of-Speech taggers: it was introduced in 1994 by Helmut Schmid [45]. The software tool is very popular among the NLP community: Google Semantic Scholar estimates that the paper was cited 2,735 times[17] (1,008 citations according to CiteSeer's database[18]); the original paper reached its peak of citations in 2014, i.e. 20 years after publication,[19] and it still gathers a healthy hundred of citations every year. While some of these citations might not correspond to experiments which actually use the software (e.g. simple mentions in literature reviews, etc.), most of them probably do.

TreeTagger was an excellent tool for POS tagging, but after more than 20 years of success the tech-

[15] According to [23], periods of "normal scientific progress" are interrupted by periods of "revolutionary science".

[16] According to Wikipedia, *the metaphor of dwarfs standing on the shoulders of giants expresses the meaning of "discovering truth by building on previous discoveries". This concept has been traced to the 12th century, attributed to Bernard of Chartres. Its most familiar expression in English is by Isaac Newton in 1675: "If I have seen further it is by standing on the shoulders of Giants."* `https://en.wikipedia.org/wiki/Standing_on_the_shoulders_of_giants` – last verified: February 2018.

[17] `https://www.semanticscholar.org/paper/Probabilistic-Part-of-Speech-Tagging-Using-Decisio-Schmid/bd0bab6fc8cd43c0ce170ad2f4cb34181b31277d` – last verified: February 2018.

[18] `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1139` – last verified: February 2018.

[19] `https://www.semanticscholar.org/paper/Probabilistic-Part-of-Speech-Tagging-Using-Decisio-Schmid/bd0bab6fc8cd43c0ce170ad2f4cb34181b31277d` – last verified: February 2018.

nology on which this venerable tagger is based is outdated:[20] there are more recent approaches in ML and especially in sequence labeling which perform better than the probabilistic decision trees method used by TreeTagger.[21] Thus it is likely that the popularity of TreeTagger nowadays has more to do with: on the one hand, its availability and the awareness thereof in the NLP community; on the other hand, its simplicity and usability as a software tool, including the fact that it is available for a fairly large set of languages (see also §4.1.3). In other words, a lot of NLP research is still based on a suboptimal POS tagger because it is popular and easy to use, potentially causing suboptimal results in the applications in which it is used. Incidentally, this implies that more recent POS taggers lack popularity (awareness) and/or usability.

This example demonstrates that there are gaps in the NLP research life cycle: the most adequate tool is not always immediately picked up by the community. Of course, such gaps would be filled eventually: new tools gain popularity progressively, and eventually the best tools are adopted, even if the qualities that determine what is "best" evolve. However, the process of adoption of the best tools can be accelerated, with potentially important scientific benefits (see 4.2).

3 Approach

3.1 Key Ideas

3.1.1 Key Objectives

The problems presented in §2 cover a broad spectrum ranging from ethical and societal questions (§2.1.1) to epistemological questions in experimental NLP like reproducibility (§2.2.2) and research instruments (§2.2.3). These issues might look like abstract problems barely related to each other, and each of them hard to address in a methodical way. This section will hopefully demonstrate the contrary: we propose a strategy intended to tackle these issues, through a framework in which they are deeply interconnected. The key objectives summarized below sketch the main ideas of this strategy:

1. Improve the reproducibility of NLP research, especially via more systematic publication of software and data resources;
2. Improve the reusability and dissemination of language resources coming from NLP research, in a way which improves openness and transparency in NLP research;
3. Reach out to a larger public, and encourage interactions between experts and the general public using open-source research resources as a support for discussion;
4. Integrate contributions from the general public into the research life cycle, i.e. allow the general public to become an actor of innovation rather than a passive user only;
5. Credit contributions from individuals from all participating sectors in a manner that respects each individuals' right to privacy as well as fair recompense.

The fact that these issues involve heterogeneous categories of people is the main challenge to be addressed. Naturally, the cornerstone to establishing a dialogue between experts and non-experts lies in making every stakeholder see the benefit they can gain from contributing. This creates an interest in identifying a shared language through which non-trivial dialogue may emerge.

3.1.2 Philosophy and Scope

Before going into further details, in such a wide-ranging proposal two main pitfalls must be avoided: on the one hand, an "idealistic" approach in which success depends solely on the individual willingness of the actors to contribute; and, on the other hand, a "bureaucratic" approach which discourages individual initiatives and small contributions. This is why this plan focuses on implementing effective structures intended to make every actor benefit from their participation, by leveraging these contributions into a global cooperative ecosystem. Such structures must also integrate seamlessly in their respective environments, in order to allow a progressive cultural shift towards the goal. Thus, the core of the proposal consists in progressively building a backbone structure which provides support and organizes the contributions of different actors, but is also flexible enough to evolve and scale up. While this plan requires resources for its deployment and sustainability, ultimately its aim is to help growing a broad self-reliant ecosystem which

[20] Additionally the datasets on which it was trained are quite old.

[21] `https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)` – last verified: February 2018.

evolves with its own dynamics.[22] This is why the proposal focuses on building from existing resources and structures and encouraging and coordinating third-party efforts which are compatible with its goals.

We propose to follow the philosophy of agile development,[23] in this case applied not only to software development but to the whole proposal, in particular in its social aspects. The main implication is that the proposal is meant to be developed iteratively over life cycles, with continuous re-evaluation of the best way to progress towards the goals taking into account any relevant circumstances: trends in research and/or society, but also available resources and partners on the proposal. Hence, the exact scope of the proposal itself across time is meant to be flexible, in line with the guiding principles explained above.

3.1.3 Trigger A "Contagion" Effect

The philosophy of the proposal entails the involvement of heterogeneous categories of people in different sectors. This relies on the assumption that the proposal can trigger a "contagion" effect[24] across these population sectors, which we explain in this section. First, a brief definition of what "these different levels of population" represent is in order. Below we present a schematic view of what is meant by that in two perspectives. In one view, from the perspective of depth of awareness and expertise with AI/NLP technology, one may see a a pyramid-shaped hierarchy (in terms of the number of individuals within a sector, relative to other sectors). In the other view, one may see something like a chain-link mesh, in that with respect to NLP science and technology, one relatively small group counts as experts, but these same individuals participate among the general public for other domains of inquiry. Thus, the sectors described in some way participate in a chain, but the links are ultimately connected in more than one dimension. Understanding that the chain-link mesh view ultimately provides part of the explanatory force of our proposal, we focus initially on a single chain of linked sectors within the mesh.[25] We start at one end (in the pyramid view, this is the top with the smallest number of NLP experts (people and organizations)) proceed through different population sectors, which get larger but less aware as one proceeds, ending with the the general public, the largest population with the least depth knowledge about AI/NLP science and technology.

- **Research Experts.** Research experts create, build upon and improve the methods used in NLP technology. New methods are tested and compared in scientific experiments (this is generally when the first software prototypes are built, using new or existing data resources). These experts have the required resources (time and skills) to stay up to date with the state of the art; they are also the first users of languages resources (software and data), that they use to compare, extend or improve the state of the art. This category includes experts from organizations in academia, public institutions, large companies with research activities as well as smaller but very specialized innovation companies.
- **Professionals.** Professional experts, in particular in the industry, are familiar with the state of the art in NLP and its potential applications. They are mostly converting research methods into end-user products, this is why they need a good understanding of the underlying concepts, without usually having the resources to do the core research themselves.
- **Direct Users.** Professional users who have some connection to NLP innovation but within a specific topic or application. They are more on the user side of NLP technology, and their vision of the field and its evolution is limited. This level includes a wide range of SMEs (Small and Medium

[22]The basis for our model is based on the insight that the economic viability of local communities is linked to diversity. Communities that are tied to the fortunes of single industries or industry players face greater risks than those that are home to a diverse employment base, when industrial "disruption" takes root. In the more specific perspective of collaborative projects, this philosophy is also inspired by other successful projects, for instance Stack Exchange or Wikipedia: the infrastructure is provided, but the evolution of the content completely depends on the community.

[23]`https://en.wikipedia.org/wiki/Agile_software_development` – last verified: February 2018.

[24]Remark: this is not meant to have pejorative associations with disease, but rather the replication/repetition behaviors associated with mirror neurons and "memes".

[25]The explanatory force mentioned is this: because experts in one field are laypeople in another field it may be anticipated that at least some of the laypeople share all of (i) curiosity about domains where they lack expertise; (ii) interest in availing of the best of technologies that emerge from those domains; (iii) empathy for the desire of domain experts to achieve even greater expert knowledge in the domain. In turn, these laypeople are visible in their lay interactions with other laypeople. Laypeople who are not socially defined as experts using traditional schemes (such as university degree qualifications) may be anticipated to involve themselves even without contact with experts acting outside their domains. Nonetheless, we do not anticipate a universal take-up of a call for participation, no more than there is universal participation in recycling initiatives. However, when people perceive a duty to participate and have information that others participate, then they are more likely to than if they lack information about others' participation [8]. A question, then, is whether the general population is most appropriately approached with a lesser duty to *contemplate* participation as citizen scientists or with a non-duty to engage with citizen science for its entertainment value, alone.

Enterprises) with some connection to the field, but which do not have the need or resources to keep up to date with the state of the art.

- **Potential Users.** This level includes heterogeneous categories of people or organizations: savvy IT professionals, technology enthusiasts, students in computer science or related topics, professionals in companies which are not specialized in NLP. This layer corresponds to people and organizations who are potential NLP tech users: they have some resources but might simply not be aware of the existing technology. In particular, this category includes SMEs which might not be aware that their business could benefit from NLP technology.
- **General Public.** Finally the largest group, which forms the foundation of the pyramid consists in the general population, i.e. people who are not educated about NLP technology, although they occasionally use its applications directly or indirectly. There are societal issues regarding the lack of awareness in the population; for example, people usually ignore what can be done (or not done) with their data, thus being potentially abused or on the contrary reluctant to use some helpful technology (see §2.1.1).

The general idea is that these layers are intertwined in such a way that changes at one level potentially affect other levels. For example, the wider availability of language resources from experts is likely to impact professionals and to some extent direct users. The approach that we detail below aims to maximize the impact, so as to extend it to potential users as well as the general public (see §4). This approach relies on leveraging the diversity of the open-source communities, as a means to drive changes across various layers of society.

3.1.4 Ethical Design

Recent discussions around the ethical ramifications of AI research point to two distinct directions in research either a) educating scientists, engineers and designers in ethics or b) incorporating social scientists and ethicists in the design process. The first approach incorporates several important techniques for integrating ethical approaches to design that have been developed in Human-Computer Interaction (HCI) studies, such as reflective design [47], participatory design, value-sensitive design [16] and so on. These approaches often focus on how to identify bias or define guiding values at the outset, although bias and values can also arise as local phenomena that are "discovered" throughout design and development [25]. The second approach suggests that social scientists must be incorporated into the research process, which may produce more disagreement but arguably more discussion and deeper reflection on where ethics happens and where there are perceived to be "ethics free zones" that require attention [51]. Both approaches suit certain design settings better than others, with some studies noting "disjunctions" between AI and big data research methods and existing research ethics paradigms [31]. Others have found a tendency for such approaches to produce different levels of ethics expertise rather than establishing ethical consideration as the responsibility of all involved in a continuing process of reflection design [48]. Meanwhile user studies attempt to include user requirements and values in design, but studies suggest they are too often constructed for abstract ideal or "intensional" users [2] rather than actual users whose responses to digital technologies and interactivity are more often individualised and inherently strategic. Further, there is an increasing attention paid to the numbers of indirect users or stakeholders in the design of technologies, particularly those using AI/NLP, whose inputs are rarely sought in the design process. The challenge is how to better educate and incorporate all stakeholders in the consideration of ethics and values in design. Recent studies on AI in society recommend that policy makers and leaders in enterprise seek ways to encourage experimentation around ethics review inside and outside of university settings and recognising that the function of pausing for independent review and deliberation is indispensable [31].

This proposal follows a complementary approach to continuous ethical reflection in research, by incorporating the general public or "community" into ethics analysis during the entire research cycle from research design to data collection, analysis, implementation and reflection on impact. According to [51], ethics must be part of the whole technological design process from the very beginning. A community participation approach to ethics empowers the general public with the tools to participate in discussions and debates about AI/NLP advances and their uses throughout the research lifecycle. Integrating the public into the ethical reflection process from the outset avoids problems with unforeseen or "emerging bias" while complementing citizen science efforts in collaboration around both the process and the product of technological innovation. It also shifts ethical reflection from being a post hoc theoretical exercise to being an integrated aspect of the research process. This will produce new knowledge around ethical issues and how they arise in an NLP project, and also fresh inputs on the essential pragmatic issue of how to identify and integrate applied ethics directly into the development process.

3.1.5 Rethinking the Innovation Chain

Traditionally, the innovation chain is thought as a linear process which goes from the initial scientific concept to its applications into end-user products, progressing through various stages from scientific development to industrial technology and finally commercial products. This linear process is often represented in the form of a scale of "Technology Readiness Levels" (TRL). For instance, the EU Commission [12] defines the following TRL scale:[26]

1. Basic principles observed
2. Technology concept formulated
3. Experimental proof of concept
4. Technology validated in lab
5. Technology validated in relevant environment (industrially relevant environment in the case of key enabling technologies)
6. Technology demonstrated in relevant environment (industrially relevant environment in the case of key enabling technologies)
7. System prototype demonstration in operational environment
8. System complete and qualified
9. Actual system proven in operational environment (competitive manufacturing in the case of key enabling technologies; or in space)

According to Wikipedia,[27] the EU initially adopted the TRL levels defined by NASA for the European Space Agency. It was later extended to every kind of technology developed within the H2020 framework program. To the authors' knowledge, the EU-defined TRL levels do not include any additional information or description, but one may find an additional short description of every level in the original NASA definition.[28] Of course, these descriptions insist on testing the technology in space; this shows that the scale was designed with a very specific purpose in mind. Thus TRL levels are supposed to apply for a very broad range of technology including software systems, whereas it was initially designed for space technology only. In particular, it understandably focuses on hardware components and follows a very thorough testing process. The life cycle of a software component differs significantly from this,[29] since:

- Software is very flexible: it can be modified, reused for a different purpose, its reproduction and dissemination is practically cost-free, and bugs can be fixed even after a product has been released.
- Software components are very rarely error-free. In the case of AI software, the risk of error is inherent to the concept of AI itself. Thus, it is not expected (nor is it possible) to make AI software infallible.

These differences make the standard TRL scale poorly fit for reliably representing software development across the innovation chain: a software system is seldom fully "proven", and its life cycle does not have to be linear; in fact it rarely is. The same point can be made for data resources, which are often developed iteratively, with versioning used to identify a particular release in time.

This point matters because language resources do not require the kind of heavily controlled top-bottom development necessary for some space shuttle component. On the contrary, it allows a lot of flexibility; for example, bottom-top feedback from users or developers can be integrated into language resources quite easily. Therefore, the representation of language resources in the form of a TRL scale tends to restrict the wide range of possibilities that are available, especially in the case of new technology for which usage and applications are still to be invented. Thus strictly following the TRL perspective introduces a bias towards centralized top-bottom technology development. For instance, this model does not offer room for interactions between the researchers at the source of a new technology and its end-users. It was designed under the assumption that only professional experts participate in the development of the technology. When applied to cutting-edge AI software, this view entails that professional experts are in charge of

[26] Remark: *key enabling technologies (KET)* (mentioned in TRLs 5,6 and 9) are a group of 6 technologies that the EU wants to focus on [11]: micro and nanoelectronics, nanotechnology, industrial biotechnology, advanced materials, photonics, and advanced manufacturing technologies. These areas are not at all or only very distantly related to software technology, in particular AI or NLP.

[27] `https://en.wikipedia.org/wiki/Technology_readiness_level` – last verified: February 2018.

[28] `https://en.wikipedia.org/wiki/Technology_readiness_level\#NASA_definitions` – last verified: February 2018.

[29] `https://en.wikipedia.org/wiki/Systems_development_life_cycle`, `https://en.wikipedia.org/wiki/Software_development_process`, `https://en.wikipedia.org/wiki/Software_prototyping` – last verified: February 2018.

developing every aspect of the technology, including how to use it, and the general public is left out of the loop in a role of mere passive customers. Moreover, a large space in the AI innovation chain is occupied by big technology companies which might be commercially interested in keeping customers in such a passive role;[30] to some extent, the control these companies have on the design process might sometimes undermine the general public's interest. In other words, behind the technical aspects of something like the TRL scale model, there are democratic and ethical issues at play. As a consequence, while TRL is a useful representation of technology development, integrating the general public at various levels of the innovation chain requires a more open and more flexible model.

3.2 Strategy

3.2.1 Open-Source Software as a Gate between AI Research and the General Public

Open-source resources (software and data) play a crucial role in the plan that we propose: they can be thought of as a vessel through which "bits of AI knowledge" are conveyed. The resources must be open-source, so as to allow exploration and scrutiny by the general public. Language resources used in NLP fit in well with this plan, since the object of study is understandable without any particular background knowledge (see §2.2.1); moreover, NLP research could really benefit from improving the quality and diversity of its resources, as motivated in §2.2.2 and §2.2.3. The publication of open-source resources is already a well established practice in the NLP research community, albeit far from universal. However, its impact so far has been mostly limited to the research community circles.

The approach that we propose relies heavily on improving the dissemination and re-use of NLP resources very early in the innovation process, i.e. at the stage of research prototypes for software and at the stage of collection for datasets. This is possible only through a well-thought open-source strategy. A common misconception about open-source is that publishing the software code (or any kind of data) in some public repository suffices to ensure that the development will be picked up by "the open-source community", provided the software is worth it. This is definitely not the case, for the same reasons that in industry a good product cannot be relied upon to sell itself. There are many reasons why a project would not get traction and would simply be abandoned, no matter its quality. Among the most obvious ones one might mention:

- Clearly a project with no visibility cannot attract contributors: if not properly indexed, or hosted on a page which receives very few visits, then nobody can find it and it cannot receive any attention;
- Similarly, if it is unclear what the piece of software aims to achieve or what its main features are, then it is unlikely that anybody would bother spending time guessing what it is about; the same applies to data released with no indication of its origin, format or potential usage.
- If the original author does not answer messages or questions about their project,[31] this will discourage interested users who might have become contributors otherwise.

Unfortunately these issues are common in the case of software prototypes originating from the research community, because prototypes are often not thought as software projects but only as requirements for an experiment, and in turn for a publication (see §2.2.1). Thus the problem does not lie so much in the lack of open-source software prototypes (even if this is also a serious issue) rather than, in most cases, in the lack of effort to integrate the software in the open-source ecosystem. In the case of innovative software, providing good documentation is an even more crucial step in order to interest people in it, since the purpose of the software might be significantly more difficult to comprehend than for some regular software.

In general, the research community is positively inclined towards the principles of open-source: openness is a cornerstone of academic research, and research projects where people work together using collaborative tools are very common. In fact, in computer-related fields a number of academics are involved in various open-source projects, related to their professional work or not. However software development is rarely seen as an important goal by itself, because it does not constitute a scientific contribution to the field, i.e. a new finding worth publication. We discuss means to encourage software development in research in §3.2.2.

[30] "It's really hard to design products by focus groups. A lot of times, people don't know what they want until you show it to them." Steve Jobs [**?**].

[31] For instance because they have moved to a new job and do not use the email address anymore, which is a frequent occurrence in academia.

A closer look at the open-source ecosystem is required in order to improve the penetration of research software prototypes and data resources into it. First it is important to remind the reader that open-source does not preclude commercialization or any commercial use of the software. There are many examples of companies which participate to developing open-source code for profit, with various possible business models [3].[32] Nevertheless the open-source world spans over a wide range of cases, from casual hobby projects to ubiquitous security-sensitive software components (e.g. OpenSSL[33]). In general, there are established advantages of open-source software development: the inherent transparency improves the reliability of the code, in particular bugs are fixed faster, thanks to the community feedback and contributions. The existence of a motivated community also helps develop the product in new directions that might not have been explored otherwise, see e.g. [49]. Interface, documentation and features can be based more directly on users' experience. In open-source development, the distinction between developers and users is not as binary: users can get involved in the development in many ways, even without any knowledge of coding; contributions can be as small as fixing a typo in the documentation, reporting a bug or proposing a new feature. This clearly allows a more direct and flexible communication between users and developers. Naturally, in order to enjoy these benefits an open-source proposal must be managed in a way which encourages the integration and participation of people who are external to the project. This involves following various kinds of good practice intended to make the proposal welcoming and motivating for its participants on the one hand, and maintain a efficient organization of work on the other hand [14].

Thus the plan that we propose aims to establish bridges between heterogeneous communities (as described in 3.1.3); summarily, the two main target populations are the research community and the general public. This will be implemented through different kinds of incentives which can be roughly classified into two groups, based on their target population.

3.2.2 Involving the Research Community

The first and perhaps most important incentive for experts to engage with non-experts on a project is the satisfaction of seeing their resources used and appreciated. While this rewarding feeling is important, concrete professional advantages are likely to play a bigger role to sustain the experts motivation to commit time in outreach activities. This is why a major part of the proposal will consist in implementing incentives to this end. Since publications represent the currency of the academic world, the proposal will offer more opportunities for researchers to publish their work and value their efforts towards making software tools more reusable; this traditionally means organizing new workshops as well as journal issues devoted to topics related to the proposal goal: reproducibility, software engineering issues in NLP research, outreach contributions, etc. The organization of these scientific events must take a special care at establishing their evaluation criteria; in particular, instead of the traditional evaluation criteria focused on scientific originality, they could give more importance to software-related aspects such as documentation quality, usability, openness. This would for instance encourage the submission of work which reproduces existing methods, or provides the community with better quality software tools. In the long term, the recognition of the workshop or journal by the scientific community would make authors more willing to be published in such reputable references, hence increasing efforts of the community towards software tools and outreach.

The case of data resources is slightly different, because it is already very common to publish those in academic venues. Thus, for data resources, the focus should be put on developing methods to involve non-experts in their creation; this would directly benefit the research community because data resources are costly to produce.

These direct academic benefits are not the only advantages experts can gain from making their software tools more available and usable, in particular for non-experts. The potential popularity of their software can induce secondary benefits: other NLP researchers, students or technology enthusiasts can provide feedback and give them ideas to improve their software, which might in turn open new scientific perspectives. The visibility offered this way to valuable research software would also benefit their authors, in the form of further citations, feedback, collaborations and project funding.

3.2.3 Involving Non-Experts Communities

Clearly the "non-expert" group includes people who might be interested in AI technology for various reasons. Everyone can have their own motivations: curiosity, an appetite for technology advances, the desire to acquire knowledge or learn new skills, the intellectual challenge, etc. AI technology in particular can have a strong appeal to the general public from this perspective. Maybe one of the main reasons why

[32] `https://en.wikipedia.org/wiki/Business_models_for_open-source_software` – last verified February 2018.
[33] `https://www.openssl.org/` – last verified February 2018.

people would want to contribute is simply the satisfaction of participating to an interesting collaborative project; when properly organized, benevolent contributors can be a great help to a project, as seen in many examples of great and valuable achievements over the Internet, e.g. Wikipedia, Stack Exchange, and many open-source software projects. The conditions of success for collaborative projects can be analyzed, in order to structurally maximize the chances of success.

First, potential contributors must be aware of the existence of the project, and know that anyone can contribute. To our knowledge, there are nowadays very few research projects really open to non-expert contributions, if any. Inviting the general public to participate in a significant way to scientific projects is likely to be seen as an opportunity by many people, as opposed for example to some quite misguided attempts at involving the public in some extra-scientific part of the proposal, with the risk of people not taking it seriously.[34] Then it is important to make people feel welcome and feel that their input is taken into account. Clearly this can happen only if the experts on the proposal are ready to spend some time answering questions, improving the documentation, etc.

However, it is also important that people feel that they are not exploited. This perception is likely if others obtain wealth on the basis of contributions beyond their own while they obtain, at best, recognition. This perception was widely attested in the aftermath of the crowd-sourced localization of Facebook's interfaces into multiple languages [37]. A perception of exploitation may be an explanation for evidently malicious behaviours that have been witnessed in calls for general public participation in crowd-sourcing in the private sector [22]. Nonetheless, one may notice that members of the public continue to demonstrate willingness to volunteer for charities that they perceive in having good causes, even when the volunteers know that executives who run the charities are paid handsome salaries, and even when executives have been discovered to have been obtaining unwarranted personal gifts at the expense of the charities. Further, as suggested above (fn. 25), if there is a means for members of the community to see accurate (yet individual privacy preserving) accounts of the extent to which the rest of the community is contributing to an endeavour, as well as the recompense correspondingly accrued, then one may reasonably expect greater levels of participation than if community-wide participation rates are not visible.

Technology enthusiasts often have programming skills and are familiar with the environment and tools around software projects. Thus they are more likely to be interested and to contribute, but it is also important not to exclude people who do not have a technology background; for example, students (not necessarily following computer science studies) might explore opportunities to participate and this can be a motivation for some of them to pursue AI studies. Finally, there can also be a interest from companies which see a professional use for the technology. Importantly, small and medium-sized businesses which do not have the same access to technology as big companies might give a valuable input to direct the technology in a direction that fits their needs. Many kinds of contributions do not require any particular skills:

- Providing feedback: ask questions, fuel discussions on usage, shortcomings, etc.;
- Testing, identifying issues, propose improvements or ideas;
- Providing or annotating data;
- Identifying documentation problems, translating documentation;
- People with programming skills can contribute to the software itself, but also code extensions or interfaces, create packages for specific systems (e.g. as a plugin, add-on, app...);
- Imagine and propose ideas about potential applications, combinations with other tools, etc.;

It should be emphasized that these interactions are meant to happen at the level of a specific project, where a community can progressively gather together around a shared interest for this proposal in particular; the dynamics of such a community depend only on the people who feel that they belong to it, whether they are experts or not. Nevertheless, various steps can be taken in the perspective of encouraging such communities to grow. This would include identifying projects which have a potential for attracting contributions from non-experts and giving them some visibility; this would involve publicize them on social media, but depending on the project more specific targets can be considered, like specialized websites and forums in the open-source world, or technology media interested in recent AI trends. Contents from the latter can occasionally trigger articles in the mainstream media, thus giving potentially extensive visibility to the projects. Some projects can also be highlighted by creating demos or applications based on their software.

The appeal to the general public is to be considered as an essential goal of the proposal. This means that a significant part of the work in delivering on our proposal will focus on how to make some specific

[34] `https://en.wikipedia.org/wiki/Boaty_McBoatface` – last verified February 2018.

NLP projects appealing (whether through entertainment value, through the self-satisfaction that arises from acts of altruism, or appeals to pure altruism), beyond making them open to the general public, for instance:

- Building demonstrations of what AI can do with language; famous examples include IBM Watson Jeopardy Challenge [4] and Personality Insights [27].
- "Serious Games" can be designed to entertain and educate, but games can even be used to collect data resources in original ways [24].
- Determining how to ensure that data obtained from crowd sourcing is reliable; see e.g. [18].
- Establishing through public discussion and debate the extent to which participation in citizen science is a public duty (in the spirit of recycling or avoiding littering).[35]

3.2.4 Design: Flexibility and Sustainability

The proposal is oriented towards leveraging existing software tools and encouraging the production of new ones. This perspective entails a few strong characteristics which differ significantly from more traditional software-oriented projects. The core part of the development does not take place in-house but in many independent projects with their own goals, design choices and coding practices. Thus the governance of the proposal aims at generating added value from multiple independent projects, by focusing on helping them grow a sustainable community of contributors as well as encouraging potential combinations and applications of projects (these two aspects being deeply interdependent). Consistently with this philosophy, the proposal favours a broad-range strategy based on the diversity of tools (in type, quality, level of development, etc.), as opposed to a centralized strategy concentrated on a closed set of tools.

An important consequence of this diversity strategy is that it does not enforce any framework or compatibility requirements, in order not to add constraints to the projects. This unusual feature deserves a particular explanation: the lack of strong standards has been seen as an issue in NLP for a long time; despite regular attempts based on various software frameworks (e.g. Gate,[36] Stanford CoreNLP,[37] Apache UIMA,[38] etc.), the NLP research community has consistently kept working for the most part with simple file formats[39] (frequently raw text or CSV files, xml when needed), on an ad-hoc basis depending on the task at hand. In other words, the community tends to favour simplicity with lightweight task-oriented components over inter-component compatibility (inter-operability); we argue that this is not by lack of standardization effort or lack of good standard, but on the contrary that this is a meaningful choice motivated by the complexity of language-related tasks. As a consequence, we choose to adopt this lack of standard as a feature for the proposal, especially since this fits particularly well with our goal of maximal flexibility and design choices on a project basis. This does not preclude the adoption of a particular standard, framework or platform by some software components. Instead, in this perspective we argue that standardisation should happen downstream, i.e. after the "raw" component has been developed and tested in various contexts; the adaptation of the component to some framework is seen as being part of the late phase of developing applications for it: at this stage, the component can be integrated into various forms which fit its potential uses. Decoupling the development of the core from the integration into a final larger system has the additional advantage of allowing different people to do these different tasks, i.e. not leaving the developer of the core necessarily in charge of the integration of their component. Again this fits well with the philosophy of the proposal, where a community of contributors can participate in different parts of the proposal. Of course, optional guidelines can still be proposed to help project developers increase the compatibility of their project and consequently its potential for reuse.

This voluntarily agnostic approach follows the principle that the general ecosystem should be very flexible about the kind of resources which are accepted, as long as they can be evaluated, explained, examplified through reproducible experiments; it is crucial that the process does not add constraints over these basic criteria. Similarly in terms of technology development level, any software starting from TRL 3 (proof of concept software, see §3.1.5) is a valid candidate, since the goal is to promote cutting-edge research methods, not final-product applications.

[35] One may anticipate a robust multi-disciplinary debate on this topic given a public perception that industrial pollution is at times treated with greater laxity than public littering.

[36] `https://gate.ac.uk/` – last verified: February 2018.

[37] `https://stanfordnlp.github.io/CoreNLP/` – last verified: February 2018.

[38] `https://uima.apache.org/` – last verified: February 2018.

[39] As seen for example in the vast majority of the shared tasks organized by the community.

3.2.5 Core Components of the Proposal

Due to its evolving nature and broad scope, the proposal cannot easily be broken down into clearly defined subtasks.[40] However one can roughly classify the nature of the actual tasks of the proposals into the four following categories (the order of which is not relevant):

- **Referencing Projects.** It is of course unavoidable for this project to progressively build a collection of open-source projects related to AI research. The exact referencing process is going to be defined progressively over time. It involves identifying projects (active or not) which fall inside the scope of the project of course, but also organizing the information in a way which facilitates the exploration of potential connections between projects. Scientific literature is naturally the main sources of pointers to software resources; some existing resources might also help in this task, e.g.:
  - ELRA[41] proposes a catalogue of language resources[42] identified through an "International Standard Language Resource Number" (ISLRN)[43].
  - Github,[44] the largest open-source repository, offers various means to explore projects: highlighted projects grouped into collections, topic tags, user ratings, etc.
- **Projects Development.** Development takes place in relation with one or several projects in accordance with the objectives of the project. Naturally a participant's role can be to engage in the core development of a particular project, but it can also involve developing extensions or enhancements to make the software usable in some a particular context, for example:
  - Testing the component for a particular task (which may or may not have been the task for which the component was originally designed);
  - Using the component to run or reproduce an experiment;
  - Combining the component with some other component(s) to achieve a different task;
  - Implementing an interface for a new usage, e.g. demonstrating the component in a webpage;
  - Make the component more user-friendly, improve the documentation, write a tutorial.
- **Outreach.** Outreach can take many different forms:
  - Social media, open-source discussion lists, online specialized press, casual conversation during human interaction in social activities, etc. The outreach strategy typically depends on the project or technology being advertised;
  - Documentation, tutorials and software demonstrations are essential to attract interest and encourage people to use and/or get involved with a project;
  - Community management: helping interested people finding a project and/or a task that they like is important for the projects to grow; sometimes it can be useful to help old or even abandoned projects finding new contributors by advertising them;
  - Community feedback: providing means for members of the community to see accurate (yet individual privacy preserving) accounts of the extent to which the rest of the community is contributing;
  - The organization of scientific events (conferences, journals) is clearly also an important part of the outreach strategy (see §3.2.2).
- **Research Contributions.** As explained in §2.2, the project is expected to produce scientific contributions in the field of NLP, in particular by improving reproducibility and providing the field with better research instrument (both in quantity and quality).

4 Impact

The impact of the social and educational side of the proposal can be measured with indicators based on the participation in its constituent projects, such as number of participants, number of projects, etc. However it is important to bear in mind that this project goes beyond short-term quantifiable results, with potentially long-term benefits in many areas.

[40] This belongs to the next stage of designing and implementing the proposed plan, taking into account available resources and specific targets.

[41] `http://www.elra.info` – last verified: February 2018.

[42] `http://catalog.elra.info/` – last verified: February 2018.

[43] `www.islrn.org` – last verified: February 2018.

[44] `github.com` – last verified: February 2018.

4.1 Long-Term Benefits of Involving Society with AI

4.1.1 Education by Exploration

A few decades from now, the proportion of people with programming skills will be much higher. It is even not absurd to imagine a future in which basic algorithmic skills have become as important as reading and writing skills.[45] Not everybody will be willing to participate to open-source AI projects, but it seems reasonable to assume that some people will be interested. In a long-term perspective, enabling and supporting this change now can only amplify its positive effects in the future, by making more and more people aware of this possibility. The educational side of the project is characterized by hands-on experience, i.e. it is clearly not designed as a course at all but rather encourages people to participate to discussions and to contribute in the way they feel the most interested. People are offered the possibility to open the "black box" of AI in an open and flexible way, to the extent of their choice depending on their motivations.

4.1.2 Creativity

AI is currently being designed and developed mostly by scientists and industrials, often with commercial applications in mind. The inclusion of a more diverse public in the development process of AI might open new perspectives of applications. By providing access to a broad range of diverse tools as well as their code, the project empowers people to develop their own variants of AI tools as well as to combine existing tools into new complex tools for any application. This could give rise to entirely new uses; it is important to emphasize that as opposed to more standard approaches (e.g. [21], [28]), our approach does not provide a limited collection of pre-made tools, but is meant to give access to every aspect of a much larger range of tools; this entails more effort on the part of the user, but it also offers much more freedom for them to build any customized AI system. This approach fits in a much wide perspective where building an AI system can be seen as a creative process which consists in the combination of smaller building blocks; the goal does not have to be utilitarian and is bound only by imagination (for instance in arts [17]).

4.1.3 Multilinguality

Major progress has been achieved in multilinguality in the recent years. In NLP applications, scalability in terms of data size has been addressed for the most part, but scalability in terms of language diversity is still a significant challenge. As of version 2.1, the Universal Dependencies corpus [38] includes 102 annotated datasets and 59 distinct languages,[46] thanks to the authors' and contributors' great effort. Packaging such a diversity of languages in a uniform format is a major step towards the ability to process multiple languages in an homogeneous way, which is the cornerstone of language-wise scalability. Researchers [5, preface] have claimed that *"Previously, to build robust and accurate multilingual natural language processing (NLP) applications, a researcher or developer had to consult several reference books and dozens, if not hundreds, of journal and conference papers."* One might add that said researcher or developer would also have to find, test and integrate multiple language-specific software tools. Thus, language scalability also requires a more streamlined engineering process: it becomes impractical to find a specific software tool for every language to process, let alone the best tool for every specific language. Instead, evaluating software tools is progressively shifting from accuracy in a specific language to robustness and adaptability to a wide range of languages.

The NLP community as well as institutions[47] are actively addressing multilinguality, since this is one of the main challenges for NLP technology to become widespread and really useful to society. In the approach that we propose, multilinguality is implied as it is a valuable consequence of the broad diversity of NLP tools that we promote and of the opening to general public contributions. In particular, the adaptation of NLP tools to various languages generally requires annotated data in the target language; given the chance, many people would certainly be happy to help preserving and promoting their own language by contributing to building linguistic resources. This would in turn help the NLP community provide them with better quality tools adapted to their language.

[45] It is important to keep in mind that the level of education of a population can evolve drastically over a few generations, as history shows: only one third of the world population was literate in 1950, and this proportion raised to 85% over sixty years [43]. Similarly, the proportion of a generation achieving tertiary education level has doubled in most western countries in the past 30 years [40].

[46] `http://universaldependencies.org/` – last verified: February 2018.

[47] For instance: `http://mlp.computing.dcu.ie/`, `http://www.meta-net.eu` – last verified: February 2018.

4.1.4 Opening the AI Black Box

As explained in §2.1.1, societal and ethics questions around AI are becoming crucially important for the future of the society. Many aspects of AI (e.g. [1], [20], [46]) require informed choices from the society; the general public should be able to question AI research blind spots. Thus it is vital that AI methods are made open to public scrutiny, and to provide the general public with means to get an understanding and even a say in the evolution of AI.

From an ethical point of view, the lack of transparency of modern ML techniques is questionable: e.g. [7], [9]; but simplistic expectations about total visibility are also unhelpful [**?**]. This proposal offers new kinds of actor-network configurations that could produce different starting points for understanding how transparency relates to accountability. AI experts are experts in their field, most notably in Machine Learning (ML), but can be subject to unconscious biases like anyone (see e.g. [15], [19], [26]). Incorporating non-expert and diverse participants in NLP research helps to deepen the potential for ethical reflection from the outset while encouraging the development of alternative evaluation strategies around the values embedded in research. Also, by facilitating a better understanding of AI, it allows researchers to take responsibility for the way research is presented to and understood by public [**?**].

4.2 Indirect Benefits to Research and Industry

The approach that we propose can be seen as a seed for exploring alternative research opportunities, and diversifying/enriching innovation in new economic areas.

The availability of a well tested and maintained collection of tools is an advantage for the research community and the whole innovation chain. The process of testing, using and extending NLP tools can lead to discoveries or improvements. For instance, testing a tool on some benchmark dataset may be published in a paper comparing the results to existing methods. Combining tools for some new task or comparing tools against each other are also likely to lead to research contributions. Overall, the richness and diversity of the tools makes it easier and faster to test different ideas or methods. This contributes to a more efficient process of selecting the best approach for a task, e.g. by filtering out unsuccessful ideas sooner. This means that the good ideas can be developed faster, hence a gain in research and development productivity.

This is true in particular for companies which do not have the resources of the technology giants: by making software tools accessible, such companies could afford to test new methods without requiring the expensive in-house knowledge of AI experts; this means that the approach we propose can lower the skill barriers for SMEs. A similar argument can be made for companies/organizations which deal with sensitive data and therefore cannot make their data available to external experts; by providing them with open software tools, they can experiment in-house instead of long and costly IP processes and/or confidentiality agreements.

5 Final remarks

The proposal for advancing NLP described above in some sense suggests a return to basics. In a simplifying analogy, we propose systematizing the reporting of knowledge and applications of NLP advances in an accessible manner in the way that Wikipedia enables contributing to archives of specific slices of general knowledge. Through this, we also propose revisiting some of the "received wisdom" in the field to ensure that assumptions, theories and technologies retain internal and external validity on further inspection, and to make them available for exaptation to other problems. While we ourselves have contributed modestly to the field's advances (for example, in parsing [42], in grammatical inference [32], etc.), we have also participated in "revisiting" activities (for example, among other results, in assessing syntactic expressivity/complexity results [52], combining string similarity measures [36], assessing quality estimation methods [33], [34], etc., and perhaps most importantly to our argument in this paper, a revisiting of the tokenization problem [35][48]) our proposal is much bigger than our own lab, and requires collaboration as a "big team" along the lines that we have described above: we hereby invite others who share productive interests in this domain (from whatever sector of activity) to make contact with us towards coordinating the proposed endeavour. More NLP components than any individual can master require advancing and revisiting in the manner that we propose. Further, additional AI domains merit analysis using the same paradigm.

[48] This is arguably most important because tokenization, individuation of the linguistic atoms to be analyzed in a text is the first step in the prototypical NLP pipeline.

Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

[1] Anonymous. AI image recognition fooled by single pixel change. *bbc.com*, November 2017. http://www.bbc.com/news/technology-41845878.

[2] J. Bardzell. Interaction criticism and aesthetics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2357–2366, New York, NY, USA, 2009. ACM.

[3] Y. Benkler. Freedom in the Commons: Towards a Political Economy of Information. *Duke Law Journal*, 52:1245–1276, 2003. https://scholarship.law.duke.edu/dlj/vol52/iss6/3.

[4] J. Best. IBM Watson: The Inside Story of how the Jeopardy-Winning Supercomputer was Born, and What it Wants to do Next. *TechRepublic*. https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-r

[5] D. Bikel and I. Zitouni. *Multilingual Natural Language Processing Applications: From Theory to Practice*. IBM Press, 1st edition, 2012.

[6] J. Bonnefon, A. Shariff, and I. Rahwan. Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? *CoRR*, abs/1510.03346, 2015.

[7] A. M. Bornstein. Is Artificial Intelligence Permanently Inscrutable? *Nautilus*, September 2016. http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable.

[8] K. A. Brekke, G. Kipperberg, and K. Nyborg. Social interaction and responsibility ascription: The case for household recycling. *Land Economics*, 86(4):766–784, 2010.

[9] D. Castelvecchi. Can we open the black box of AI? *Nature*, 1(538):20–23, October 2016. http://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

[10] C. Cookson. DeepMind computer teaches itself to become world's best Go player. *Financial Times*, October 2017. https://www.newscientist.com/article/2132086-deepminds-ai-beats-worlds-best-go-player-in-latest-face-off.

[11] European Commission. Key Enabling Technologies. https://ec.europa.eu/growth/industry/policy/key-enabling-technologies_en.

[12] European Commission. Horizon 2020 – Work Programme 2014-2015, General Annexes, G. Technology readiness levels (TRL), July 2014. https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf.

[13] N. Evans and S. C. Levinson. The myth of language universals: Language diversity and its im portance for cognitive science. *Behavioral and Brain Sciences*, 32:429–492, 2009.

[14] K. Fogel. *Producing Open Source Software: How to Run a Successful Free Software Project*. O'Reilly Media, second edition, Jan. 2017. http://www.producingoss.com/.

[15] A. Fokkens, M. van Erp, M. Postma, T. Pedersen, P. Vossen, and N. Freire. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1691–1701, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[16] B. Friedman. Value-sensitive design. *interactions*, 3(6):16–23, Dec. 1996.

[17] M. Gayford. Robot Art Raises Questions about Human Creativity. *MIT Technology Review*, February 2016. https://www.technologyreview.com/s/600762/robot-art-raises-questions-about-human-creativity.

[18] Y. Graham, Q. Ma, T. Baldwin, Q. Liu, C. Parra, and C. Scarton. Improving evaluation of document-level machine translation quality estimation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 356–361, Valencia, Spain, 2017. Association for Computational Linguistics.

[19] D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, February 2006. https://doi.org/10.1214/088342306000000060.

[20] K. Kelly. The Myth of a Superhuman AI. *Wired*, April 2017. https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai.

[21] W. Knight. Google's Self-Training AI Turns Coders into Machine-Learning Masters. *MIT Technology Review*, January 2018. https://www.technologyreview.com/s/609996/googles-self-training-ai-turns-coders-into-machine-learning-masters.

[22] A. Kroulek. Crowd-Sourced Translation Goes Awry For Facebook. *k international: the Language Blog*, August 2010. https://www.k-international.com/blog/wrong-translation-for-facebook/.

[23] T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

[24] M. Lafourcade, A. Joubert, and N. Le Brun. *GWAPs for Natural Language Processing*, pages 47–72. John Wiley & Sons, Inc., 2015.

[25] C. Le Dantec, E. Poole, and S. Wyche. Values as Lived Experience: Evolving Value Sensitive Design in Support of Value Discovery. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 1141–1150, 2009.

[26] O. Levy, Y. Goldberg, and I. Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[27] T. Lewis. IBM's Watson says it can analyze your personality in seconds — but the results are all over the place. *Business Insider UK*, July 2015. http://uk.businessinsider.com/ibms-supercomputer-can-now-analyze-your-personality-based-on-a-writing-sample-heres-how-you-try-it-2015-

[28] I. Lunden. AWS ramps up in AI with new consultancy services and Rekognition features. *TechCrunch*, November 2017. https://techcrunch.com/2017/11/22/aws-ai/.

[29] M. Masterman. The nature of a paradigm. In I. Lakatos and A. Musgrave, editors, *Criticism and the Grown of Knowledge*, pages 59–89. Cambridge University Press, 1970.

[30] R. McMillan. AI has arrived, and that really worries the world's brightest minds. *Wired*, January 2015. https://www.wired.com/2015/01/ai-arrived-really-worries-worlds-brightest-minds.

[31] J. Metcalf, E. F. Keller, and D. Boyd. Perspectives on Big Data, Ethics, and Society, White Paper, 2017.

[32] E. Moreau. Identification of Natural Languages in the Limit: Exploring Frontiers of Finite Elasticity for General Combinatory Grammars. In *12th conference on Formal Grammars (FG 2007)*, page Online Proceedings, Dublin, Ireland, France, Aug. 2007. CSLI Publications Online Proceedings.

[33] E. Moreau and C. Vogel. Weakly Supervised Approaches for Quality Estimation. *Machine Translation*, 27(3):pp 257–280, Sept. 2013.

[34] E. Moreau and C. Vogel. Limitations of MT Quality Estimation Supervised Systems: The Tails Prediction Problem. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 2205–2216, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics.

[35] E. Moreau and C. Vogel. Multilingual word segmentation: Training many language-specific tokenizers smoothly thanks to the universal dependencies corpus. In *LREC2018: 11th edition of the Language Resources and Evaluation Conference*, 2018. to appear.

[36] E. Moreau, F. Yvon, and O. Cappé. Robust Similarity Measures for Named Entities Matching. In *COLING 2008*, pages 593–600, Manchester, United Kingdom, Aug. 2008. ACL.

[37] A. A. Newman. Translators Scoff at LinkedIn's Offer of $0 an Hour. *New York Times*, June 2009. http://www.nytimes.com/2009/06/29/technology/start-ups/29linkedin.html.

[38] J. Nivre, Ž. Agić, L. Ahrenberg, M. J. Aranzabe, M. Asahara, A. Atutxa, M. Ballesteros, J. Bauer, K. Bengoetxea, R. A. Bhat, E. Bick, C. Bosco, G. Bouma, S. Bowman, M. Candito, G. Cebiroğlu Eryiğit, G. G. A. Celano, F. Chalub, J. Choi, Ç. Çöltekin, M. Connor, E. Davidson, M.-C. de Marneffe, V. de Paiva, A. Diaz de Ilarraza, K. Dobrovoljc, T. Dozat, K. Droganova, P. Dwivedi, M. Eli, T. Erjavec, R. Farkas, J. Foster, C. Freitas, K. Gajdošová, D. Galbraith, M. Garcia, F. Ginter, I. Goenaga, K. Gojenola, M. Gökırmak, Y. Goldberg, X. Gómez Guinovart, B. Gonzáles Saavedra, M. Grioni, N. Grūzītis, B. Guillaume, N. Habash, J. Hajič, L. Hà Mỹ, D. Haug, B. Hladká, P. Hohle, R. Ion, E. Irimia, A. Johannsen, F. Jørgensen, H. Kaşıkara, H. Kanayama, J. Kanerva, N. Kotsyba, S. Krek, V. Laippala, P. Lê Hồng, A. Lenci, N. Ljubešić, O. Lyashevskaya, T. Lynn, A. Makazhanov, C. Manning, C. Mărănduc, D. Mareček, H. Martínez Alonso, A. Martins, J. Mašek, Y. Matsumoto, R. McDonald, A. Missilä, V. Mititelu, Y. Miyao, S. Montemagni, A. More,

S. Mori, B. Moskalevskyi, K. Muischnek, N. Mustafina, K. Müürisep, L. Nguyễn Thị, H. Nguyễn Thị Minh, V. Nikolaev, H. Nurmi, S. Ojala, P. Osenova, L. Øvrelid, E. Pascual, M. Passarotti, C.-A. Perez, G. Perrier, S. Petrov, J. Piitulainen, B. Plank, M. Popel, L. Pretkalniņa, P. Prokopidis, T. Puolakainen, S. Pyysalo, A. Rademaker, L. Ramasamy, L. Real, L. Rituma, R. Rosa, S. Saleh, M. Sanguinetti, B. Saulīte, S. Schuster, D. Seddah, W. Seeker, M. Seraji, L. Shakurova, M. Shen, D. Sichinava, N. Silveira, M. Simi, R. Simionescu, K. Simkó, M. Šimková, K. Simov, A. Smith, A. Suhr, U. Sulubacak, Z. Szántó, D. Taji, T. Tanaka, R. Tsarfaty, F. Tyers, S. Uematsu, L. Uria, G. van Noord, V. Varga, V. Vincze, J. N. Washington, Z. Žabokrtský, A. Zeldes, D. Zeman, and H. Zhu. Universal dependencies 2.0. 2017. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

[39] R. Nuzzo. How scientists fool themselves – and how they can stop. *Nature*, October 2015. `http://www.nature.com/news/how-scientists-fool-themselves-and-how-they-can-stop-1.18517`.

[40] OECD Data. Population with tertiary education, 2017. `https://data.oecd.org/eduatt/population-with-tertiary-education.htm`.

[41] J. Pearson. Why An AI-Judged Beauty Contest Picked Nearly All White Winners. *Motherboard*, September 2016. `https://motherboard.vice.com/en_us/article/78k7de/why-an-ai-judged-beauty-contest-picked-nearly-all-white-winners`.

[42] F. Popowich and C. Vogel. A logic based implementation of head-driven phrase structure grammar. In C. Brown and G. Koch, editors, *Natural Language Understanding and Logic Programming, III*, pages 227–246. Elsevier, North-Holland, 1991.

[43] M. Roser and E. Ortiz-Ospina. Literacy, 2017. `https://ourworldindata.org/literacy/`.

[44] I. Sample. Artificial Intelligence risks GM-style public backlash, experts warn. *theguardian.com*, November 2017. `https://www.theguardian.com/science/2017/nov/01/artificial-intelligence-risks-gm-style-public-backlash-experts-warn`.

[45] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[46] J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980.

[47] P. Sengers, K. Boehner, S. David, and J. J. Kaye. Reflective design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, CC '05, pages 49–58, New York, NY, USA, 2005. ACM.

[48] K. Shilton and S. Anderson. Blended, not bossy: Ethics roles, responsibilities and expertise in design. *Interacting with Computers*, 29(1):71–79, 2017.

[49] J. Sneddon. Why Linux Users Make the Most Valuable Customers. *OMG! Ubuntu!*, July 2017. `http://www.omgubuntu.co.uk/2017/07/linux-users-are-more-valuable-customers`.

[50] J. Titcomb. 'Facebook is listening to me': Why this conspiracy theory refuses to die. *telegraph.co.uk*, October 2017. `http://www.telegraph.co.uk/technology/2017/10/30/facebook-listening-conspiracy-theory-refuses-die`.

[51] A. van Wynsberghe and S. Robbins. Ethicist as designer: A pragmatic approach to ethics in the lab. *Science and Engineering Ethics*, 20(4):947–961, Dec 2014.

[52] C. M. Vogel, U. Hahn, and H. Branigan. Cross-serial dependencies are not hard to process. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 157–62, 1996. COLING'96, Copenhagen, Denmark.

[53] R. Yuste, S. Goering, B. A. y Arcas, G. Bi, J. M. Carmena, A. Carter, J. J. Fins, P. Friesen, J. Gallant, J. E. Huggins, J. Illes, P. Kellmeyer, E. Klein, A. Marblestone, C. Mitchell, E. Parens, M. Pham, A. Rubel, N. Sadato, L. S. Sullivan, M. Teicher, D. Wasserman, A. Wexler, M. Whittaker, and J. Wolpaw. Four ethical priorities for neurotechnologies and AI. *Nature*, 1(551):159–163, November 2017. `https://www.nature.com/news/four-ethical-priorities-for-neurotechnologies-and-ai-1.22960`.