LREC 2018 Workshop

Improving Social Inclusion using NLP: Tools, Methods and Resources ISI-NLP 2

PROCEEDINGS

Edited by

Ineke Schuurman, Leen Sevens, Victoria Yaneva, John O'Flaherty

ISBN: 979-10-95546-12-2
EAN: 9791095546122

7 May 2018

Proceedings of the LREC 2018 Workshop
"Improving Social Inclusion using NLP: Tools, Methods and Resources"
(ISI-NLP 2)
7 May 2019 – Miyazaki, Japan

Edited by Ineke Schuurman, Leen Sevens, Victoria Yaneva, John O'Flaherty

http://www.ccl.kuleuven.be/ISINLP2

Organising Committee

- Ineke Schuurman, KU Leuven (BE)
- Leen Sevens, KU Leuven (BE)
- Victoria Yaneva, University of Wolverhampton (UK)
- John O'Flaherty, The National Microelectronics Applications Centre (IE)

Programme Committee

- Bram Bulté, KU Leuven (BE)
- Heidi Christensen, University of Sheffield (UK)
- Onno Crasborn, Radboud University (NL)
- Orphée De Clercq, Ghent University (BE)
- Kris Demuynck, Ghent University (BE)
- Koenraad De Smedt, University of Bergen (NO)
- Camilla Lindholm, University of Helsinky (FI)
- Peter Ljunglöf, University of Gothenburg (SE)
- Isa Maks, VU University Amsterdam (NL)
- Davy Nijs, UC Leuven-Limburg (BE)
- John O'Flaherty, The National Microelectronics Applications Centre (IE)
- Martin Reynaert, Tilburg University and Radboud University (NL)
- Horacio Saggion, Universitat Pompeu Fabra (ES)
- Ineke Schuurman, KU Leuven (BE)
- Leen Sevens, KU Leuven (BE)
- Liz Tilly, University of Wolverhampton (UK)
- Vincent Vandeghinste, KU Leuven (BE)
- Hugo Van hamme, KU Leuven (BE)
- Victoria Yaneva, University of Wolverhampton (UK)

Preface

Social media are an essential component of the XXI century information society, however, and in spite of their wide adoption they still present many barriers for specific types of users. Information shared in applications such as Twitter, Facebook, or Instagram, just to name a few, is far too complicated to be understood by people with special needs such as people with intellectual and/or developmental disabilities, like Fragile X-syndrome, Down syndrome, Specific Language Impairment, dementia, but also for people with limited communication skills due to illness or accident. It can also be problematic for people such as immigrants who want to be integrated in the digital society of their host country but do not master the language of their new home. Speakers of under-resourced languages wanting to communicate in their own language are likely to face problems as well.

Several technologies can make a difference in making information more accessible for different types of users. For example, a complicated text can be converted into a simpler version by the application of lexical or syntactic simplification or extra linguistic information such as definitions can be used to clarify the content. In the case of people who are functionally illiterate, augmentative and alternative non-verbal input methods can be automatically converted to natural language and provide a means to take part in social interaction. Hard to understand user generated texts, which usually contain abbreviation and social media jargon, can be normalized to make them more accessible.

The objective of the second Workshop on Improving Social Inclusion using Natural Language Processing is to bring together researchers and practitioners in the areas of social inclusion and natural language processing to understand problems faced by different social groups with text accessibility in social media, describe current development in language resources and methods for this problem, and discuss future research directions. Of particular interest if how techniques and resources developed for one language and domain could be ported to a different language.

We thank all the authors for their contributions and also the LREC 2018 Conference for hosting our workshop.

Ineke Schuurman
Leen Sevens
Victoria Yaneva
John O'Flaherty

Programme

14:00 – 14:20 Introduction

Oral Presentations

14:20 – 14:45 Anna Matamala, Pilar Orero, Sara Rovira-Esteva, Helena Casas-Tost, Luis Fernando Morales, Olga Soler-Vilageliu, Belén Agulló, Anita Fidyka, Daniel Segura and Irene Tor-Carroggio
User-centric approaches in Access Services Evaluation: Profiling the End User

14:45 – 15:10 Lilia Georgieva
Digital Inclusion And The Elderly: The Case of Online Banking

15:10 – 15:35 Sanja Štajner
How to Make Troubleshooting Simpler? Assessing Differences in Perceived Sentence Simplicity by Native and Non-native Speakers

15:35 – 16:00 Rodolfo Zevallos, Luis Camacho, Ronald Cardenas and Reynaldo Baquerizo
Siminchik: A Speech Corpus for Preservation of Southern Quechua

Coffee break

Invited Talk and Panel Discussion

16:30 – 17:20 Pawel Kamocki (ELRA)
"No injury is done to a willing person?" Legal aspects of personal data collection and sharing for research purposes

17:20 – 17:50 Panel Discussion (*topics to be announced*)

17:50 – 18:00 Closing

***

ISI-NLP2 and LiNCR are collaborating to jointly support the invited talk and panel, please refer to ISI-NLP2 and LiNCR websites for more details

Table of Contents

User-centric Approaches in Access Services Evaluation: Profiling the End User

Anna Matamala, Pilar Orero, Sara Rovira-Esteva, Helena Casas-Tost, Luis Fernando Morales, Olga Soler-Vilageliu, Belén Agulló, Anita Fidyka, Daniel Segura, Irene Tor-Carroggio
Universitat Autònoma de Barcelona
MRA/126, 08193 Bellaterra (Barcelona)
{anna.matamala, pilar.orero, sara.rovira, helena.casas, fernando.morales, olga.soler, belen.agullo, anita.fidyka, daniel.segura, irene.tor}@uab.cat

Abstract

This article presents best practices for the design of a user-centric approach in accessibility research projects, taking two European projects as an example: ImAc (Immersive Accessibility) and EasyTV (Easing the access of Europeans with disabilities to converging media and content). Both H2020 projects aim to investigate access services such as audio description, audio subtitling, subtitling/captioning or sign language interpreting in media content. The paper explains the many and varied documentation required to comply with existing ethical issues in Europe. Designing alternative means of obtaining information will be explained, since interaction with participants had to cater for the needs of diverse users. The second part of the article presents an overview of user profiling in previous accessibility projects in the field of media accessibility, and shows good practices based on the two ongoing projects. Finally, the article presents an example of how users have been defined and how they have been involved in the initial stages of both ImAc and EasyTV, by summarizing the methodology developed for a series of focus groups with end users. The article concludes with some recommendations when involving users in accessibility research.

Keywords: media accessibility, user-centric methodologies, user profiling

1. Introduction

ImAc (Immersive Accessibility) and EasyTV (Easing the access of Europeans with disabilities to converging media and content) are two European projects funded by the European Commission in the ICT19 2016 call. These projects, which started in October 2017, research access services in media content, both in traditional and social media. Both projects aim to improve social inclusion by offering accessible media content through four access services: audio description, audio subtitling, sign language interpreting, and subtitling/captioning.

Working within the United Nations Convention on the Rights of Persons with Disabilities (CRPD) paradigm means to consult end users, following the "nothing about us without us" approach. Therefore, taking into account the access services researched, users involved in the tests and applications of both projects are mainly people with visual and hearing impairments.

In order to enhance the user experience, both projects take a user-centric approach when defining system requirements and when testing the technologies and services to be implemented (Harte et al., 2017). Adopting a user-centric approach requires that all procedures, from user selection to results dissemination, comply with ethical requirements. Ethical considerations in human research have been a major concern since the critical articles written by Pappworth (1967) on medical research. They are nowadays regulated by official guidelines that take into account not only the participant well-being (Human Subjects Protection, HSP), but also the communities where both participants and researchers belong to (Community-Engaged Research (CEnR) (Ross et al., 2010; Singleton et al., 2015).

Regarding ImAc and EasyTV, the research group TransMedia Catalonia (grupsderecerca.uab.cat/transmedia/) at Universitat Autònoma de Barcelona (UAB) is in charge of developing the methodology for user testing, due to the previous experience gained testing diverse users in European projects such as DTV4ALL (Romero-Fresco, 2015) or HBB4ALL (Orero et al., 2015). This paper aims to describe the first stages taken in both projects, the challenges found, and how they have been overcome. More specifically, the paper describes how ethical issues have been fulfilled and how alternative means of obtaining informed consent have been developed to comply with current European legislation. It also describes how user profiling has been carried out in previous projects and the approach taken in ImAc and EasyTV. Finally, an example of how users have been defined and involved during the initial stages of the projects is presented. In sum, this paper aims to describe what could be termed best practices for profiling end users in international accessibility-related projects.

1.1 The ImAc Project

ImAc (www.imac-project.eu) is a 30-month project funded by the European Commission aiming to research how access services can be integrated in immersive media, more specifically in 360º content. The aim is to move from current technology to hyper-personalised environments where end users can customise their experiences to meet their needs.

ImAc challenges existing subtitling guidelines (BBC, 2017; Díaz-Cintas and Remael, 2007) drafted for non-immersive media content, i.e. standard movies or TV content. ImAc also poses a new challenge to audio description since sound is immersive, and many audio description solutions can be delivered by object-based sound. Sign language interpreting also defies current practices and guidelines, because it becomes an immersive picture in immersive content.

To work with such complex human and technical challenges, the project consortium is multidisciplinary. The partners include, on the one hand, technical experts, concerned with the development of platforms, players, user interfaces, and access services production tools, namely subtitle production tools, audio production tools, and a sign language editor. On the other hand, academic partners with a background in humanities and social sciences are involved with user experience establishing use cases and user requirements, as well as running pilots to test project results with real end users: persons with disabilities.

1.2 The EasyTV Project

EasyTV is a 30-month project funded by the European Commission aiming to offer easier access to converging media content to persons with disabilities. The project will work on different accessibility-related aspects. First of all, it will work on improving access services by focusing on two aspects: image adaptation, by presenting contrast/edge enhancement or modification, and improved content description, by developing narratives which can be adapted to different playing paces and by providing a cleaner audio. Secondly, the project will work towards improved personalisation of content experiencing and interaction. It will include the development of an "auto-personalisation-from-profile". User profiles together with context information (application, device, content) will allow for assistive technologies and automatic user interface personalisations and adaptations. EasyTV will demonstrate the ability of cloud-based hyper-personalisation to automatically turn on and configure accessibility features built into different TV operating systems, applications and embedded devices. A third area in which EasyTV will focus is the development of novel technologies to break the sign language barrier. Work will be carried out on translation in different sign languages, through a multilingual ontology that will map signs to ontology concepts and realistic sign language avatar animations. Additionally, crowdsourcing technology is going to be developed in combination to the sign language avatar animations. This will allow non-professional users to contribute with their own translations of audiovisual content to several sign languages and share the resulting avatar animations. These new technologies could significantly increase current sign language offer in the media. Finally, a last outcome of the project will be the improvement and development of voice and gesture/gaze recognition to control the TV set and TV applications.

2. Ethical Procedures

Ethics is an integral part of research and projects need to comply with existing regulations and codes of conduct. This is especially relevant when so-called vulnerable populations such as persons with disabilities are involved. According to the European textbook on ethics research, vulnerability is a very complex concept but when "the voluntariness of the subject's consent is compromised, this may similarly prevent them from choosing to give or withhold consent in a way that would protect their interests" (European Commission, 2010: 53). The same document goes on to acknowledge that the "physical (or psychological) condition of some subjects leaves them especially liable to harm, for example as a result of frailty through age, disability, or illness" (*ibid*). To cater for the needs of vulnerable populations, generally due to a disability or to age, both ImAc and EasyTV have followed some specific procedures that can be regarded as best practices when doing research on accessibility with end users: first of all, special care has been taken to write down the information and consent sheets in an easy-to-understand and non-technical way. Participation in both project tests has been and will be voluntary. End users are explicitly informed they can refuse to participate or withdraw their participation at any time without any consequences. Steps are taken to ensure that participants are not subjected to any form of coercion. Participants are also informed they can request additional information about the project results in case they are interested. Although the departing point is a written document, alternative means of communicating information and obtaining consent are always planned. For instance, when participants are blind or visually impaired or have difficulties reading, an electronic form is offered or an alternative oral version is provided and recorded. When deaf participants whose mother tongue is sign language take part in experiments, alternative signed informed consent forms are also provided. Finally, due to the international nature of both projects, the needs of participants in terms of languages is also catered for, providing translations when needed. The final aim is to protect the participants' rights and to make sure that all subjects are aware of the implications of their participation in the research. It must also be highlighted that avoiding any harm that might occur and ensuring the participant health and safety is and will be a priority throughout testing. Partners have been asked to identify any potential risks their technological developments might have for different user profiles. If any risk is identified, such as motion sickness in 360º videos, participants are warned about them through the information sheets and the consent forms. Appropriate measures are always taken to guarantee the participant safety and well-being, and participants thought to be unstable or under the influence of drugs or alcohol are not admitted to the experiments. Last but not least, a crucial element for both the ImAc and EasyTV projects is that ethical forms are approved by UAB's Ethical Committee.

3. User Profiling

User profiling is often carried out through questionnaires which gather demographic information. However, deciding on the specific questions and phrasing is not always as easy as it may seem. Before developing demographic questionnaires for ImAc and EasyTV, a systematic analysis of existing questionnaires from the field of audiovisual translation (AVT) and media accessibility in which access services are tested with persons with disabilities was carried out. Special attention was paid to the following 14 user reception studies dealing with audio description which can be considered representative of recent research in the field: Fernández-Torné and Matamala, 2015; Szarkowska, 2011; Szarkowska and Jankowska, 2012; Walczak, 2010; Romero-Fresco and Fryer, 2013; Fresno et al., 2014; Fryer and Freeman, 2012; Fryer and Freeman, 2014; Szarkowska and Wasylczyk, 2014; Udo and Fels, 2009; Walczak and Fryer 2017; Walczak and Fryer 2018; Walczak and Rubaj, 2014; Chmiel and Mazur, 2012; and three experimental PhD dissertations: Fryer, 2013; Cabeza-Cáceres, 2013; and Walczak, 2017. Information from the projects DTV4ALL, HBB4ALL, ADLAB, OpenArt, and AD-Verba was also gathered for the analysis. A summary of results is presented in the following subsections.

3.1 Sex/Gender and Age

When asking about sex/gender, in the literature under analysis there is always a choice between male/female but the option of not answering the question or selecting another option is not generally included. More recent approaches to this question come from medical and health

research, where sex and gender might be of crucial importance on interpreting the experimental results. Therefore, some guidelines have been developed giving directions on how to consider gender issues at all research stages (Day et al., 2017). In ImAc and EasyTV participants will be able to select between "female/male/other/I prefer not to reply" to account for the various options users may want to choose (Zukerman, 1973). The reason to extend to four the traditional duality male/female is twofold. First, gender is not the object of study in both projects and is not expected any relevance in the participation beyond aiming for parity. Secondly, we are moving from an attributed surface-level approach towards a self-reporting or deep-level attitudinal style (Harrison et al., 1998). Since we are dealing with persons with disabilities and vulnerable groups, self-reporting is more engaging for participants since they will be able to reflect their diversity. This diversity or context may have some relevance towards attitudinal differences, which is important, since defining end user expectations is one of the objectives of the questionnaires.

In relation to age, in the investigations under analysis, it is generally asked by offering some intervals, although in some cases it can also be an open question in which a figure has to be entered. In ImAc and EasyTV we have decided to leave it as an open question, as it gives more flexibility for data analysis.

3.2 Educational Level, Occupation, and Language

Concerning the educational level, it is not always asked in the literature under analysis. When asked, the question is presented in various forms: items can be very detailed (Fernández-Torné and Matamala, 2015), a choice of three options (Szarkowska, 2011) or something in between (ADLAB project). In the current projects it has been agreed to ask about the level of finished studies, differentiating between "no studies/primary education/secondary education/ further education/university".

As for the occupation of the participants, it is not generally asked except for one study in which this was considered to be relevant. Therefore, in ImAc and EasyTV it has been decided not to ask about this.

Regarding the language participants generally use, most of the questionnaires do not refer to it. The exception are the questionnaires in DTV4ALL and the Pear Tree project. In ImAc and EasyTV, it has been considered relevant to gather information about the participants main language as this may have an impact in the reception of media content and the opinion of users on system requirements. One of the reasons for this is the fact that, for some participants, sign language (SL) is their natural way of communication. This has a direct implication in both projects. Being visual languages, SL has a special consideration in broadcast since it is considered as a video object or a picture. Specific provisions will be taken for the picture in picture (PIP) challenges arisen from offering SL services in both projects.

It must be noted that, when doing research involving deaf users, data about their native language should not be taken only as a mere demographic fact, but also as information about the participant particular needs. The reception of subtitles by deaf users might be different depending on whether the participant has prelocutive hearing impairment and their mother tongue is a sign language (McIlroy and Storbeck, 2011; Serrat-Manén, 2013). Additionally, it should be noted that just asking users about this topic will not necessarily involve that the researcher will receive the correct answer. Sign language users generally consider themselves bilingual, knowing both their SL and the oral language of their community. However, this is not always the case, as it was demonstrated in a study where some users—who considered themselves to be bilingual—made some mistakes that entailed a difference in skill between the SL and the oral language when writing answers to open-ended questions (Romero-Fresco, 2015). Due to this finding, in Miquel Iriarte (2017) the user level of written comprehension in Spanish was determined by a standardised reading proficiency test. Thus, the question on language is not straightforward when profiling users with disabilities and needs to be taken into consideration as it goes beyond demographic issues.

3.3 Disabilities

The studies under analysis show different approaches to profiling users with disabilities. How to formulate questions is very often related to the model of disability adopted (Berghs et al., 2016). The medical model tends to define "disability in terms of a biological pathology located in an individual body, which requires medical technology, medicine or rehabilitation to make a person well" (*ibid*: xix)". Yet, this model has been criticised on different grounds by activists and academics since focusing on intellectual and bodily functions this approach fails to acknowledge environmental conditioners (Marks, 1997). This approach has been shown to be beneficial to improve medical diagnosis and treatment, but it has a series of weaknesses such as the unbalanced situation between doctor/patient leading to uneven results. Doctors are the experts, whereas patients are passive and not collaborators. Doctors "fix" what is "wrong" aiming at "normality" (Edler, 2009).

The medical model of disability is often referred to as 'the old paradigm' and stands in contrast to the social model of disability. The latter, which has at least nine different versions (Mitra, 2006), believes the medical explanation is insufficient to understand the relation between people and their environment avoiding human diversity (Edler, 2009). The social model of disability "makes a distinction between disability as the experience of oppression and disadvantage and impairment as a physical, sensory, cognitive or mental health condition" (Berghs et al., 2016: xix). If someone refers to himself or herself as a disabled person, s/he is referring to his or her identification with the experience of disablement. From critical disability approaches, as Berghs et al. (2016) explain, terms such as 'differently able' are used, and disability is viewed along a continuum of human diversity. According to this approach, disability is not the result of having a physical impairment, but the failure of society to consider individual differences (Bøttcher and Dammeyer, 2016). In other words, disability is not an attribute of the individual but a creation of the social environment requiring social change (Mitra, 2006).

The social model of disability was developed against the medical model of disability; however, within Disability Studies, the social model of disability was also under scrutiny (Degener, 2016). The UN CRPD was initially drafted as a human rights document aiming to substitute

the medical model of disability for the social model of disability. Yet, according to Degener (2016), who in 2001 was a legal expert to the UN High Commissioner for Human Rights as co-author of the background study to the United Nations CRPD in 2001, the final outcome was a treaty based on the human rights model of disability. Human rights approaches, as explained by Bergh et al. (2016: xix), use "person-first definitions, such as 'persons with disabilities', establishing legal, political, cultural, social and economic rights, consistent with the normative values associated with the society within which a disabled person lives."

The International Classification of Functioning, Disability and Health (ICF), approved by the UN World Health Organization in 2001, embodies what is now called the biopsychosocial model. This is a combination of the medical and social approaches to disability (Lundälv et al., 2015). This was a response to the over-medicalisation of the International Classification of Impairments, Disabilities and Handicaps (ICIDH) and the tendency of the social model to detach disablement from its biomedical origins (Imrie, 2004). It is widely used nowadays but falls short to reproduce the social and personal context.

When designing the methodology for these two projects, we took inspiration from Amartya Sen capabilities or capability approach (Mitra, 2006), which can be applied to disability too. Under Sen's approach, disability can be understood as a deprivation in terms of what he calls capabilities (understood as "practical opportunities") or functionings (understood as "actual achievements") resulting from the interaction of an individual (a) personal characteristics (age, impairment, etc.) and (b) available goods (assets, income) and (c) environment (social, economic, political, cultural) (Mitra, 2006). Disability means lacking certain capabilities/functionings due to the interaction of the above-mentioned factors. Disability depends on whether the impairment places restrictions on the individual functionings or capabilities. It is worth to retrieve the example Mitra (2016) mentions in her article: a 19-year-old boy who suffers a brain injury is considered disabled if his practical opportunity to attend university is restricted (potential disability), in contrast to an individual with a similar basket of goods, in the same environment, and with similar personal characteristics except for the impairment. In case the 19-year-old cannot finally attend college, we would be facing actual disability but in case he finally goes to university, then he would not be considered as disabled. Thus, having a health impairment does not make a person disabled.

When designing the questionnaires for the projects, special emphasis was and will be put in formulating questions that allow us to find out where the deprivation of capabilities or functionings comes from, instead of taking for granted that the problem is the health issue. For example, reading subtitles is not related to being deaf or not, but to the user reading skills, which is closely linked to education. For this reason, and away from a medical or social classification for humans, we try to get a broader overview of user capabilities/functionings through other questions. Concerning their disability, we ask them one single question: how they "define" themselves. The choice of the verb "define" was made on purpose, to respect the user self-perception. An additional question about the age when disability started has also been added. This is due to the fact that disabilities acquired since birth or at early stages in life affect in a direct way the cognitive development and communication of an individual. There are differences, for example, in the education, language acquisition process and cognitive evolution between people with prelocutive hearing impairment and people who lost hearing at a later age (D'Albis and Collard, 2013; Orfanidou et al., 2015).

3.4 Technological Abilities

Questions about technology and audio description exposure of participants were asked in most of the questionnaires under analysis. The aim of such questions was to verify whether the participants were familiar with a given technology and service, how well they knew it, and how regularly they used it. Information about participant habits regarding consumption of audiovisual content was also a regular feature of the questionnaires, by means of closed questions or multiple-choice questions.

In the ImAc and EasyTV projects, it was decided to ask what technological devices participants use on a daily basis, giving the possibility of selecting more than one (the options being television/PC/laptop/mobile phone/tablet). In the ImAc project, they were also asked whether they have any device to access virtual reality content. It was also considered relevant for the current projects to ask about their preferred device for watching online video content. All these elements help us construct a more thorough profile of the users and their capabilities/functionings, to use Mitra's terminology.

4. Involving Users: Focus Groups in ImAc and EasyTV

In ImAc the implementation of access services in immersive media will be tested, whereas EasyTV will focus on testing user interaction on improved and customisable media access services. What is especially challenging in both projects is that access services will not be developed after the technology is fully implemented but will be discussed during and after the development, prior to implementation. At the beginning of the projects, input from a reduced number of users has been sought through focus groups, whilst pilots will aim to carry out experiments with bigger samples at a later stage.

The first step in the focus group preparation was to identify the various stages in the workflow. Four main stages were identified: content management, content production, content delivery, and presentation. The second step was to identify the users who would be interacting with access services at those different stages, and two main categories were identified: on the one hand, professional users (those who will be producing the access services) and, on the other, home users (those who will be consuming accessible media content). The third step was to identify user scenarios linked to various technological components in the different stages. The last step was to derive a list of specific questions related to the technological components to be developed. These questions were used as a guide during focus group development.

A common methodology was developed for the focus groups of both projects. Regarding ImAc, focus groups dealt with audio description and audio subtitling in the UK and Catalonia and on subtitling and sign language in

Germany and Catalonia. In EasyTV, focus groups dealt with technologies related to avatars, crowdsourcing, audio narratives, image magnification, speech recognition, gesture-gaze and sign language translation, distributed among their developers from Italy, Greece and Spain.

It was agreed that focus groups would include between 6 and 12 participants in ImAc and EasyTV, where both professional and home users with technical expertise would be involved. A balance in terms of age and gender was sought.

A specific feature of all focus groups was the final agreement in the form of a series of conclusions approved by all participants. These conclusions referred to end users wishes, expectations, needs, and recommendations in relation to the creation or consumption of access services. The logistics for being able to deliver such written conclusions on the spot included, in the case of ImAc, a team composed by: a facilitator to manage the focus group, one note-taker to take general notes, and a second note-taker to structure the results of the focus group in the form of conclusions. This was possible through sharing and editing live a common e-document. Similarly, the focus groups from EasyTV included two facilitators: one who dealt with group members and kept the discussion on track, and a second taking notes and drafting the final jointly approved conclusions.

Focus groups proved a useful tool to identify user needs and had an impact when developing user requirements. During focus groups users came up with innovative solutions and put forward challenges that academic and technological partners had not considered. The aim of this paper is not to present the results of such focus group but to highlight the usefulness and the lessons learnt when involving persons with disabilities.

5. Conclusions

This paper has presented the user-centric methodology adopted at the initial stages of two on-going European projects. Special emphasis has been put on the key issues to be taken into account when involving end users from the so-called vulnerable populations in accessibility-related research.

First of all, investigations on users with disabilities should not be developed without end users' involvement. In this regard, recruitment through user associations proves to be useful to guarantee a wider reach.

Secondly, ethical requirements should be fulfilled taken into account the needs of diverse participants in terms of communication. An informed consent cannot be considered valid if the end user cannot fully access and understand it.

Thirdly, when asking about disability, researchers should be aware of the different approaches to the concept of disability, and the consequences of their choices when phrasing the questions about disability. Moreover, they should be aware that sometimes factors beyond the specific disability may have a higher impact on the capabilities of the users. It is recommended to pilot the questionnaires with end users representative and agree on the phrasing of specific questions with them, as it was done in the projects presented in this paper.

Regarding focus groups, it was very useful to have one facilitator, one note-taker and another note-taker that summarised the conclusions of the focus groups. These conclusions were approved by participants at the end of the session, and if requested, they were also shared with them via email. This was valued by focus group participants, who had an immediate feedback on their contribution.

The room arrangement in a U-shaped form also proved to be a good practice in the case of deaf and hard-of-hearing participants, because it allowed to clearly see other persons speaking (and therefore read their lips) or see the sign language interpreter.

Although our research is limited in scope because it only deals with persons with visual or hearing impairments and because it is still at an initial stage, we hope it can contribute to define best practices in profiling end users in user-centric research projects with persons with disabilities, a field in which more extensive research is needed.

6. Acknowledgements

ImAc has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 761974. EasyTV has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 76199. This paper reflects only the authors' views. The European Commission is not liable for any use that may be made of the information contained therein.

TransMedia Catalonia is a research group funded by the Catalan Government (2017SGR113, funding from Secretaria d'Universitats i Recerca, Departament d'Empresa i Coneixement, Generalitat de Catalunya).

7. Bibliographical References

Berghs, M., Atkins, K. and Graham, H. (2016). Scoping models and theories of disability. Implications for public health research of models and theories of disabilities: a scoping study and evidence synthesis. *Public Health Research*, 4(8), pp. 23-40.

Bøttcher L. and Dammeyer J. (2016). Beyond a biomedical and social model of disability: a cultural-historical approach. *Development and learning of young children with disabilities*. *International perspectives on early childhood education and development*, 13. Cham: Springer, pp. 3-23.

BBC (2017). *Subtitle guidelines*. http://bbc.github.io/subtitle-guidelines/ (last access 24.01.2018).

Cabeza-Cáceres, C. (2013). *Audiodescripció i recepció. Efecte de la velocitat de narració, l'entonació i l'explicitació en la comprensió fílmica*. PhD dissertation. UAB.

Chmiel, A. and I. Mazur (2012). Audio description research: some methodological considerations. In E. Perego (Ed.), *Emerging topics in translation: audio description*. Trieste: EUT, pp. 57-80.

D'Albis, H. and Collard, F. (2013). Age groups and the measure of population aging. *Demographic Research*, 29, pp. 617-640.

Day, S., Mason, R., Tannenbaum, C. and Rochon, P.A. (2017). Essential metrics for assessing sex & gender Integration in health research proposals involving human participants. *PLOS ONE*, 12(8): e0182812.

https://d oi.org/10.1371/journal.pone.0182812 (last access 24.01.2018).

Degener, T. (2016). Disability in a human rights context. *Laws*, 5(3), pp. 5-35.

Díaz-Cintas, J. and Remael, A. (2007). *Audiovisual translation: subtitling*. Manchester: St. Jerome.

Edler, R. (2009). La clasificación de la funcionalidad y su influencia en el imaginario social sobre las discapacidades. In P. Brogna (Ed.), *Visiones y revisiones de la discapacidad*. México D.F.: Fondo de Cultura Económica, pp. 137-154.

European Commission (2010). *Guidance note for researchers and evaluators of social sciences and humanities research*. Retrieved from http://ec.europa.eu/research/participants/data/ref/fp7/89867/social-sciences-humanities_en.pdf (last access 22.12.17).

Fernández-Torné, A. and Matamala, A. (2015). Text-to-speech versus human voiced audio description: a reception study in films dubbed into Catalan. *The Journal of Specialised Translation*, 24, pp. 61-88.

Fresno, N., Castellà, J. and Soler-Vilageliu, O. (2014). Less is more. Effects of the amount of information and its presentation in the recall and reception of audio described characters. *International Journal of Sciences: Basic and Applied Research*, 14(2), pp. 169-196.

Fryer, L. (2013). *Putting it into words: the impact of visual impairment on perception, experience and presence*. PhD dissertation. University of London.

Fryer, L. and Freeman, J. (2012). Cinematic language and the description of film: keeping AD users in the frame. *Perspectives. Studies in Translatology*, 21(3), pp. 412-426.

Fryer, L. and Freeman, J. (2014). Can you feel what I'm saying? The impact of verbal information on emotion elicitation and presence in people with a visual impairment. In A. Felnhofer, O.D. Kothgassner (Eds.), *Challenging presence: Proceedings of the 15th international conference on presence*. Wien: facultas.wuv, pp. 99-107.

Harrison, D., Price, K. and Bell, M. (1998). Beyond relational demography: time and the effects of surface- and deep-level diversity on work group cohesion. *Academy of Management Journal*, 41(1), pp. 96-107.

Harte, R., Glynn, L., Rodríguez-Molinero, A., Baker, P. M., Scharf, T., Quinlan, L. R., and Ólaighin, G. (2017). A human-centered design methodology to enhance the usability, human factors, and user experience of connected health systems: a three-phase methodology. *JMIR Human Factors*, 4(1), e8. http://doi.org/10.2196/humanfactors.5443 (last access 24.01.2018).

Imrie, R. (2004). Demystifying disability: a review of the International Classification of Functioning, Disability and Health. *Sociology of Health & Illness*, 26(3), pp. 287-305.

Lundälv, J., Törnbom, M., Larsson, P., and Sunnerhagen, K. (2015). Awareness and the arguments for and against the international classification of functioning, disability and health among representatives of disability organisations. *International Journal of Environmental Research and Public Health*, 12, pp. 3293-3300.

McIlroy, G. and Storbeck, C. (2011). Development of deaf identity: an ethnographic study. *Journal of Deaf Studies and Deaf Education*, 16, pp. 494-511.

Marks, D. (1997). Models of disability. *Disability and Rehabilitation*, 19(3), pp. 85-91.

Miquel Iriarte, M. (2017). *The reception of subtitling for the deaf and hard of hearing: viewer's hearing and communication profile and subtitling speed of exposure*. PhD dissertation. Universitat Autònoma de Barcelona.

Mitra, S. (2006). The capability approach and disability. *Journal of Disability Policy Studies*, 16(4), pp. 236-247.

Orero, P., Martín, C. A. and Zorrilla, M. (2015). HBB4ALL: Deployment of HbbTV services for all. *2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*. Ghent (Belgium).

Orfanidou, E., Woll, B. and Morgan, G. (2015). *Research methods in Sign Language studies: a practical guide*. Oxford: John Wiley & sons.

Pappworth, M.H. (1967). *Human guinea pigs: experimentation on man*. London: Routledge and Kegan Paul.

Romero-Fresco, P. and Fryer, L. (2013). Could audio described films benefit from audio introductions? A reception study with audio description users. *Journal of Visual Impairment and Blindness*, 107(4), pp. 287-295.

Romero-Fresco, P. (Ed.) (2015). *The reception of subtitles for the deaf and hard of hearing in Europe*. Bern: Peter Lang.

Ross, L.F., Loup, A., Nelson, R.M., Botkin, J.R., Kost, R., Smith Jr., G.R., and Gehlert, S. (2010). Nine key functions for a human subjects protection program for community-engaged research: points to consider. *Journal of Empirical Research on Human Research Ethics*, 5(1), pp. 33-47.

Serrat-Manén, J. (2011). *La percepció que tenen les persones sordes signants de l'actualitat periodística (2005-2009)*. PhD dissertation. Universitat Autònoma de Barcelona.

Singleton, J., Martin, A., and Morgan, G. (2015). Ethics, deaf-friendly research, and good practice when studying Sign Language. In E. Orfanidou, B. Woll and G. Morgan (Eds.), *Research methods in sign language studies: a practical guide*. Oxford: John Wiley & sons, pp. 7-20.

Szarkowska, A. (2011). Text-to-speech audio description: towards wider availability of audio description. *The Journal of Specialised Translation*, 15, pp. 142-163.

Szarkowska, A. and Jankowska, A. (2012). Text-to-speech audio description of voice-over films. A case study of audio described *Volver* in Polish. In E. Perego (Ed.), *Emerging topics in translation: audio description*. Trieste: EUT, pp. 81-98.

Szarkowska, A. and Wasylczyk, P. (2014). Audiodeskrypcja autorska. *Przekladaniec*, 28, pp. 48-62.

Udo, P. and Fels, D. (2009). Suit the action to the word, the word to the action. An unconventional approach to describing Shakespeare's Hamlet. *Journal of Visual Impairment and Blindness*, 103(3), pp. 178-183.

Walczak, A. (2010). *Audio description for children. A case study of text-to-speech audio description of*

educational animation series Once Upon a Time… Life. MA Thesis. University of Warsaw.

Walczak, A. (2017). *Immersion in audio description. The impact of style and vocal delivery on users' experience*. PhD dissertation. UAB.

Walczak, A. and Fryer, L. (2017). Creative description. The impact of audio description style on presence in visually impaired audiences. *British Journal of Visual Impairment,* pp. 6-17.

Walczak, A. and Fryer, L. (2018). Vocal delivery of audio description by genre. *Perspectives. Studies in Translatology*, 26(1), pp. 69-83.

Walczak, A. and Rubaj, X. (2014). Audiodeskrypcja na lekcji historii, biologii i fizki w klasie uczniów z dysfunkcja wzroku. *Przekladaniec*, 28, pp. 63-79.

Zuckerman, M. (1973). Scales for sex experience for males and females. *Journal of Consulting and Clinical Psychology*, 41(1), pp. 27-29.

Digital Inclusion and the Elderly: The Case of Online Banking

L. Georgieva
School of Mathematical and Computer Sciences
Heriot Watt University
Edinburgh, EH14, UK
L.Georgieva@hw.ac.uk

Abstract

Digital inclusion is recognized as a significant issue in the UK and is affecting all aspects of the society, including access to jobs and education, community structure and services. Recent research by the BBC has found that 21% of Britain's population lack basic digital skills and as a result are not able to take advantage of the digital technology and benefit from the internet. Among this population group, the elderly users are over-represented. Worrying statistics indicate that out of the 5.9 million adults, living in the UK who have never used the Internet, 85% are over 65 years old. In this paper we study the challenges that are faced by the older generation in the digital world. We survey a group of elderly users and identify issues that prevent them from engaging with digital technology. We discuss the role of NLP for improving perception of usability and ensuring optimal experience and propose measures which would address bridging the divide.

Keywords: digital inclusion, digital divide, NLP

1. Introduction

This paper aims to identify the main factors that are causing limited adoption of digital services for elderly users, and to consider the role of NLP in defining the digital divide in modern society. We consider the case of use of online banking by people aged 65 and over in the UK.

There are many factors that cause usability issues of technology for the elderly, including ease of access, compatibility with reduced range of motor skills, limited technical skills, resistance to willingness to change, and reliance on and overly-emphasizing of prior knowledge (Czaja, 2006, Czaja, 2007, Kleinberger, 2007). A way to improve customer engagement and retention is to anticipate client concerns and improve communication by effectively addressing these concerns (Carroll, 1997). Banks and other financial institutions can use NLP to discover and parse customer sentiment, for example, by monitoring social media and analysing conversations about their services and policies. A significant proportion of senior citizens are, however, excluded in this recommendation as they are underrepresented in users of social media and online platforms (Morrell, 2002).

In this paper we discuss fear as a factor and the potential role of NLP for transforming the fear into engagement and inclusion.

The elderly represent significant proportion of the UK population. Their number of people in the UK is estimated to be 65.4 million in 2018. People, aged 65 and over amount to 11.8 million (Office of National Statistics, UK, 2018). A dominant trend is that population is continuing to age; life expectancy has improved considerably in the past 25 years and so has health care resulting in a significant number of elderly and ageing users of technology. Current trends indicate that the population is said to increase by 5.76 million by 2035 with the percentage of people over 65, rising to 24% of the total population over the same period. In the absence of enough data relevant to the use of social media by the elderly, NLP provides an alternative channel for assessing their needs. We discuss the emergence of chatbots as a communication tool for this population group and recommend design features that would promote engagement.

The structure of this paper is as follows. In Section 2 we define the concept of digital divide and the respective categories of users of technology, affected by the divide. Factors, affecting elderly users of technology are discussed in Section 3. Steps for overcoming the divide are given in Section 4. Section 5 discusses the role of NLP.

2. Digital divide

Digital divide is defined as the gap between individuals, households, businesses and geographic areas at different socio-economic levels with regard both to their opportunities to access information and communication technologies (ICTs) and use of the internet for a wide variety of activities (Ragnedda and Muschert, 2013). Investigation into digital divide in Europe by Brandtzæg, et. al (Brandtzæg, Heim, and Karahasanović, 2011)) studied habits, attitudes and social status of users from the most technologically advanced European countries, including Norway, Sweden, Austria, UK, and Spain. Different categories of users, based on frequency of use of technology: non users, sporadic users, instrumental users, entertainment users, and advanced users were identified. A significant percentage of the population of the advanced European countries (approximately 60%) are identified by surveys or self-identify to be either non-users or sporadic users. This indicates that advantages of using digital technologies are not accessible by a significant percentage of the population (Karahasanović, et al., 2009). These advantages include, for example, access to free and high

quality education, offered exclusively online, which would encourage acquisition of new skills and allow for development of hobbies, consequently fighting cognitive decline and providing a platform for innovation, access to digital services, essential to daily lives, such as digital payments to local authorities, online shopping, and information lookup services.

The characteristics of the technology determine whether or not it will be adopted by targeted user groups. The following user groups were identified in (Yousafzai & Yani-de-Soriano, 2012).

Laggards are defined as users who use the internet for private causes and do not use any of the e-government services. Globally, people living in France, Germany, Ireland, and the UK are classified as laggards, according to this research.

Confused and adverse. These users fit best the user group that we study. They show high variability of habits and have low usage of internet. The highest numbers of confused and adverse users have been identified in Austria and UK.

Advanced users. They show a frequent and continuous use of the Internet and use the Internet for e-commerce and for administrative tasks. Advanced users are proportionally represented best in UK, Holland and Nordic countries.

The followers. The followers tend to use the internet frequently but not on a daily basis. They also use e-government services, but do not engage in e-commerce activities as frequently, when compare with the advanced users. EU countries with comparatively high percentage of the population, identified as , are Denmark and Holland.

Non-users. 44% or the **largest** group in the research were identified or self-identified as non-users of the internet and related technologies. Predominantly, this included high percentage of population form the Southern part of Europe, geographically represented in Spain, Greece, Portugal, and Italy.

3. Influential Factors

There are many factors that case the digital divide, that are pervasive across the user groups, including

Economic factors:

Income and personal wealth are factors influencing the digital divide (Vicente and Lopez, 2011). Personal income is positively correlated to persistent digital technology infiltration rates, independent of age.

Research by Helsper and Reisdorf, (2016) compared the use of digital technologies for the uptake of digital services (including online banking) in Great Britain and Sweden. *Low income* and *unemployment* were identified as significant socio-economic characteristics in UK. In Sweden, a third factor, namely, *family status* was also identified: being single is a barrier to the uptake of digital technologies. This factor, even though of less significance in the UK is particularly relevant for older people. The increasing number of elderly people, living alone, places them at risk of being excluded from access wide range of digital technologies and services and their respective benefits. As *non-users*, they do not see the benefits in using digital services. As people get older and retire, their earning power commonly diminishes and this causes additional issue. They perceive the computer or portable electronic devices, such as laptops, notepads or notebooks as expensive items which they cannot neither afford, not have the knowledge how to use or maintain.

Demographic factors: Age and gender

Keeping up with advancements of digital technologies is an ongoing and persistent issue for elderly people who are not IT savvy. Financial services have continuously become prevailing. With significant number of local banks closing down, due to cost cutting incentives, online banking becomes often the only option for the daily needs of an ageing population.

When considering exclusively age as a factor, researchers also focus on health-related issues, which include eyesight deterioration and coordination and motor skill issues (Vicente and Lopez, 2011), (Helsper and Reisdorf, 2011). Older users are less familiar with the technology, and their ability to adopt new technology depends on their willingness, computer self-efficacy, and dependence on prior knowledge. They have lower confidence in their own cognitive capabilities, often acting as a self-fulfilling prophecy when adopting new technologies (Yousafzai and Soriano, 2012).

Urbanization:

Internet access is cheaper and faster in urban areas in comparison to the countryside. Higher numbers of skilled and knowledgeable workers are available for support in case of technical issues (Vicente and Lopez, 2011). We identified that, while not of highest significance for the uptake of new technology, support with technical issues is a decisive factor for the elderly users. *Education and language barriers:*

Vicente and Lopez, 2011 link language barriers to technology uptake. In Scotland, part of the older generation use Gaelic which has limited language support online. Breadth of usage, self-efficacy, experience and education are also factors (Helsper & Eynon, 2010).

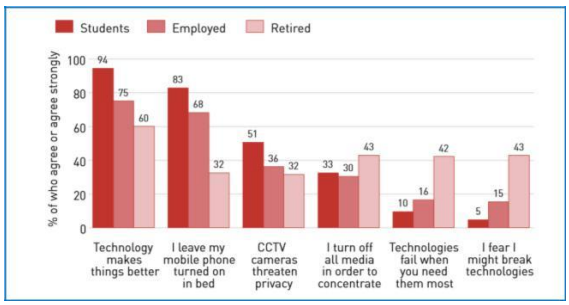


Figure 1 Age- dependent attitude to technology.

4. Overcoming the divide

Online banking offers advantages for users who live in the remote places, for example in areas, predominately populated by elderly people (in the UK one such example is the Scottish islands in the North). Most of the banks are now committed to promoting online service over the traditional methods as it beneficial. Banks, which operate exclusively online exist (see for example smile.co.uk or Monzo) and are expected to become more numerous.

The concept of digital divide is also closely linked to the generation gap, and is influenced by factors such as education, income, social mobility (IBM, 2016).

The Oxford internet survey identifies two types of users. The *next generation*, who use technology every day of their lives, and the *first generation* of users, which comprises of people who are unable to use the technology and internet services, and consists of a large proportion of retired and unemployed people. Figure 2 is adapted from the survey, and categorises 71% of retired people as *first generation* and only 29% as *next generation*.

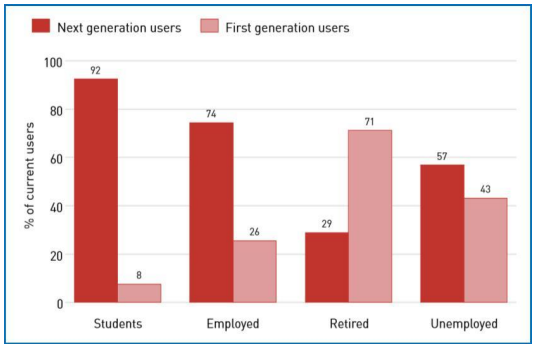


Figure 2 First versus next generation of users

We aim to determine the possible reasons for rejection and disapproval of online banking among the elderly users, focusing on factors that can be influenced by NLP. We next consider are fear and usability.

Fear of technology is excessive and disproportionate to the situation (Graham and Bronwyn, 2011). The Oxford Internet survey identified that 46% of the elderly participants strongly agreed they don't use the technology as they are scared they might 'break something' and 68% suggested it was easier to cope without using the technology.

We surveyed only people aged 65 and over, both genders were equally represented. We compared the users' attitudes towards traditional banking, which is identified as the most popular among senior citizens (Yousafzai & Yani-de-Soriano, 2012) and online banking.

88% of our respondents identified themselves as users of exclusively in-person banking. Telephone banking was not preferred option for any of our respondents. Online banking was taken up by 12%. The reasons for using in-person banking by elderly users include the need for human interaction, the need to be sociable and the need to engage in an in-person communication to preserve cognitive functions. Our respondents identified the need to be sociable and to exchange as the most significant factor.

None of our respondents identified themselves as users of telephone banking, indicating that the usefulness of this channel is potentially decreasing. Elderly users of online banking ranked highly also *ease of use* and *safety* as dominant determining factors.

Advanced age, lower income, education and having no family support when using online banking have been identified as barrier factors for other user groups (Vicente and Lopez, 2011) , (Yousafzai & Yani-de-Soriano, 2012).

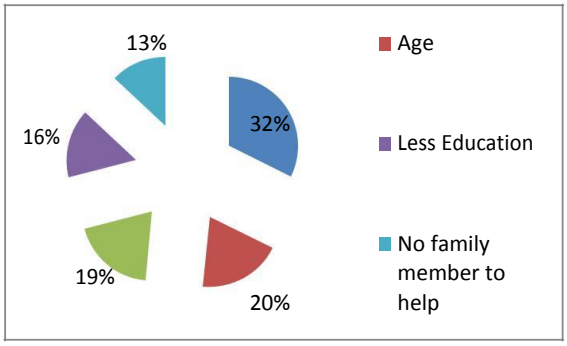


Figure 3. Factors affecting usability

Factors, affecting usability as identified by our survey respondents are given in Figure 3. Predominantly, less experience with use of online banking was the most significant factor, with 32% of the respondents indicating it as the most significant reason for avoidance. *« No family member to help «* was given as a reason for avoidance by only 13% of the respondents, indicating that the need for independence when using online banking is strong and few of the people aged 65 and over relies on a younger relative to assist them.

Not having the right experience is identified as the problem: 32% of the people identified themselves as not having the correct experience to carry out the online banking. This can be resolved by providing elderly users with access to computers, for example in community centres or local libraries and with the suitable training. 17% of the participants identified online banking as 'confusing and complicated' and 20% said age is factor for avoidance. The 'customer relations lost' were seen as a important factor among the elderly users, which resulted in them not favouring online banking.

Our survey participants are all customers of main banks in Scotland. They identified experience; trust in online banking, and banking in general, usability of website, reliability of online banking and security issues to be dominant in their reluctance to join the growing number of online banking customers (see Figure 5, which details the most significant reasons for use of traditional banking). When customers can learn to trust the online

banking systems and will start to use it. Many elderly users feared the risk of losing their physical bank to an online bank, and as a result depriving them of valued human interaction. Factors, positively affecting usability identified by our survey respondents included cheaper broadband, training, easy accessibility, easy user interface and higher aware ness. A related question asked what the banks need to do in order to improve the online banking. Online security and trust was identified as essential, user friendly website in order to do their transactions and online banking.

A cheaper broadband and computer facilities and training were identified as secondary, but significant factors.

5. The role of NLP

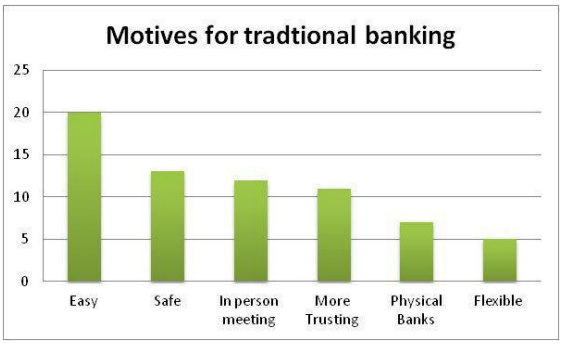
Elderly users identified the need for social interaction and assurance as decisive when using traditional banking. They indicated that online banking and digital media does not give them the same level of reassurance. One possibility to address this to providing customer reassurance by chatbots, which are become a ubiquitous component of the customer service. With the demand for phone banking declining, digital interactions is expected to be the preferred communication. Chatbots will enhance chat conversations by helping humans with micro-tasks and automatic replies, though it's unlikely that bots will replace humans entirely. Chatbots have been around since the 1966 and a currently used by, for example, Scottish Gas or online services such as Dell. Chatbots provide hybrid experience between human customer service agents and bots and provide a simulation to a human interaction.

Simple tasks, such as information gathering, like asking a customer for their name and account number can be carried out by chatbots, but need to be adjusted with better natural language processing to address the specific needs of elderly customers. Using machine learning techniques, the chatbots are expected smarter over time (IBM, 2016). The unique needs of elderly users, including better voice recognition, clearer instructions, and better responsiveness will also need to be addressed.

Small screens are another motor skill-related issue for the elderly. Reminders, sent by text messages tend to not be as effective for this reason.

Initiatives to design chatbots, specifically for the elderly exist (Fomitchev, 2017). In online banking services, a chatbot can assist, for example with planning of payments, reminder of outstanding transaction. Chatbots search for relevant and information and presented in a suitable way, providing assistance when needed and can store communication history, which can serve as evidence of communication and transaction records. Routine communication with a chatbot and simple user interface can address if not all, at least some of the fear factors and be a simulation of a positive interaction.

Figure 5 Motives for using traditional banking



1. Bibliography

Brandtzæg, P. B., Heim, J. and Karahasanović, A. (2011) *Understanding the new digital divide*. A typology of Internet users in Europe, International Journal of Human-Computer Studies, Volume 69(3)

Carroll, J. M. (1997) *Human-computer interaction: psychology as a science of design.* International Journal of Human-Computer Studies Volume 46(4), p. 501–522.

Castor, A. and Pollux, L. E. (1992). *The use of user modelling to guide inference and learning*. Applied Intelligence, 2(1):37–53.

Cultures of the Internet: The Internet in Britain. Oxford Internet Survey 2013. Internet Institute, University of Oxford.

Czaja, S. J., Charness, N., Fisk, A.D., Hertzog, C., Nair, S.N., Rogers, W.A., & Sharit, J. (2006). *Factors predicting the use of technology: Findings from the Center for Research and Education on Aging and Technology Enhancement.* Psychology and Aging, 21, p. 333-352.

Czaja, S. J., & Lee, C. C. (2007). *The impact of aging on access to technology.* Universal Access in the Information Society, Volume 5, p. 341-349.

Formichev, G. (2017) *ChatBots for Senior People and Patients with Alzheimer's Disease.*

Graham, M. and Bronwyn, B. (2011) *The Study of Fear Extinction: Implications for Anxiety Disorders.* The American Journal of Psychatry.

Helsper, E. J. and Reisdorf, B. C. (2016*) The emergence of a "digital underclass" in Great Britain and Sweden: changing reasons for digital exclusion.* New Media & Society.

IBM, Banking Technology Division, *The future of banking is here: cognitive banking*, 2016.

Karahasanović, A., Brandtzæg, P. B., Heim, J., Lüders, M., Vermeir, L., Pierson, J. , Lievens, B. , Vanattenhoven, J. and Jans, G. (2009). *Co-creation and user-generated content-elderly people's user requirements*. Computational Human Behaviour, 25(3) p. 655-678.

Kleinberger, T., Becker, M., Ras, E., Holzinger, A., Müller, P. (2007) *Ambient Intelligence in Assisted Living: Enable Elderly People to Handle Future Interfaces*. UAHCI, LNCS, Volume 4555, p. 103–112. Springer.

Morrell, R. W. (Ed.). (2002). *Older Adults, Health Information and the World Wide Web*. Mahwah, New Jersey: Lawrence Erlbaum Associates. Office for National Statistics, Internet Users, 2015.

Ragnedda, M. and Muschert, G. (2013) *The digital divide.* The Internet and Social Inequality in International Perspective.

Vicente, M. and López-Menéndez, A. (2011). Assessing the regional digital divide across the European Union. Telecommunications Policy, Vol. 35(3), p. 220-237.

Schmid, B. F. (2001). What is new about the digital economy? Electronic Markets, 11 (1), 44–51.

Yousafzai, S. and Yani-de-Soriano, M. (2012). "Understanding customer-specific factors underpinning internet banking adoption", International Journal of Bank Marketing , Vol. 30(1), p. 60-81.

How to Make Troubleshooting Simpler? Assessing Differences in Perceived Sentence Simplicity by Native and Non-native Speakers

Sanja Štajner
Data and Web Science Group
University of Mannheim, Germany
sanja@informatik.uni-mannheim.de

Abstract

Text Simplification (TS) aims to make texts more accessible to various readers by reducing reading time and/or by facilitating understanding and access to the relevant information. In this study, we address two topics which are important for TS but have not received much attention so far. First, we explore which types of syntactic simplification transformations make troubleshooting in IT domain easier for both native and non-native English speakers. Second, we explore how the choice of TS evaluation strategy influences final results. Our experiments show that grammaticality of a sentence influences the perceived simplicity by native and by non-native speakers differently. We also find that a high inter-annotator agreement can be achieved only in the case of the relative assessment of the sentence pairs in which one sentence is significantly simpler than the other one.

Keywords: evaluation, text simplification, sentence simplicity

1. Introduction

Text simplification (TS) has the goal of making texts more accessible by reducing reading time and/or improving understanding of the information contained in them. So far, TS systems have been proposed for many languages, e.g. English (Carroll et al., 1999; Coster and Kauchak, 2011a; Siddharthan and Angrosh, 2014; Štajner and Glavaš, 2017; Nisioi et al., 2017), Spanish (Saggion et al., 2015; Štajner et al., 2015b), Portuguese (Aluísio and Gasperin, 2010; Specia, 2010), Italian (Barlacchi and Tonelli, 2013), Basque (Aranzabe et al., 2012). The proposed TS systems had various target populations in mind, e.g. children (Barlacchi and Tonelli, 2013), people with low literacy levels (Aluísio and Gasperin, 2010), non-native speakers (Paetzold and Specia, 2016b), and people with various cognitive or reading impairments (Saggion et al., 2015; Rello et al., 2013; Orasan et al., 2013). The majority of the proposed TS systems focused on simplifying either news articles, or Wikipedia articles, or both. The latest state-of-the-art TS systems (Nisioi et al., 2017; Zhang and Lapata, 2017) use neural architectures and are trained on the English Wikipedia – Simple English Wikipedia (EW–SEW) TS datasets (Coster and Kauchak, 2011b; Hwang et al., 2015). By being fully supervised and trained on the EW–SEW TS dataset, they represent the 'general' TS systems which should make texts more accessible to everyone.

1.1. Evaluation of TS systems

In spite of the recent increased interest in text simplification, there are no common standards in evaluation of TS systems. Ideally, TS systems should be evaluated at the text level, by measuring reading time and understanding by the final users. However, such an evaluation is time-consuming, and in the case of vulnerable target populations, the access to the final users might be difficult. Therefore, in practice, TS systems are usually evaluated at the sentence level by native or non-native speakers, or a mixture of both. This already raises some important issues, as randomly-selected or crowdsourced evaluators might not be good proxies for the target populations. Apart from that, such human evaluations have varied from one study to another with regards to the number of annotators, the type of annotators (native vs. non-native), the type of evaluation (absolute score vs. relative comparison), the evaluation scale (0/1, 1–3, or 1–5), etc. (see Table 1 in Section 2), which brings additional problems in comparing the performances of TS systems from different studies.

1.2. Goals and Contributions

This work has two main goals. The first is to better understand how different syntactic transformations and the grammaticality of a sentence influence the perceived sentence simplicity by native and non-native English speakers. The second goal is to explore how the type of annotators and the type of evaluation influence final results and the inter-annotator agreement (IAA). To achieve those goals, we explore the following research questions:

- **RQ1**: How does the grammaticality of a text snippet influence its simplicity and whether this influence vary depending on the type of evaluators (native vs. non-native speakers) or not?
- **RQ2**: Is the absolute simplicity of a text snippet perceived differently by native and by non-native speakers?
- **RQ3**: Is there a difference in simplification gain after applying particular simplification operations on a text snippet (i.e. its relative simplicity) depending on whether it is evaluated by a native or by a non-native speakers?
- **RQ4**: How does the type of evaluators (native vs. non-native speakers) and the evaluation type (absolute vs. relative simplicity) reflect on the inter-annotator agreement?

Study – Language	Simp.type		Readab.	MTeval	Cover.	Human evaluation of sentence/word simplicity							
	Synt.	Lex.				Native	1–5	1–3	0–1	Rel.	#annot.	#sent.	mod.
(Specia, 2010) – PT	+	+	−	+	−	?	−	+	−	−	?	20	+
(Yatskar et al., 2010) – EN	−	+	−	−	+	+	−	+	−	−	3	100*	−
(Coster and Kauchak, 2011a) – EN	+	+	−	+	−	−	−	−	−	−	−	−	−
(Wubben et al., 2012) – EN	+	+	+	+	−	−	+	−	−	−	46	20	+
(Glavaš and Štajner, 2013) – EN	+	−	+	−	−	−	−	+	−	−	3	70	?
(Angrosh et al., 2014) – EN	+	+	−	−	−	±?	+	−	−	−	?	50	+
(Saggion et al., 2015) – ES	+	+	+	−	−	+	+	+	−	−	25	48	+
(Baeza-Yates et al., 2015) – ES	−	+	−	−	+	+	−	−	+	−	3	200*	+
(Štajner et al., 2015b) – ES	+	+	−	+	−	±	+	−	−	+	13	20-40	+
(Glavaš and Štajner, 2015) – EN	−	+	−	−	+	−	+	−	−	+	2	80	−
(Paetzold and Specia, 2016b) – EN	−	+	−	−	+	−	−	−	−	−	−	−	−
(Xu et al., 2016) – EN	+	+	−	+	−	?	−	−	−	+	5	?	?

Table 1: The types of evaluations used in various text simplification studies ('*' signifies the number of examples/words instead of the number of sentences; '?' that the answer cannot be found in the paper; and '±?' that the evaluators are most probably a mixture of native and non-native speakers).

To better place our work into the current TS literature and clarify our research goals and contributions, we present the most relevant related work in Section 2. In Section 3, we describe the procedure for collecting the new dataset used in this study. The next four sections (Sections 4–7) describe the experimental setup and provide results and detailed discussions for each of the four research questions (RQ1–4) separately. Section 8 summarises the most important results and discuss their potential impact on TS research.

2. Related Work

In this section, we present the most relevant related work with regards to the automatic evaluation methods in TS (Section 2.1), human evaluation procedures in TS (Section 2.2), inter-annotator agreement (Section 2.3), and the existing evaluation datasets (Section 2.4).

2.1. Automatic Evaluation in TS

The main way of evaluating text simplification (TS) systems, either manual or automated, is by human evaluation. Some studies, however, additionally include an automatic evaluation of TS systems, either by using readability formulae (e.g. (Saggion et al., 2015; Drndarević et al., 2013; Woodsend and Lapata, 2011)), or in the case of machine translation (MT) based TS (e.g. (Specia, 2010; Xu et al., 2016)), by using some of the automatic machine translation evaluation measures such as BLEU (Papineni et al., 2002) or NIST score (Doddington, 2002). Although such automatic evaluation measures allow for a greater number of instances to be evaluated than by means of human evaluation, they still have a number of shortcomings. Readability metrics are reliable only at the text level and not at the sentence level (and are oblivalent to grammaticality and meaning), while machine translation evaluation metrics, although showing good correlation with some human judgements (Wubben et al., 2012; Štajner et al., 2014; Xu et al., 2016; Štajner et al., 2016), do not reward for TS-specific transformations such as sentence shortening or strong paraphrasing, and do not take into account the input sentences (Štajner et al., 2015a; Xu et al., 2016; Štajner and Nisioi, 2018). Additionally, the MT-evaluation metrics require 'gold standard' simplifications which are rarely available in TS.

The systems which only perform lexical simplification (no syntactic simplification) are usually evaluated automatically by the number of changes performed and the coverage of changes over a 'gold standard' dataset of complex words for non-native speakers (Horn et al., 2014; Glavaš and Štajner, 2015; Paetzold and Specia, 2016a; Paetzold and Specia, 2016b). This type of evaluation is thus limited to those particular test instances.

2.2. Human Evaluation in TS

The main and most reliable method for evaluating TS systems is still considered to be human evaluation. Although it should ideally be performed at the text level and by the final target users, given the time such evaluation would require and difficulties to reach some of the target populations, human evaluation of TS systems is usually done only at the sentence level, by either native or non-native speakers (or, less often, a mixture of both). The annotators are asked to rate the simplified sentences for their grammaticality, meaning preservation, and simplicity, either on an absolute scale or in a pairwise comparison. The problem is that different TS studies use different evaluation strategies and thus it is not possible to compare their results. Table 1 presents an overview of different types of evaluation strategies used for assessing the simplicity of TS output so far. As can be seen, human evaluation is sometimes performed in terms of an absolute simplicity score on a 1–5 or 1–3 level scale, sometimes as a 0–1 labeling (simpler/not simpler than the original), and sometimes as a relative pairwise comparison of the output of different systems or system configurations (*Rel.*). The number of annotators (*#annot.*), and the number of sentences (*#sent.*) also vary, as well as whether the evaluation is performed only on the sentences which underwent some change (*mod.*) or a random subset of sentences (including the unchanged ones).

2.3. Inter-Annotator Agreement

Although Yatskar et al. (2010) reported a better inter-annotator agreement among native than among non-native evaluators on a 1–5 level evaluation task, no other stud-

Answer type	Text
Original (O)	Your payment received not and application submitted
Grammatical (G)	Your payment **was not received** and **the** application **was** submitted.
Split (S)	Your payment was not received**. However,** the application was submitted.
All (A)	**We did not receive** your payment. **But, we received your** application.

Table 2: An example of the original answer (O), grammatically corrected answer (G), answer with split sentences (S), and the answer with applied *additional operations* (A). Differences between the answers are shown in bold. The corresponding question was: *Was my payment received on time?*

ies tried to further explore the influence of the annotators type on the obtained evaluation results. The differences in the obtained results could be particularly pronounced in the case of relative comparisons, as the original sentence might already be simple enough for native speakers and therefore, its further simplification would not be rewarded by native annotators as much as by non-native annotators. Another problem could be that the original sentence might be so complex for the non-native annotators that any of the performed simplification transformations (usually very few in automatic text simplification systems) could not make it any simpler.

Furthermore, it has never been explored how much the grammaticality of a sentence influences its perceived simplicity. This might be an important factor, as the current state-of-the-art TS systems still produce many ungrammatical sentences. An output sentence which is rated as complex might be rated that way for various reasons: (1) because it is ungrammatical (in spite of correctly applied lexical or syntactic simplification operations); (2) because it was left unchanged and the original sentence was already complex; (3) because the simplification operations were incorrectly applied and led to generating a more complex sentence (which might be completely grammatical). Those three cases could be differently rated by native and by non-native annotators, as for example, the native evaluators might penalize the ungrammaticality more severely, while the non-native annotators might reward lexical and syntactic simplicity regardless of the grammaticality.

2.4. Existing Datasets for TS Evaluation

The existing datasets with human evaluation scores (e.g. those systems shown in Table 1) are not convenient for assessing the influence of grammaticality and various sentence simplification operations on the perceived simplicity of a text snippet, as they do not control for one variable at the time. The majority of those datasets only contain those sentences which underwent at least one syntactic or lexical transformation, and such sentences are often ungrammatical. By using such a dataset we would not be able to know whether the change in simplicity score between the original and automatically simplified (ungrammatical) sentence comes from the fact that the grammaticality of the sentence was damaged, or from the changes that were made (and which kind of changes were made exactly), while all those factors may play a role. Additionally, only one of those datasets (Štajner et al., 2015b) contains the evaluation scores assigned by both native and non-native speakers and would thus allow for comparison of scores obtained by native vs. non-native evaluators. Yet again, that dataset does not have a sufficient number of sentences to allow controlling for the grammaticality and the sentence transformation type. To control for all those factors at the same time, we create a new dataset described in the next section.

3. Data Collection

We collect questions and answers (Q&A) from an IT service in a hospital in India and from the WMT 2016 shared task for the IT domain.[1] We opt for having a Q&A type of dataset to have a larger context for better simplicity assessment, and for testing the previously mentioned issues in a real-world scenario, in which a reader is required to understand the answer and find the necessary information. Out of all Q&A from those two sources, we selected 30 (20 from the first source and 10 from the second) which fulfilled the following two conditions: (1) the answer was grammatically incorrect (containing more than one grammatical error) and (2) the answer was sufficiently complex to allow applying sentence splitting and at least one of the following simplification operations (*additional transformations*):

- removing superfluous words
- conversion of passive to active voice
- disambiguation of meaning
- conversion to the canonical subject-verb-object form

As *additional transformations* we choose the most frequently used operations in various guidelines for producing easy-to-read texts (PlainLanguage, 2011; Mencap, 2002; Freyhoff et al., 1998) and in rule-based text simplification modules (Siddharthan and Angrosh, 2014; Saggion et al., 2015). We do not focus on lexical simplification on purpose, as LS systems are usually evaluated for their coverage on the existing, specially designed datasets (Horn et al., 2014; Glavaš and Štajner, 2015; Paetzold and Specia, 2016a; Paetzold and Specia, 2016b). Instead, we focus on the simplification operations which operate on a syntactic or discourse level.

For each of the 30 answers, in addition to the original answer (O), we produce three simplified versions: (1) the grammatically corrected version (G); (2) the sentence-split version (S) by performing sentence splitting on the grammatically corrected version as many times as possible to satisfy the main rule of easy-to-read texts (keeping the sentences as short as possible and covering only one main idea per sentence); and (3) the (grammatically corrected) sentence simplified by using sentence splitting and as many as

[1] http://www.statmt.org/wmt16/it-translation-task.html

possible *additional transformations* (A). All three versions were produced by a linguist, native English speaker, and checked by another. Both editors were experienced in manual text simplification. An example of the original sentence (O), its manually corrected version (G), and the two manually simplified versions: after the sentence splitting (S) and after performing additional simplification transformations (A) is presented in Table 2.

The 30 Q&A were divided into two sets of 15 Q&A randomly. Next, we ask 40 native and 40 non-native English speakers to rate each of the four answers for the given 15 questions on a 1–5 Likert scale by saying how easy to understand was the answer ($1 \rightarrow$ very simple; $5 \rightarrow$ very complex). This way, each answer is evaluated by 20 native and 20 non-native English speakers, and each evaluator only evaluates 15 answers. In this way, we follow the common TS evaluation practices of not giving more than 20 tasks/items to the evaluators in the case of crowdsourced evaluation, to avoid the fatigue effect.

Different versions of the same answers were always shown one after another in random order, always together with their corresponding question. We opted for this way of presenting answers, which allows for their direct comparison and ranking, but we did not instruct the annotators to rank them. Instead, we asked them to evaluate them independently of each other. We chose this evaluation setup as it is the most common way of evaluating TS systems.

The native English speakers were from the UK, USA, and Australia. The non-native English speakers were from India and Germany. The English proficiency of non-native speakers was not checked by any kind of qualification tests. However, the language used at their workplace is English. All 80 evaluators (40 native and 40 non-native English speakers) are familiar with the IT support procedures, as they all use computers in their daily tasks at work and have contacted the IT support at least once before. We did not collect any additional information about the evaluators (e.g. gender, age, or name).

4. Influence of Grammaticality (RQ1)

To explore how the grammaticality of an answer influences its perceived simplicity, we conduct two analyses.

First, we calculate the difference between the simplicity score assigned to the grammatically corrected answer and the one assigned to the original answer (where positive number indicates higher simplicity of the grammatically corrected version) for each answer and for each annotator (a total of $30 \times 20 = 600$ data points). The first row in Table 3 presents the average difference with standard deviation. We find that grammaticality has a significantly (Mann-Whitney U test in SPSS; $p < 0.05$) stronger influence on perceived simplicity within the native speakers than within the non-native speakers.

Second, we calculate the percentage of cases in which the grammatically correct version (G) was rated as simpler than the original (ungrammatical) version (O), and the percentage of cases in which both versions (O and G) were rated as equally simple (the last two rows in Table 3).

The results indicate that the native speakers penalize the simplicity score of ungrammatical sentences significantly more than the non-native speakers do.

Measure	Native	Non-native
Average difference $\pm$ st.dev.	0.50 ± 1.07	0.39 ± 1.20
G simpler than O	44.50%	37.50%
G equally simple as O	43.60%	45.83%

Table 3: The average difference (with standard deviation) between the simplicity score of the grammatically correct answer (G) and the simplicity score of the original answer (O), where positive values signify that the original answer was perceived as more complex, and the percentage of cases (out of 600) in which G was rated as simpler than O, or as equally simple as O.

Type	Native	Non-native
Original	2.65 ± 1.12	2.66 ± 1.22
Grammatically correct	2.14 ± 0.99	2.27 ± 1.09
With splitting	$\mathbf{1.95 \pm 0.91}$	$\mathbf{2.15 \pm 1.10}$
With addition. transform.	$\mathbf{1.82 \pm 0.91}$	$\mathbf{2.14 \pm 1.20}$

Table 4: The mean value of the simplicity score (with standard deviation) for different variants of the answers (the lower the score, the simpler the answer). Statistically significant differences (Mann-Whitney U test in SPSS; $p < 0.02$ and $p < 0.001$, respectively) between the two groups of evaluators (native and non-native) are presented in bold.

5. Absolute Simplicity (RQ2)

The influence of the annotators group (native vs. non-native) on the perceived absolute simplicity of the answer (before and after various modifications) is presented in Table 4. Although both groups start from the similar average simplicity scores for the original answers (2.65 and 2.66, respectively), after the application of simplification transformations (both sentence splitting and additional transformations), the native speakers perceive the simplified answers significantly simpler than the non-native speakers do. Additionally, we notice a lower standard deviation in the scores within the native speakers. This indicates a greater homogeneity within the native than within the non-native annotators, which is in line with the findings of Yatskar et al. (2010) and Yimam et al. (2017) on similar tasks.

6. Relative Simplicity (RQ3)

To investigate the influence of the annotators group (native vs. non-native) on the results of the relative simplicity assessment, for each of the 600 data points (30 Q&A $\times$ 20 annotators), we calculate the difference between the scores obtained for two different versions of the same answer. The mean values and standard deviations are presented in Table 5, where positive score indicates that the second version is simpler. To investigate whether there are significant differences in simplicity scores assigned to the answer before and after certain simplification/correction operation, within the same group of annotators, we compare the scores using marginal homogeneity test for two related samples in SPSS, following the methodology for the relative simplicity assessment used by Štajner et al. (2015b).

Comparing	Native	Non-native
Original−Grammatical	***0.50 ± 1.07**	***0.39 ± 1.20**
Grammatical−Split	*0.19 ± 1.01	*0.12 ± 1.07
Grammatical−Additional	***0.32 ± 1.25**	***0.13 ± 1.17**
Original−Split	***0.70 ± 1.20**	***0.51 ± 1.31**
Original−Additional	***0.82 ± 1.32**	***0.52 ± 1.37**
Split−Additional	*0.13 ± 1.12	0.01 ± 1.10

Table 5: The average change (with standard deviation) in simplicity score between the two versions of the same answer. Statistically significant (Mann-Whitney U test in SPSS; $p < 0.05$) differences between the two groups of evaluators are presented in bold. Statistically significant (marginal homogeneity test for two repeated samples in SPSS; $p < 0.01$) differences in simplicity scores between the two versions of the same answer, within the same group of annotators, are marked with an '*'.

For almost all relative comparisons, we find a significant difference in the obtained simplicity gain between the two groups of annotators. The differences are mostly influenced by the way the grammaticality and the application of *additional transformations* change the score. Sentence splitting seems to influence simplicity gain similarly in both groups of annotators. The difference in simplicity gain after applying sentence splitting and multiple transformations on the grammatically corrected versions (*Grammatical−Additional*) is almost three times more pronounced within the native than the non-native annotators.

The results also indicate that the grammaticality of a sentence has much stronger influence on the simplicity gain than any simplification transformation. This calls for attention in current practices in text simplification evaluation. As we saw earlier (Table 1), some studies evaluate only the sentences which have been changed, while the others also evaluate the unchanged sentences. Transformed sentences are often ungrammatical, and according to our results, this can influence the perceived simplicity more than any simplification operation, thus blurring the TS evaluation results. Furthermore, the application of multiple transformations on the already short sentences (*Split−Additional*) seems not to have much influence, on average, in any of the two annotator groups.

The large standard deviations within the groups indicate a high heterogeneity in perceived simplicity gain within both groups, indicating the need for having a large number of annotators for a reliable evaluation of simplicity gain (relative comparisons).

7. Inter-Annotator Agreement (RQ4)

In this set of experiments, we calculate the unweighted Cohen's kappa (κ) as a measure of inter-annotator agreement among each pair of annotators within the group of native annotators, and within the group of non-native annotators (a total of 190 pairs in each group).[2] At the same time, we control for the type of evaluation (absolute vs. relative), and for the sentence type (or, in the case of the relative simplicity evaluation, for the sentence pair type).

[2]Although the IAA can be calculated for multiple annotators, in TS evaluation, it is a common practice to report the average pairwise IAA calculated as the (un)weighted Cohen's kappa.

κ range	Original	Gramm.	Split	Addit.
$0 < \kappa \leq 0.2$	139	168	158	152
$0.2 < \kappa \leq 0.4$	41	19	25	33
$0.4 < \kappa \leq 0.6$	7	2	5	4
$0.6 < \kappa \leq 0.8$	2	0	1	0
$0.8 < \kappa \leq 1.0$	1	1	1	1

Table 6: The number of pairs of native annotators with the IAA scores (Cohen's κ) in each of five score ranges, for each of the four types of sentences (or text snippets).

κ range	Original	Gramm.	Split	Addit.
$0 < \kappa \leq 0.2$	153	175	160	154
$0.2 < \kappa \leq 0.4$	34	13	22	28
$0.4 < \kappa \leq 0.6$	2	1	6	7
$0.6 < \kappa \leq 0.8$	1	1	2	1
$0.8 < \kappa \leq 1.0$	0	0	0	0

Table 7: The number of pairs of non-native annotators with the IAA scores (Cohen's κ) in each of the five score ranges, for each of the four types of sentences (or text snippets).

7.1. IAA in Absolute Evaluations

In the case of the absolute simplicity evaluation, each sentence was marked with a 1–5 score. For each group (native or non-native) of 190 annotator pairs, and for each of the five IAA score ranges, we count the number of annotator pairs with the IAA in that range (Tables 6 and 7).

The obtained results confirm the earlier findings that native speakers have better IAA than non-native speakers (Yatskar et al., 2010; Yimam et al., 2017). The most striking, however, is that in both annotator groups, over 73% of annotator pairs (regardless of the annotators group and the sentence type) have a very low IAA agreement ($0 < \kappa \leq 0.2$).

Both annotator groups show similar trends in how sentence type influences the IAA (see Figures 1 and 2). In both cases, the lowest IAA is achieved for the grammatical sentences and those that were split.

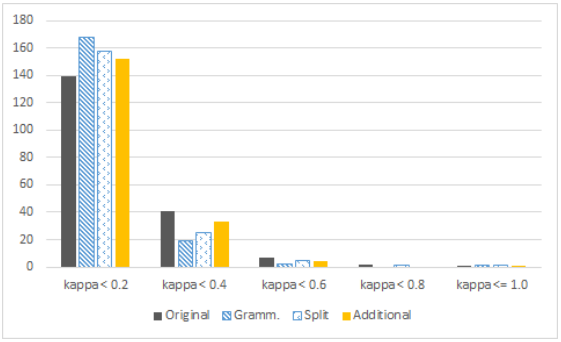


Figure 1: Distribution of the IAA scores depending on the sentence type for the native annotators.

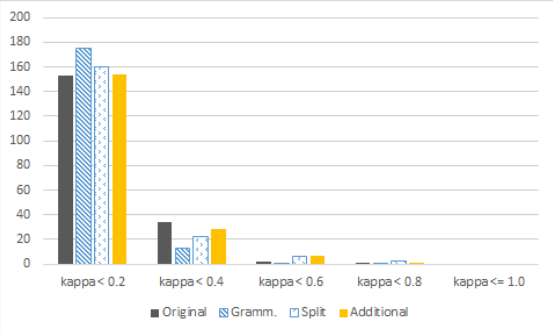


Figure 2: Distribution of the IAA scores depending on the sentence type for the non-native annotators.

κ range	O>G	O>S	O>A	G>S	G>A	S>A
$0 < \kappa \leq 0.2$	133	118	23	128	125	149
$0.2 < \kappa \leq 0.4$	40	49	0	40	49	21
$0.4 < \kappa \leq 0.6$	14	18	40	17	12	18
$0.6 < \kappa \leq 0.8$	1	3	70	4	1	0
$0.8 < \kappa \leq 1.0$	2	2	57	1	3	2

Table 8: The number of pairs of native annotators with the IAA scores (Cohen's κ) in each of the five score ranges, for each of the four types of sentences (or text snippets).

7.2. IAA in Relative Evaluations

In the case of the relative evaluation of simplicity, the annotators are asked to answer 'yes'/'no' to the question whether one sentence is simpler than the other. For each group (native or non-native) of 190 annotator pairs, and each of the five IAA score ranges, we count the number of annotator pairs with the IAA in that range (Tables 8 and 9). As expected, the IAA is better for the relative evaluations than for the absolute evaluations within both groups of annotators. We found significantly higher number of annotator pairs with the κ between 0.4 and 0.6 in the relative evaluations than in the absolute evaluations. The most striking is, however, that only the relative evaluations that compare the original sentences with the fully simplified sentences (O>A) achieve a very high number of good IAA scores, within both annotator groups (with better inter-annotator agreements within the native than within the non-native annotators).

These results call for caution in current TS evaluation methods, where human evaluations are either crowdsourced or where the IAA is not checked (those studies that use very few annotators usually have more experienced and well trained annotators, with a high IAA reported).

κ range	O>G	O>S	O>A	G>S	G>A	S>A
$0 < \kappa \leq 0.2$	120	142	59	141	152	148
$0.2 < \kappa \leq 0.4$	66	43	6	43	34	40
$0.4 < \kappa \leq 0.6$	4	5	40	5	4	2
$0.6 < \kappa \leq 0.8$	0	0	54	1	0	0
$0.8 < \kappa \leq 1.0$	0	0	31	0	0	0

Table 9: The number of pairs of non-native annotators with the IAA scores (Cohen's κ) in each of five score ranges, for each of the four types of sentences (or text snippets).

Additionally, these results show that we can have reliable relative comparisons of simplicity only in those cases where the differences between the two sentences are obvious (as in the case of the original sentence being compared with the fully simplified sentence).

8. Conclusions

In this study, we explored the differences in how native and non-native speakers perceive sentence simplicity, aiming for better understanding of their simplification needs and for better understanding how the choice of evaluators (native or non-native speakers) can influence the results of simplicity assessment. To do so, we built a new dataset with human evaluation of simplicity of text snippets, carefully controlling for various factors that could influence the perceived simplicity.

We found that native and non-native annotators differently reward both grammaticality and various simplification operations in their simplicity scores, and that grammaticality influences simplicity score more than any other (non-lexical) simplification transformation. These results imply that we should be cautions when mixing native and non-native annotators in TS evaluation, or when we use native annotators for evaluating the simplicity of the sentences simplified for non-native speakers, and vice versa.

The presented results also show that native and non-native speakers have different needs for a better understanding of the instructions in the IT troubleshooting domain. Grammatical correctness of the sentences influences the perceived simplicity of the sentences more in native than in non-native speakers. After the sentences have been split, additional simplification operations (removing superfluous words, conversion of passive to active voice, disambiguation of meaning, and conversion to the canonical subject-verb-object form) only improves the simplicity for native speakers.

We further found that inter-annotator agreements measured as the Cohen's kappa (κ) are, in most cases, very low for the majority of annotator pairs, regardless of the annotators group (native vs. non-native) and the type of evaluation performed (absolute evaluation of sentences on a 1–5 level scale vs. relative evaluation with a 'yes'/'no' answer to the question whether one sentence is simpler than the other). The only reliable results (in terms of the Cohen's kappa) seem to be those for the relative comparison of the original and fully simplified sentences.

These results call for special caution in TS evaluation, showing that crowdsourced evaluation without checking the inter-annotator agreements can result in misleading results. They also show that the currently used absolute evaluation on a 1–5 Likert scale might not be suitable for the sentence simplicity assessment, as it leads to high differences in scores across annotators even if they all belong to the same annotators group (native or non-native).

9. Acknowledgements

We thank Nikitha Sukumaran for helping in data collection and evaluation.

10. Bibliographical References

Aluísio, S. M. and Gasperin, C. (2010). Fostering digital inclusion and accessibility: The porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, YIWCALA '10, pages 46–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Angrosh, M., Nomoto, T., and Siddharthan, A. (2014). Lexico-syntactic text simplification and compression with typed dependencies. In *Procedings of COLING 2014*, pages 1996–2006.

Aranzabe, M. J., Díaz De Ilarraza, A., and González, I. (2012). First Approach to Automatic Text Simplification in Basque. In *Proceedings of the first Natural Language Processing for Improving Textual Accessibility Workshop (NLP4ITA)*.

Baeza-Yates, R., Rello, L., and Dembowski, J. (2015). Cassa: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1385. ACL.

Barlacchi, G. and Tonelli, S. (2013). ERNESTA: A sentence simplification tool for childrens stories in italian. In *Computational Linguistics and Intelligent Text Processing*.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the ACL (EACL'99)*, pages 269–270.

Coster, W. and Kauchak, D. (2011a). Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–9.

Coster, W. and Kauchak, D. (2011b). Simple English Wikipedia: a new text simplification task. In *Proceedings of ACL&HLT*, pages 665–669.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram coocurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Drndarević, B., Štajner, S., Bott, S., Bautista, S., and Saggion, H. (2013). Automatic Text Simplication in Spanish: A Comparative Evaluation of Complementing Components. In *Proceedings of CICLing 2012*, Lecture Notes in Computer Science, pages 488–500. Springer.

Freyhoff, G., Hess, G., Kerr, L., Tronbacke, B., and Van Der Veken, K., (1998). *Make it Simple, European Guidelines for the Production of Easy-toRead Information for People with Learning Disability*. ILSMH European Association, Brussels.

Glavaš, G. and Štajner, S. (2013). Event-Centered Simplication of News Stories. In *Proceedings of the Student Workshop at RANLP 2013*, pages 71–78.

Glavaš, G. and Štajner, S. (2015). Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of the ACL&IJCNLP 2015 (Volume 2: Short Papers)*, pages 63–68.

Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a lexical simplifier using wikipedia. In *Proceedings of ACL 2014 (Short Papers)*, pages 458–463.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL&HLT*, pages 211–217.

Mencap, (2002). *Am I making myself clear? Mencap's guidelines for accessible writing*.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 85–91.

Orasan, C., Evans, R., and Dornescu, I., (2013). *Towards Multilingual Europe 2020: A Romanian Perspective*, chapter Text Simplification for People with Autistic Spectrum Disorders, pages 287–312. Romanian Academy Publishing House, Bucharest.

Paetzold, G. H. and Specia, L. (2016a). Benchmarking lexical simplification systems. In *Proceedings of LREC*, pages 3074–3080.

Paetzold, G. H. and Specia, L. (2016b). Unsupervised lexical simplification for non-native speakers. In *Proceedings of the 30th AAAI*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

PlainLanguage. (2011). Federal plain language guidelines.

Rello, L., Baeza-Yates, R., Bott, S., and Saggion, H. (2013). Simplify or help? Text simplification strategies for people with dyslexia. In *Proceedings of W4A conference*.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., and Drndarevic, B. (2015). Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, 6(4):14:1–14:36.

Siddharthan, A. and Angrosh, M. A. (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 722–731.

Specia, L. (2010). Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR)*, volume 6001 of *Lecture Notes in Computer Science*, pages 30–39. Springer Berlin Heidelberg.

Štajner, S. and Glavaš, G. (2017). Leveraging event-based semantics for automated text simplification. *Expert Systems With Applications, Elsevier*, 82:383–395.

Štajner, S. and Nisioi, S. (2018). A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*.

Štajner, S., Mitkov, R., and Saggion, H. (2014). One Step Closer to Automatic Evaluation of Text Simplification Systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) at EACL*.

Štajner, S., Bechara, H., and Saggion, H. (2015a). A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. In *Proceedings of ACL&IJCNLP (Volume 2: Short Papers)*, pages 823–828.

Štajner, S., Calixto, I., and Saggion, H. (2015b). Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies. In *Proceedings of RANLP 2015*, pages 618–626.

Štajner, S., Popović, M., Saggion, H., Specia, L., and Fishel, M. (2016). Shared Task on Quality Assessment for Text Simplification. In *Proceedings of the LREC Workshop on Quality Assessment for Text Simplification (QATS)*, pages 22–31.

Woodsend, K. and Lapata, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 409–420.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*, pages 1015–1024. Association for Computational Linguistics.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the sake of simplicity: unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of RANLP*, pages 813–822, Varna, Bulgaria.

Zhang, X. and Lapata, M. (2017). Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.

Siminchik: A Speech Corpus for Preservation of Southern Quechua

Ronald Cardenas[†], Rodolfo Zevallos[⋆], Reynaldo Baquerizo[‡], Luis Camacho[∓]
[†]*Institute of Formal and Applied Linguistics Charles University in Prague, Czech Republic*
[⋆‡∓]*Grupo de Telecomunicaciones Rurales Pontificia Universidad Católica del Perú*
[†]ronald.cardenas@matfyz.cz, [⋆]rjzevallos.salazar@gmail.com
[‡]baquerizo.reynaldo@pucp.edu.pe, [∓]camacho.l@pucp.pe

Abstract

Languages are disappearing at an alarming rate, linguistics rights of speakers of most of the 7000 languages are under risk. ICT play a key role for the preservation of endangered languages; as ultimate use of ICT, natural language processing must be highlighted since in this century the lack of such support hampers literacy acquisition as well as prevents the use of Internet and any electronic means. The first step is the building of resources for processing, therefore we introduce the first speech corpus of Southern Quechua, *Siminchik*, suitable for training and evaluating speech recognition systems. The corpus consists of 97 hours of spontaneous conversations recorded in radio programs in the Southern regions of Peru. The annotation task was carried out by native speakers from those regions using the unified written convention. We present initial experiments on speech recognition and language modeling and explain the challenges inherent to the nature and current status of this ancestral language.

Keywords: Quechua, endangered languages, corpus, speech recognition

1. Introduction

Peru is a multicultural country, not by modern immigration like USA or Europe but due the presence of many native first nations making a total of 10% of the population; the bulk of remainder has the same roots that these first nations but they lost ethnic identification decades or even centuries ago. First nations are speakers of 47 languages still alive but under risk of extinction.
Linguistic research of Peruvian native languages has achieved the creation of official standard alphabets of thirty five of these languages (Cáceres et al., 2016). Unfortunately, there is insufficient private and public funding to help increase, maintain and disseminate that knowledge. As a result, Peruvian native languages, amongst which Quechua is included, present scarce written footprint and even today they are predominantly orally transmitted. Even worse, the amount of digital content in Peruvian languages is extremely low, all of them are considered underresourced, that means they meet one or more of the following characteristics:

1. they lack a unique writing system or an established grammar
2. they lack significant presence on Internet
3. they lack a critical mass of expert linguists
4. they lack electronic resources: monolingual corpora, bilingual electronic dictionaries, databases of transcribed speeches, pronunciation dictionaries or lexica.

Beyond linguistics, computational portability involves creating scalable language technology. However, language processing technologies require large datasets of joined and aligned text and speech, and parallel corpus. We can conclude that increasing the size of these kinds of corpora is an unavoidable task, despite the operational difficulties in doing so due to the lack of public funding.
In this paper, we present an initial effort to address computational portability for the Quechua family of languages. This way, we hope to boost the development of the complete set of software tools for this endangered language and finally to contribute to its revitalization.
Our main contribution is twofold:

- We present a 97 hour long speech corpus containing audio recordings of spontaneous conversations recorded in radio programs, for the two most widely spoken Quechua languages in Peru QUECHUA CHANCA and QUECHUA COLLAO.
- We perform experiments in traditional (acoustic-based) speech recognition and language modeling for this dataset.

2. Linguistics rights of Peruvian first nations

First nations have managed to survive colonial and post-colonial repression but are facing constant pressure from the side of the monolingual Spanish-speaking Peruvian society that drives them towards the abandonment of their ancestral languages.
A situation that is much less effectively addressed by programs for endangered languages is the case of historically important languages with large numbers of speakers but unclear official status, as Quechua. Such languages are located in regions where, for many centuries, their speakers have occupied an inferior socioeconomic position in trade with speakers of dominant languages; exposed to a strong economic dependence on the use of the dominant language, their feeling of pride of belonging to a language community has been hurt and now they are much more difficult to support than small speaker communities with a high degree of internal cohesion. Most of these languages are subject to a dramatic language shift that threatens to interrupt their transmission to future generations.(Adelaar, 2014).
The root of the problem is a ”structural inequality”, the legacy of colonialism is the inequality due to ethnicity. In South America, colonialism ended two hundred years ago but its consequences are still felt. In the last fifty years,

the Peruvian State issued laws to make compulsory the teaching of languages Quechua and Aymara and to declare Quechua as official language at national level (but later it was restricted just to the regional level). None of these laws really has been put into practice.
Issuance of laws is not enough if these are not enforced. All these good initiatives clashed with reality, a State apparatus neither convinced nor prepared. In spite of existing laws, in practice the State still ignores the multiculturalism and behaves as a monoculture and monolingual organization. Since this wrong paradigm is still in force, the State has not invested enough to build the linguistics skills to serve everyone equally. The consequences of this are the lack of promotion, discrimination and finally the isolation that leads to extinction of our native languages.
Top down decisions are essential but also is needed bottom up action from the grassroots. Our initiative is to change the wrong paradigm, to arouse national pride for our native roots, and to do so on three fronts:

1. demonstrate that our languages can be used in the modern technological world just like well established languages
2. demonstrate that our languages can carry contemporary culture and entertainment
3. demonstrate that our languages bring economic value to the nation, which justifies its preservation beyond the rights.

Nowadays social media should be accessible to everyone. People who are in some respect and/or to some extent functionally illiterate in a specific language are currently excluded from properly using these media. However, until now the high valuation of social media and any kind of ICT tool has reinforced the vicious circle of diglossia and even worse, created "cyberglossia". We expect turning the vicious circle into a virtuous circle by developing the three fronts mentioned, generating horizontal attitudes and practices towards all languages and reinforcing the potentialities and opportunities offered by technology related to endangered languages.
Quechua language is our first target. As said before, the basic standarization of this language is already done, the next steps to be undertaken are the building of corpus and the development of a lexical database and a spelling checker. In this paper we face the first of those challenges.

3. Background and Related Work

Even though there does not exist any Quechua speech dataset, to the best of our knowledge, various groups in Latin America and abroad have been working on Quechua language technology for the last few years. The Instituto de Lengua y Literatura Andina Amazónica (ILLA)[1] has been working on the construction of electronic dictionaries for Quechua, Aymara and Guaraní; the group Hinantin[2] at the Universidad Nacional San Antonio Abad del Cusco (UNSAAC) has produced a text-to-speech system for Cusco Quechua, a Quechua spell checker plug-in for LibreOffice (Rios, 2011) and a morphological analyzer for Ashaninka, an aboriginal language whose population is scattered across the Amazonian rainforest in Peru and Brazil.
Rios (Rios, 2016) describes a language technology toolkit that includes several things worth mentioning, such as the first morphological analyzer for Quechua, a hybrid machine translation in the direction Spanish-Quechua, and the first Quechua dependency treebank.
The AVENUE-Mapudungun project developed a machine translation system for Quechua and Mapudungun (Monson et al., 2006), and included textual and speech corpora. However, the speech dataset was only collected for Mapudungun. Resources and code can still be found in the website of the project Human Language Technology and the Democratization of information.[3]

4. The Quechua Language Familiy

Quechua is a family of languages spoken in South America with around 10 million speakers, not only in the Andean regions but also along the valleys and plains connecting the Amazonian Forest to the Pacific Ocean coastline.
Quechua languages are considered highly agglutinative with sentence structure subject-object-verb (SOV) and mostly post-positional. Table 1 contains an example of standard Quechua.

Quechua	Qichwa siminchik kan Qichwa simi-nchik ka-n
Lit. trans.	Quechua mouth-ours is
Translation	Quechua is our language.

Table 1: Sentence example of standard QUECHUA CHANCA

Even though the classification of Quechua languages remains open to research (Heggarty et al., 2005; Landerman, 1992), recent work in language technology for Quechua (Rios, 2016; Rios and Mamani, 2014) have adopted the categorization system described by Torero (Torero, 1964). This categorization divides the Quechua languages into two main branches, QI and QII. Branch QI corresponds to the dialects spoken in central Peru. QII is further divided in three branches, QIIA, QIIB and QIIC. QIIA groups the dialects spoken in Northern Peru, while QIIB the ones in Ecuador and Colombia. In this paper, we focus in the QIIC dialects, which correspond to the ones spoken in Southern Peru, Bolivia and Argentina. Mutual intelligibility between speakers of QI and QII dialects is not always given. However, QII dialects are close enough to allow mutual intelligibility (see Figure 1)
Siminchik gathers audio and text of two dialects from Southern Peru. The first one, QUECHUA CHANCA, is mainly spoken in Ayacucho and surrounding departments of Peru. The second one, QUECHUA COLLAO, is spoken in the departments of Cusco and Puno, and some Northern

[1] http://www.illa-a.org/wp/
[2] http://hinant.in
[3] https://code.google.com/archive/p/hltdi-l3/wikis/AvenueQuechuaCorpus.wiki

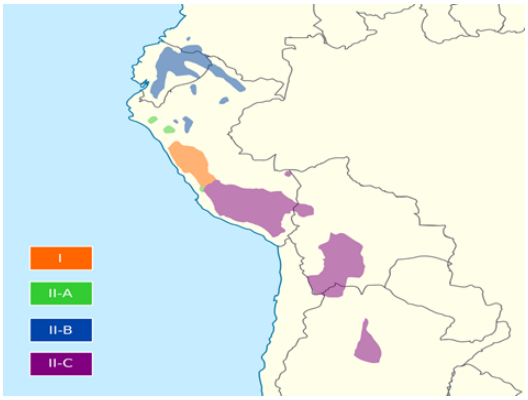


Figure 1: The four branches of Quechua language

regions of Bolivia. The main difference between these dialects is the occurrence of glottalized and aspirated stops in QUECHUA COLLAO, a phonetic distinction that QUECHUA CHANCA lacks.

4.1. Normalization for Dialect Variation

While the semantic meaning and function of most of the 130 recognized morpheme units are shared by all languages in the family, the final forms (spoken and written) differs. In an effort to create a unified standard for written language, several standards were proposed, most notably the ones proposed by Profesor Cerron-Palomino (Cerrón-Palomino, 1994) and by the *Academy of Cusco Quechua*. In 1985, Ministry of Education of Peru issued the order 1218-85-ED setting the official alphabet of Quechua (and also of Aymara) which is still in force and this is the normalization system we reference as standard as well in the rest of this paper.

The standard proposed by the *Academy of Cusco Quechua* exclusively for QUECHUA COLLAO, presents a five-vowel system. Despite the fact this standard was rejected by the Ministry of Education and by linguists, because of the lack of formal and extensive study, *Academy of Cusco Quechua* keeps spreading its ideology (Coronel-Molina, 2007) causing interference in the dissemination of official standard.

It's worth to say that Cerron-Palomino's standardized orthography was adopted by the Bolivian government with only one change: the glottal fricative consonant is written as *j* instead of *h*.

4.2. Phonemic Inventory

QUECHUA CHANCA has a total of 15 consonants, most of them voiceless, as shown in Table 2. As in Spanish, the phoneme [t͡ʃ] is written as *ch*, [ɲ] as *ñ* , and [ʎ] as *ll*.

QUECHUA COLLAO also has a glottal and an aspirated version of each plosive consonant, leading to a total of 25 consonants. However, the use of voiced consonants present in the Spanish phonemic inventory is common due to the large number of borrowings present in all dialects.

Both quechua dialects present only the vowels *a* [æ], *i* [ɪ] and *u* [ʊ], although in proximity of /q/, they are pronounced as [ɑ], [ɛ] and [ɔ], respectively. The online database OMNIGLOT has entries for both the Chanca [4] and the Collao standard [5].

	Bil	Alv	Pal	Vel	Uvu	Glo
Plosive	p	t	t͡ʃ	k	q	
Nasal	m	n	ɲ			
Fricative		s				h
Lat. Approx.		l	ʎ			
Approximant		ɹ				
Semi-consonants	w		y			

Table 2: Consonants in the phonemic inventory of QUECHUA CHANCA (IPA)

5. Siminchik Speech Corpus

The collection and curation of SIMINCHIK (Quechua word for *"our language"*) speech corpus is part of a much bigger initiative that seeks to digitally preserve ancestral Quechua and revitalize it for daily use by current and future generations. Such a colossal enterprise requires thousands of hours of quality audio from every dialect along with their transcriptions.

5.1. Crowd-Sourcing Transcription Annotation

In our effort to construct an audio corpus and a corresponding transcribed text corpus, we first began collecting audio recorded from radio programs. These included QUECHUA CHANKA from Ayacucho, Apurimac and QUECHUA COLLAO from Cusco, Puno.

The initial collection comprised of hundreds of raw audio hours that contained spot advertisements, music and Spanish speech. To separate speech from music we employed a voice detector using pyAudioAnalysis (Giannakopoulos, 2015). The audio was converted to mono channel, resampled at 16 kHz, encoding with 16 bits precision and saved in the WAV format. In preliminary experiments, it was found that radio locutors speak fast enough for the voice detector to not be able to locate pauses between sentences. Hence, we opted for splitting any audio detected as voice in segments no longer than 30 seconds each. The main reason for this was to avoid dealing with long sequences, as it is known that sequence models' performance decay rapidly with the length of the sequence. This resulted in words quite possibly being truncated at the beginning and at the end of the segment.

Given the large amount of data, we crowdsourced the task of annotating orthographic transcriptions for each audio segment. We built a specialized website for this purpose [6] and called for native speakers of both dialects to voluntarily participate. Each audio clip was annotated by exactly two annotators. The number of annotators reached and transcribed hours are presented in Table 3.

The website addressed the accuracy limitations of the voice detector by allowing the users to mark a given audio as containing only Spanish speech, only music, Quechua speech

[4] https://www.omniglot.com/writing/ayacuchoquechua.htm
[5] https://www.omniglot.com/writing/quechua.htm
[6] https://siminchikkunarayku.pe/corpus/transcribir

	Female		Male	
	Quechua Chanca	Quechua Collao	Quechua Chanca	Quechua Collao
# annotators	160	63	153	56
# transcribed hours	58.28	12.95	17.74	8.03

Table 3: Details of volunteer annotators

(possibly with non-loud music in the background) or entirely unintelligible. The interface also includes the option to indicate in which second the transcription begins and ends. This is in order to make up for a potentially truncated initial or final word because of the audio's hard splitting.

The statistics of the speech corpus are presented in Table 4. A total of 97 hours have been transcribed so far.

5.2. Text Corpus

Although the collected speech data was sufficient for acoustic modeling, it was insufficient for n-gram estimation. Hence, we gathered text corpora from legal documents, such as the Peruvian Constitution and the Declaration of Human Rights; as well as Andean tales for children prepared by the Ministry of Education of Peru, and the New Testament of the Bible. The text is normalized for further downstream usage using (Rios, 2016)' normalizer.

The statistics of the final text corpus are shown in Table 5. For our language modeling experiments, this corpus was split into three corpora for training, evaluation (validation) and testing, in an 80/10/10 ratio.

5.3. Preprocessing and Normalization of Transcriptions

Besides standard preprocessing for speech recognition, such as punctuation removal and case normalization, special care had to be taken for the normalization of interjections, those common to all dialects and those rooted in the culture of the community of the speaker. For this purpose we created a dictionary of said interjections (e.g. *mamay, ayy, mmm, ahaa*) and crafted regular expressions in order to map each orthographic variation to a fixed word form. Annotators were encouraged to use the Spanish writing system for proper names, since almost all of these are loan words from the Spanish language.

After cleaning the text, it was necessary to normalize dialectal variations to the standard written format. We used the morphological analyzer and normalizer developed by (Rios, 2016), which normalizes to the CHANCA standard. However, spelling error still remained unchanged, since annotators were not provided a spell checker in the web tool. We dealt with this by using a language model to score the two transcriptions of each audio segment and choosing the most probable one. By doing this, we keep only the less noisy transcription for each audio segment. This language model was trained over the normalized corpus presented in section 5.2.. Section 6.2. presents details of the design and experimental results of this model. Then, we manually curated all remaining transcriptions.

6. Experiments

In this section we analyze the quality of our speech corpus by conducting acoustic and language modeling experiments.

6.1. Acoustic Modeling

The acoustic model for SIMINCHIK was built using the Hidden Markov Model Toolkit (HTK) (Young et al., 2002) because it has a better performance in dynamic decoder (dictionary, language model and acoustic model) (Ganesh and Sahu, 2015) and phonetic segmentation with a tolerance region less than 30ms (Matoušek and Klíma, 2017).

We sub-sampled the dataset, obtained 8 hours of training and 2 hours of testing data for a total of 16,340 instances. The training set includes audio with 9 speakers (3 male and 6 female), while the test set includes two (one male and one female). We work just with 10 out of a total of 97 hours of audio because only this 10 hours were speech-text aligned at word level, as required to run the experiment; the 87 hours left were aligned at sentence level.

Each utterance was split into Hamming windows of 25ms with an offset of 10ms. Acoustic parameters were 39 MFCCs with 12 Mel cepstrum plus log energy and their delta (first order derivative) and acceleration (second order derivative) coefficients.

Our monophone models consist of 5-state HMMs in which the first and last state are non-emitting states. The prototype model were initialized using HCompV which allows for a fast and precise convergence of the training algorithm. Next, HMM parameters were re-estimated 5 times using the HERest and HHEd which computes the optimum values of HMM parameters using gaussian mixtures. These parameters were re-estimated repeatedly using the training data until re-estimation converged.

The experiment follow the procedures mentioned in (Odriozola et al., 2014; Dua et al., 2012).

Table 6 shows promising results for word level accuracy and word error rate, obtaining averages of 82.8% and 17.2%, respectively.

6.2. Language Modeling

Given the high presence of rare words in morphologically rich languages such as Quechua, we use a singleton pruning rate κ of 0.05 as proposed by (Botha, 2015), in order to randomly replace only a fraction κ of words occurring only once in the training data with a global UNK symbol.

We built a 5-gram interpolated Modified Kneser-Ney model at the word level, obtaining a perplexity of 298.79. A vocabulary intersection analysis further revealed a 63.46% intersection between the validation set's and training set's vocabulary. This rather high value of perplexity and low vo-

	Total	Training	Validation	Test
Words	448,919	404,625	22,526	21,768
Vocabulary	89,767	83,258	9,538	9,433
Length	97.5h	88h	4.8h	4.7h

Table 4: Statistics of the Quechua dataset.

	Total	Training	Validation	Test
Words	1'211,936	1'011,335	103,478	97,123
Vocabulary	174,151	155,655	30,776	29,024
Sentences	70,790	57,735	6,300	6,755

Table 5: Text corpus statistics

cabulary intersection reveal the morphological richness of this family of languages. A more thorough inspection of the vocabulary showed that 64.42% of words have a frequency of one.

Speaker	# utterances	WACC	WER
S1	1,509	83.2	16.8
S2	3,000	82.2	17.8
Total	4,509	82.8	17.2

Table 6: Acoustic model results for two speakers (one male and one female) in per-word accuracy (WACC) and word error rate (WER)

7. Conclusions & Further Work

Driven by social and attitudinal factors deeply rooted in history, massive language shift of endangered languages with millions of speakers is difficult to reverse, and it requires more complex techniques than language maintenance and revitalization in small language communities.

By the fact that Quechua official spelling hasn't been well disseminated, Quechua speakers are to some extent functionally illiterate in their own language. Our approach will in the end enable them to use spoken input, that will be converted automatically in standard written format.

As first step of that medium-term goal, we have collected and curated raw audio from radio shows to build the first dataset of Quechua speech: SIMINCHIK. It consists of 97 hours of spontaneous Quechua speech and the corresponding transcribed text for the Chanca and Collao, dialects from the southern regions of Peru.

We present initial ASR experiments using a HMM acoustic model, yielding a promising word error rate of 17.2%. The highly agglutinative nature of the language is challenging for word form based tasks, such as n-gram language modeling, POS tagging, speech recognition, among others. Similar challenges can be found in ASR systems for typologically similar languages such as Turkish (Carki et al., 2000) and Basque (Odriozola et al., 2014).

We are currently introducing neural networks instead a HMM acoustic model. The new model is to reproduce the work of (Graves and Jaitly, 2014) including extensions made by (Amodei et al., 2016), that means to introduce neural networks in each stage of the process.

Further work on ASR of Quechua language will proceed in use Knowledge Transfer described by (Zhao et al., 2014) to build a model capable of learning several languages, in this case Basque and Quechua.

Finally, this research is a part of Siminchikkunarayku [7] initiative which vision is that the future of the languages of America depends upon the conservation of the native languages, but also upon the polyglotism of the citizens. Siminchikkunarayku follows two orientations:

1. "language as a right" views minority languages as a right to which their speakers are entitled. Following this, preservation of Quechua is an identity reinforcement and recognition of that population.

2. "language as a resource" sees the minority languages as potential economical resources for the whole nation. A lot of research must be done to make visible the profitability of preservation.

8. Acknowledgements

This project was supported by CONCYTEC CIENCIACTIVA of the Peruvian government through grant 164-2015-FONDECYT and by PUCP through grant 2017-3-0039/436

9. Bibliographical References

Adelaar, W. F. (2014). Endangered languages with millions of speakers: Focus on quechua in peru. *JournaLIPP*, (3):1–12.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182.

Botha, J. A. (2015). Probabilistic modelling of morphologically rich languages. *arXiv preprint arXiv:1508.04271*.

Cáceres, R., Cavero Cornejo, O., Gutiérrez, D., et al. (2016). Diagnóstico descriptivo de la situación de los pueblos originarios y de la política de educación intercultural bilingüe en el perú.

[7] https://siminchikkunarayku.pe

Carki, K., Geutner, P., and Schultz, T. (2000). Turkish lvcsr: towards better speech recognition for agglutinative languages. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1563–1566. IEEE.

Cerrón-Palomino, R. (1994). Quechua sureño diccionario unificada quechua-castellano castellano-quechua [unified dictionary of southern quechua, quechua-spanish spanish-quechua]. *Lima: Biblioteca Nacional del Perú*.

Coronel-Molina, S. M. (2007). *Language policy and planning, and language ideologies in Peru: The case of Cuzco's High Academy of the Quechua Language (Qheswa simi hamut'ana kuraq suntur)*. University of Pennsylvania.

Dua, M., Aggarwal, R., Kadyan, V., and Dua, S. (2012). Punjabi automatic speech recognition using htk. *IJCSI International Journal of Computer Science Issues*, 9(4):1694–0814.

Ganesh, D. S. and Sahu, P. K. (2015). A study on automatic speech recognition toolkits. In *International Conference on Microwave, Optical and Communcation Engineering*.

Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12).

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772.

Heggarty, P., Valko, M. L., Huarcaya, S. M., Jerez, O., Pilares, G., Paz, E. P., Noli, E., and Usandizaga, H. (2005). Enigmas en el origen de las lenguas andinas: aplicando nuevas técnicas a las incógnitas por resolver. *Revista Andina*, 40:9–57.

Landerman, P. N. (1992). Quechua dialects and their classification. *PhD Thesis*.

Matoušek, J. and Klíma, M. (2017). Automatic phonetic segmentation using the kaldi toolkit. In *International Conference on Text, Speech, and Dialogue*, pages 138–146. Springer.

Monson, C., Llitjós, A. F., Aranovich, R., Levin, L., Brown, R., Peterson, E., Carbonell, J., and Lavie, A. (2006). Building nlp systems for two resource-scarce indigenous languages: Mapudungun and quechua. *Strategies for developing machine translation for minority languages*, page 15.

Odriozola, I., Serrano, L., Hernaez, I., and Navas, E. (2014). The ahosr automatic speech recognition system. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 279–288. Springer.

Rios, A. and Mamani, R. C. (2014). Morphological disambiguation and text normalization for southern quechua varieties. *COLING 2014*, page 39.

Rios, A. (2011). Spell checking an agglutinative language: Quechua. *University of Zurich. Zurich Open Repository and Archive*.

Rios, A. (2016). A basic language technology toolkit for quechua.

Torero, A. (1964). Los dialectos quechua.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2002). The htk book. *Cambridge university engineering department*, 3:175.

Zhao, Y., Xu, Y. M., Sun, M. J., Xu, X. N., Wang, H., Yang, G. S., and Ji, Q. (2014). Cross-language transfer speech recognition using deep learning. In *Control & Automation (ICCA), 11th IEEE International Conference on*, pages 1422–1426. IEEE.

==========