LREC 2016 Workshop

Improving Social Inclusion Using NLP: Tools and Resources

PROCEEDINGS

Edited by

Ineke Schuurman, Vincent Vandeghinste, Horacio Saggion

23 May 2016

Proceedings of the LREC 2016 Workshop
"Improving Social Inclusion Using NLP: Tools and Resources"

23 May 2016 – Portorož, Slovenia

Edited by Ineke Schuurman, Vincent Vandeghinste, Horacio Saggion

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

Acknowledgments: This workshop is organized in the context of the Able-To-Include project (European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme), http://able-to-include.com

Organizing Committee

- Ineke Schuurman, KU Leuven (BE)
- Vincent vandeghinste, KU Leuven (BE)
- Horacio Saggion, Universitat Pompeu Fabra, Barcelona (ES)

Programme Committee

- Susana Bautista, Federal University of Rio Grande do Sul (BR)
- Heidi Christensen, University of Sheffield (UK)
- Onno Crasborn, Radboud University (NL)
- Koenraad De Smedt, University of Bergen (NO)
- Nuria Gala, Aix-Marseille Université (FR)
- Peter Ljunglöf, University of Gothenburg (SE)
- Isa Maks, VU University Amsterdam (NL)
- Davy Nijs, UC Leuven-Limburg (BE)
- Jean-Pierre Martens, Ghent University (BE)
- Martin Reynaert, Tilburg University and Radboud University (NL)
- Horacio Saggion, Universitat Pompeu Fabra (ES)
- Ineke Schuurman, University of Leuven (BE)
- Liz Tilly, University of Wolverhampton (UK)
- Vincent Vandeghinste, University of Leuven (BE)
- Hugo Van hamme, University of Leuven (BE)

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. On the one hand, 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. On the other hand, it can be problematic for immigrants who want to be integrated in the digital society of their host country but do not master the language of their new home.

Several technologies can make a difference in the accessibility of information 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. Extra linguistic information, such as definitions, can be used to clarify the content. In the case of people who are to some extent 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 Workshop on Improving Social Inclusion using Natural Language Processing (NLP) is to bring together researchers and practitioners in the areas of social inclusion and natural language processing to understand problems faced with text accessibility in social media by different social groups, describe current development in language resources and methods for these problems, and discuss future research directions. Of particular interest is how techniques and resources developed for one language and domain can be ported to a different language or domain.

In particular the workshop aimed at the following topics, in relation to social inclusion:
input methods of non-verbal input, input normalization, text adaptation, semantic representations, generation, evaluation, ethics, and reusability of Social Inclusion-approaches.

We thank Lucia Specia from the University of Sheffield for giving an invited presentation entitled "Text Simplification for Social Inclusion".

We thank all the authors for their contributions, the members of the programme committee for their extended (and timely) reviews, and the LREC 2016 Conference for hosting our workshop.

the organizers,

Ineke Schuurman
Vincent Vandeghinste
Horacio Saggion

Programme

09.00 – 09.15 Opening

09.15 – 09.40 Liz Tilly
Issues relating to using a co-productive approach in an accessible technology project

09.40 – 10.05 Krzysztof Wróbel, Dawid Smoleń, Dorota Szulc, Jakub Gałka
Development of the First Polish Sign Language Part-of-Speech Tagger

10.05 – 10.30 Leen Sevens, Tom Vanallemeersch, Ineke Schuurman, Vincent Vandeghinste, FrankVan Eynde
Automated Spelling Correction for Dutch Internet Users with Intellectual Disabilities

11.00 – 11.25 Victoria Yaneva, Richard Evans,Irina Temnikova
Predicting Reading Difficulty for Readers with Autism Spectrum Disorder

11.25 – 11.50 Estela Saquete, Ruben Izquierdo Bevia, Sonia Vazquez
SimplexEduReading: Simplification of Natural Language for Reading Comprehension Improvement in Education

11.50 – 12.40 Lucia Specia
Text Simplification for Social Inclusion (invited talk)

12.40 – 12.55 General discussion

12.55 – 13.00 Closing

Table of Contents

Issues Relating to Using a Co-productive Approach in an Accessible Technology Project

Liz Tilly
University of Wolverhampton
Room MH 125
Faculty of Education, Health and Wellbeing
Mary Seacole Building
City Campus - North
Wolverhampton
WV1 1AD
UK
E-mail: Liz.Tilly@wlv.ac.uk

Abstract

This paper discusses the issue of accessibility to being online and using social media for people with a learning disability, and the challenges to using a co-production approach in an accessible technology project. While an increasing number of daily living tasks are now completed online, people with a learning disability frequently experience digital exclusion due to limited literacy and IT skills. The Able to Include project sought to engage people with a learning disability as active partners to test and feedback on the use and development of a pictogram app used to make social media more accessible. The challenges mainly related to the feedback needing to be sent electronically to the partners; there was only minimal contact with them and no face to face contact. The paper also outlines how other challenges were overcome to enable genuine and meaningful co-production. These included addressing online safety and ethical issues regarding anonymity.

Keywords: People with a learning disability, digital inclusion, accessible technology, co-production, well-being, online safety

1. Terminology

This article uses the term *people with a learning disability* as this is the way that the Building Bridges Training group chose to describe themselves. It is noted that Inclusion Europe and other organisations and institutions would use the terminology *people with intellectual disability* to describe this same group.

2. Introduction

Ongoing developments in information and communication technology and especially the internet, are changing all aspects of life. The advent of smartphones and tablets has made the internet more portable, convenient and accessible, and this includes benefitting people with a learning disability (Foley and Ferri, 2012). However they also experience digital exclusion, and the concept of the digital divide has been used in connection with this group (McKenzie, 2007). Many of them do not have access to computers and other devices, or the internet, to the same extent as the general population (Chadwick, Wesson and Fullwood, 2013; Hoppestad, 2013). There are several reasons for their digital exclusion, including: poverty resulting in lack of access to computers, limited access to the internet, lower skills, and fewer learning and training opportunities, with access often being controlled by parents or staff.

Being online is now a central part of everyday life for many, and social media websites such as Facebook and Twitter enable billions of people worldwide to interact with others instantaneously on the internet. In 2015 Facebook announced it had in excess of 1.44 billion monthly active users worldwide, an increase of 13% on the previous year (Protalinski, 2015).

Ninety per cent of people with a learning disability live independently and without the support of specialist services (Emerson and Hatton, 2008). Due to recent austerity measures in the UK there has been a decrease in the amount of proactive and preventative community-based support, which has resulted in more people with a learning disability having to try to cope with independent living with little or no staff support (Money Friends and Making Ends Meet Research Group, 2011; Tilly, 2012). Added to this, there is an increasing expectation that people are managing various aspects of their daily lives such as travel planning and financial services through online services.

3. Background

Building Bridges Training is a partner in a European-funded research and development project called Able to Include, part of the EU CIP ICT Programme, reference number CIP-ICT-PSP-2013-7 (see https://ec.europa.eu/digital-agenda/en/ict-policy-support-programme) which runs from 2014 to 2017. The aim is to develop and pilot accessible technology for smartphones, tablets, and similar devices, for people with a learning disability, focusing on the use of social media and independent travel. It provides an accessibility layer to make any smartphones and tablets accessible, so has wide-reaching benefits. There are nine partners involved, comprising six universities and technical companies who are developing the software and three not-for-profit organisations who are enabling their people with a learning disability to pilot the technology in real situations. These organisations are based in Belgium and Spain, with Building Bridges Training as the UK partner (www.abletoinclude.eu).

4. The Right to Accessible Information

People with a learning disability have challenges in both expressive and receptive communication and often have literacy difficulties, but they have a right to information in an accessible format. The UK Equality Act 2010 places a legal duty on service providers to make reasonable adjustments to ensure accessibility for disabled people. This includes services and information from service-provider agencies. The Convention on the Rights of Persons with Disabilities (United Nations General Assembly, 2006) outlines the fundamental rights for people with a disability.

Article 21 focuses on freedom of expression and opinion, and access to information which include:

a) Providing information intended for the general public to persons with disabilities in accessible formats and technologies appropriate to different kinds of disabilities in a timely manner and without additional cost;

b) Accepting and facilitating the use of sign languages, Braille, augmentative and alternative communication, and all other accessible means, modes and formats of communication of their choice by persons with disabilities in official interactions;

c) Urging private entities that provide services to the general public, including through the internet, to provide information and services in accessible and usable formats for persons with disabilities;

d) Encouraging the mass media, including providers of information through the internet, to make their services accessible to persons with disabilities;

e) Recognising and promoting the use of sign languages.

This outlines the need for proactive interventions and product development to enable people with a learning disability to have access to IT and social media.

5. People with a Learning Disability and Access to the Internet

Access to the internet can contribute to social inclusion; it can help people with a learning disability to keep in touch with others and reduce social isolation (Holmes and O'Loughlin, 2014), learn new skills, and gain access to information in a more accessible format, and with visual rather than written information. This can help them with living more independently and feeling more in control of their lives, which is the primary focus for people with a learning disability (Department of Health, 2001).

Hoppestad (2013) highlights how people with intellectual disabilities have limited use of technology, particularly those who are adults and those with more severe disabilities. One reason for this is that it takes time and a considerable amount of effort to help teach these individuals to use computers. Chadwick, Wesson and Fullwood (2013) reported inequalities and fewer opportunities available to individuals with an intellectual disability to go online, noting how people with a learning disability often find it hard to gain full access to the internet. This can be due to a number of factors including the physical and cognitive impairments of the person with disabilities but also, taking a more social model stance, because the internet is designed with little consideration for the needs of people with intellectual disabilities and because the carers who support them may act as gatekeepers to accessing the internet (Chadwick et al., 2013).

Enabling accessibility to social media is therefore essential for people with a learning disability and has been the focus of several technology projects (Davies et al., 2015).

6. Co-production

The concept of co-production has evolved since the late 1990s and has been a UK government approach for the design and delivery of social care services since the previous New Labour Government (Department of Health, 2010). The term was specifically referred to in the 2010 Government learning disability strategy (Department of Health, 2010).

While there is currently no one definition of 'co-production' as it is a concept still developing and changing, co-production recognises that people with learning disabilities are experts in their own lives and therefore essential partners in identifying, designing, delivering, monitoring and reviewing both commissioned and universally available services. In order to achieve the ambitions of the UK Care Act 2014 around prevention, well-being and a strong focus on outcomes, within the current climate of austerity and budget cuts, transformational co-production is recognised as having a vital role in shaping future services and opportunities for adults with learning disabilities.

7. Able to Include Project

The Able to Include project was developed with the aim of enabling people with a learning disability to read and understand simple written text (Able to Include, 2016). It has created an open source context aware accessibility layer using three technologies; text simplification, text to speech and text to pictograms tools. This paper will focus on this **Text2Picto** technology, co-developed by partners KU Leuven, it provides a text-to-pictogram and a pictogram-to-text translation, which are standardised image-based representations of words or concepts. Two pictogram sets are used in the project; Beta and Sclera, the former using colour and the later black and white more simplistic images.

The project recruited participants with a learning disability living in the West Midlands, UK, through email invitations to colleagues working in learning disability provider services and advocacy groups from the statutory and voluntary sector. No specific definition of learning disability was given, and eligibility for participation was based on identifying with an organisation or service that supported people with a learning disability. Invitations to participate and consent forms were all developed in an easy read format. Ethical approval was given by the FEHW ethics committee, University of Wolverhampton.

The project was delivered in two stages. The first stage was to gauge internet and smart device access and use, through questionnaires and pilot groups, exploring the participants' current use of the internet, IT, smartphone devices and social media.

The second stage used feedback and observations from group-work sessions piloting the Able to Include app on tablets to explore:

- how this group engaged with learning how to use tablets and the new apps;

- how this new technology impacted their lives, with a particular focus on independent living and citizenship;
- how they experienced their role as co-designers with the technical partners.

This feedback was shared with the relevant project partners to enable them to make adaptions to the app and the pictogram translation service.

7.1 Stage One – Information-gathering

In the autumn of 2014, 53 people with a learning disability were recruited to participate in the project, after completing an initial questionnaire in an easy read format on their internet use. Five focus groups, totalling 30 people were then held to further discuss the topics. The participants were all aged over 18, from a range of settings including family homes, residential care, supported living and living independently in the community. All were able to communicate in the group settings and contribute to the discussions, but displayed a range of literacy skills. The following information was established through analysis of the questionnaires and the focus group discussions:

7.1.1. Access to the Internet

Around 58% (31) had access to the internet at home, using their own or their parents' devices and broadband. Another 16% (9) were able to access the internet via friends or family or community facilities, while 24% (13) said they did not use the internet. Reasons given for not using the internet, included the cost, respondents finding the internet too complicated or perceiving a risk to their own or their financial well-being, and lack of access, suggesting that they did not have the means or assistance to enable them to gain access.

7.1.2. Website Usage

The majority of respondents who used the internet had visited Facebook, Google, YouTube and gaming sites. Using the internet for shopping, general interests, hobbies and sport were the most popular searches, and a few had visited music and TV sites.

7.1.3. Mobile Phones

Of the 53 respondents, 86% (46) had mobile phones, 25 of which were smartphones. For those with phones, 43% indicated they were on a contract, with 54% using Pay As You Go. The reasons given for the respective payment methods were based on ease of use and perceived cheaper costs. Most of the respondents with smartphones (88%) used apps, including for TV and films, Facebook, Twitter and YouTube.

7.1.4. WhatsApp

Ten respondents indicated a preference for text messaging compared with two who preferred WhatsApp, while eight thought WhatsApp was as easy to use as text messaging.

7.1.5. Facebook

Of the 53 respondents, 22 (41%) had Facebook accounts. The number of 'friends' each had ranged from 2 to 400. Eight of the respondents reported finding it hard to post messages or photographs, and it was suggested that better instructions, pictures and training would help to make posting messages and photographs easier. It was also noted that not changing the Facebook format would make it easier to use.

7.1.6. Twitter and Instagram

Only 8 (15%) of the respondents had Twitter accounts, which they used for following celebrities or sport, and 3 (5%) of the respondents used Instagram.

7.1.7. Tablets

Tablet devices had been used by 23 (43%) of the respondents, with 17 (32%) expressing a preference for them, citing their portability, speed of start-up and the number of available apps.

7.1.8 Skills

There was a range of skill levels among the respondents, with 11 (20%) indicating no ability to communicate via text messaging or e-mail, 10 (18%) with no capability to make use of device manual interfaces (keyboards, touchpads, mice etc.), 19 (35%) saying they were unable to browse the internet or shop online, and 7 (13%) respondents reporting incapability to access or communicate via the internet.

7.1.9. Reported Difficulties Using the Internet

Among the reported difficulties in using the internet were that the search engine results were heavily reliant on verbal accuracy; there was a requirement for too many passwords; and online payment systems could be complicated. There was also the issue of the visibility of personal details which meant that security was unclear.

7.1.10. What Would Make Using the Internet Easier

Participants discussed things that would make using the internet easier such as including vocal and pictorial search and response mechanisms, being able to have training on the use of passwords, and having speech options introduced for online payments.

7.2. Stage Two – Testing the Pictograms

Those who wished to continue with the project were invited to attend a regular group to learn how to use 10" and 12" tablets, and to continue to test out the pictogram translation through the demonstration website and the new app on Facebook.

The group were given opportunities to reflect on their experiences at key stages in the project through short semi-structured interviews, and evaluation activities such as completing response sheets with smiley faces etc. This data was forwarded electronically to the project partners to inform them of future technical developments.

A range of issues were found with using both sets of pictogram programmes for example using local words such as 'mash' for mashed potato or phrases such as 'coach trip'. The participants reported that they preferred coloured images as in Sclera and photos in preference to line drawings, and did not like childish representations. In the pilot of the text to pictogram tools it was found that participants mainly wanted to make statements about what they had recently done or were going to do, which was difficult to clarify in the pictogram translation. Another

difficulty was that much of the content that people wanted to communicate was about their daily lives and experiences and so included many proper nouns such as local towns, shopping malls and people's names, which meant that their messages had limited success. They also wanted to use expressions such as 'Christmas fayre' or 'Blackpool illuminations'. There were no pictos for some of the words they used such as curry, England or even learning disability. Some words produced the wrong picto for the context eg the score in 'football score' was a music score, match as in 'football match' produced a picto of a match as in a source of ignition, and soap as in soap opera produced a picto of a bar of soap. The people involved in the sessions understood the results were being fed back to colleagues in Belgium, and they found the errors mainly comical, and gave them a sense of superiority in that they could see 'the computer got it wrong again'.

8. Personal Safety

Issues of safety, risk and protection online for people with a learning disability are a major concern (Holmes and O'Loughlin, 2014) and it is recognised that further investigation is required. Such issues were noted by staff, family carers and even people with a learning disability themselves as hindrances to gaining online access, and especially engaging with social media. Being online should enable people with a learning disability to be anonymous and have a different identity (McClimens, 2003). However, literary skills and lack of knowledge of current trending online language will inevitably reveal their learning disability, especially if they also need to use accessibility tools such as the pictos. But vulnerabilities exist regardless of whether online contacts are aware of the users' learning disability.

The issue of vulnerability was given due consideration in the Able to Include piloting sessions (ARC, 2012) with the following risks identified:

- Having a relatively expensive device either at home or in the community. This was managed by risk assessments and keeping the tablets only for group activities, but it was recognised as hampering the opportunity to develop skills.
- Travelling in the community independently to attend the piloting sessions, with the risk from traffic and potential harassment. This was managed by travel plans and risk assessments being put in place and checking up on people if they did not arrive on time.
- Online grooming and the risk of meeting people online who they may then arrange to meet and who may be abusive.

The above risks were managed by delivering community safety training and online safety training at the beginning of the project and reinforcing it at key stages through fun activities such as quizzes and games. Two Facebook accounts were opened specifically to be used in the piloting sessions, to enable the participants to communicate with each other in a restricted online environment.

9. Collaboration and Communication with Partners

This was delivered via the non-disabled workers from Building Bridges Training rather than the participants themselves which therefore added a layer of interpretation to the findings. Team meetings with the partners only included non-disabled people involved in the project. To date there has only been one opportunity for the users to communicate with a technical partner using Skype, the success of which suggests this should be used further. In subsequent sessions one of the participants would frequently ask if we would be making a Skype call. The overall lack of direct contact between the technical partners and the end users, however, meant that all the feedback was communicated via a third party and this failed to give the end users a sense of making a contribution, or any opportunity for a response to their feedback. Nevertheless small steps were made to make these relationships more real, such as looking at the Able to Include website, and at the partners information and even enabling the participants to email the other partners themselves.

10. Contribution to Research Output

Inclusive or collaborative research has its own particular challenges since people with a learning disability often choose to be openly acknowledged for their contribution to a research project (Iriarte, O'Brien and Chadwick, 2014; Tilly and Building Bridges Research Group, 2015). This can cause some tension with an ethical approach to research which typically seeks to enable participants, especially those deemed vulnerable, to remain anonymous. It was agreed that this aspect would be handled sensitively to enable the project participants to produce a paper themselves on their involvement in the project and so to be the authors of their own work while also affording them privacy. In previous publications by Building Bridges Training first names only have been used and no addresses included.

11. Conclusion

Direct involvement of the end users in assisted technology projects is essential for the project to have reliability and be widely applicable to the end users. A range of practical measures need to be put in place to ensure any risks to personal safety are managed well. Planning also needs to be implemented to enable more first-hand feedback rather than via a third party so that the participants feel their contribution to be valid and relevant.

12. References

Able to Include. (2016). Text to pictograms. Retrieved 18 March 2016 from http://able-to-include.com/text-to-pictograms/.

ARC. (2012). Safety Net Project. Retrieved 28 February 2012 from www.arcsafety.net/.

Care Act (2014) www.legislation.gov.uk/ ukpga/2014/23/ pdfs/ukpga _20140023_en.pdf.

Chadwick, D., Wesson, C. and Fullwood, C. (2013). Internet access by people with intellectual disabilities: inequalities and opportunities. *Future Internet*, 2, pp. 376–397.

Davies, D., Stock, S., King, L., Brown, R., Wehmeyer, M. and Shogren, K. (2015). An interface to support independent use of Facebook by people with intellectual disability. *Intellectual & Developmental Disabilities*,

53(1), pp. 30–41.

Department of Health. (2001*). Valuing People: A New Strategy for Learning Disability for the 21st Century*. London: The Stationery Office.

Department of Health. (2010). *Practical Approaches to Co-production: Building Effective Partnerships with People using Services, Carers, Families and Citizen*s. London: Department of Health.

Department of Health. (2010). *Valuing People Now: The Delivery Plan 2010-2011*. London: Department of Health.

Emerson, E. and Hatton, C. (2008). *People with Learning Disabilities in England*. Lancaster University: Centre for Disability Research (CeDR).

Foley, A. and Ferri, B.A. (2012). Technology for people, not disabilities: ensuring access and inclusion. *Journal of Research in Special Educational Needs*, 12(4), pp. 192–200.

Holmes, K.M. and O'Loughlin, N. (2014). The experiences of people with learning disabilities on social networking sites. *British Journal of Learning Disabilities*, 42(1) pp.1–5.

Hoppestad, B.S. (2013). Current perspective regarding adults with intellectual and developmental disabilities accessing computer technology. *Disability and Rehabilitation: Assistive Technology*, 8(3), pp. 190–194.

Iriarte, E.G., O'Brien, P. and Chadwick, D. (2014). Involving people with intellectual disabilities within research teams: lessons learned from an Irish experience. *Journal of Policy and Practice in Intellectual Disabilities*, 11(2), pp.149–157.

McClimens, A. (2003). Mixing on the net. *Nursing Standard*, 17(38), p. 26.

McKenzie, K. (2007). Digital divides: the implications for social inclusion. *Learning Disability Practice***,** 10(6): 16-21.

Money Friends and Making Ends Meet Research Group (2011). Making ends meet - what it's like for people with learning difficulties living in the community on low incomes. *Community Living*, 25(2), pp.16–17.

Protalinski, E. (2015). Facebook passes 1.44B monthly active users and 1.25B mobile users; 65% are now daily users. From http://venturebeat.com/2015/04/22/facebook-passes1-44b-monthly-active-users-1-25b-mobile-users-and-936-million-daily-users/.

The Equalities Act (2010). London: Stationery Office.

Tilly, L. (2012). *Making Ends Meet*. Norah Fry Research Centre, Bristol, University of Bristol. Unpublished PhD.

Tilly, L. and Building Bridges Research Group. (2015). Being researchers for the first time: reflections on the development of an inclusive research group. *British Journal of Learning Disabilities*, 43(2), pp. 121–127.

United Nations General Assembly. (2006). *Convention on the Rights of Persons with Disabilities. Operation Protocol to the Convention*. New York: United Nations, Centre d'Orsay.

Development of the First Polish Sign Language Part-of-Speech Tagger

Krzysztof Wróbel[1,2], **Dawid Smoleń**[1], **Dorota Szulc**[1], **Jakub Gałka**[1]
[1]AGH University of Science and Technology, Department of Electronics, Krakow, Poland
[2]Jagiellonian University, Department of Computational Linguistics, Krakow, Poland
{kwrobel, jgalka}@agh.edu.pl

Abstract

This article presents the development of the first part-of-speech (POS) tagger for Polish Sign Language (PJM). Due to the lack of PJM corpora, a data set consisting of 34.5 thousand sentences was automatically created and annotated. It was done using a machine translation (MT) system, from Polish to PJM. The annotation with POS tags is done concurrently by transferring and mapping them from Polish. The POS tagger is trained using a sequence classifier and tested on a manually-developed PJM corpus. The results are compared to other taggers for various languages, and error analysis is performed. This paper shows that it is possible to develop a POS tagger with promising results using a transfer-based MT system. The created PJM corpus will be publicly shared.

Keywords: part-of-speech tagger, Polish Sign Language

1. Introduction

Polish Sign Language (PJM) is still not well explored. The analysis of PJM is an especially difficult task, primarily due to the lack of sufficient corpora (the PJM corpus is currently at an early development stage, it is being developed by the Section for Sign Linguistic of the University of Warsaw). Also, there are no standard natural language processing (NLP) tools available for PJM, such as part-of-speech (POS) taggers or dependency parsers.

The available resources do not correspond to the needs. Recent statistics show that there is around 50-100 thousand users of PJM (Świdziński, 2014), while hearing loss is a common problem that concerns about 850 thousand people in Poland (these estimates do not include people who lost their hearing with advancing age, e.g. cases of presbycusis) (Główny Urząd Statystyczny, 2011). Although the number of users may seem to be extremely large, it should be noted, that PJM experiences currently a renaissance, finds new users, and spreads widely. Deaf constitute a considerable language minority in Poland (Świdziński, 2014). The pursuit to break down the communication barriers between the hearing and hearing-impaired people – also by promoting the design, development, and production of information and communications technologies and systems (Lawson, 2007) – is an important policy.

In this article, we present a POS tagger for PJM. The tagger is a part of WiTKoM (Virtual Sign Language Translator) – an interdisciplinary research project funded by the Polish government and carried out by the AGH University of Science and Technology and VoicePIN.com LLC, which aims to create a PJM translator. Such translator would be a huge step in connecting the worlds of Deaf and hearing people, and could lead to a promotion of social inclusion. Working on the translator, the tagger was found useful for analysis and application of PJM corpora. POS might be used e.g. as a source of additional information while developing a PJM to Polish machine translation system or as an additional model supporting the visual sequence of glosses recognition system (POS model in automatic speech recognition reduces the error rate by more than 10 percentage points (Pohl and Ziółko, 2013)). It can be also used for developing a dependency parser for PJM.

2. Related Works

Currently no NLP tools are available for PJM. Worldwide, it is hard to find similar attempts of creating NLP tools for Sign Language (SL). In (Östling et al., 2015), having parallel corpora, the authors transferred the POS from Swedish to Swedish Sign Language using the translation and transferring model. Transferring annotation enriches SL, but it can be used only for data in parallel corpora.

We can observe that amongst many of the ongoing SL corpus projects around the world, grammatical category annotation is either not included or is annotated manually (Östling et al., 2015).

Unlike PJM, there are various POS taggers for the Polish language (Waszczuk, 2012; Radziszewski, 2013; Radziszewski and Śniatowski, 2011; Acedański, 2010; Piasecki, 2007). The quality of Polish taggers is comparable to taggers from other languages and achieves more than 90% accuracy, even though Polish is considered to be a far more complex language than many others due to its advanced morphology and variety of cases (Lewandowska-Tomaszczyk et al., 2012). From the linguistic side, there are several publications on PJM grammatical structures. The most in-depth analysis is probably (Czajkowska-Kisil, 2014), which directly attempts to describe the entire variety of the grammar of PJM.

3. Cross-Reference POS in Polish and PJM

The adaptation of categories used in spoken language grammar into the description of visual-spatial languages causes many difficulties. The discussion regarding parts of speech occurring in sign languages and the most effective basis for their extraction (syntactic, semantic, or morphological) is still ongoing (Schwager and Zeshan, 2008; Filipczak, 2014). According to (Czajkowska-Kisil, 2014), the application of a semantic criterion allows only for partial SL signs categorization, whereas morphological (inflectional) categorization is ineffective. Filipczak (2014), working on

the analysis of corpora data from the Section of Sign Linguistics at the University of Warsaw, cast doubt on the possibility of the clear separation the parts of speech of PJM. She also indicates that further works would be beneficial and encourages their development.

It is often the case that the same sign has different grammatical roles, depending on the context. For example, the sign [MOWA] (Eng. [SPEECH]) (noun) is signed the same way as the sign [MÓWIĆ] (Eng. [SPEAK]) (verb), [MÓWIONA] (Eng. [SPOKEN]) (adjective) or [MÓWIĄCY] (Eng. [SPEAKING]) (present participle). However, the distinctive features of SL signs may provide additional information which may lead to assigning the sign as a particular part of speech. A sign may also incorporate the features of a place in sequence of signs. In literature, the differences between verbs and nouns in national sign languages are widely described (eg. (Johnston, 2001; Hunger, 2006; Kimmelman, 2009; Tkachman and Sandler, 2013; Łozińska, 2015) in the case of PJM). Researchers especially point out the movement – manner, repetition, duration, size – and mouthing as features which can differentiate verbs from nouns.

The identification of signs functioning as adjectives is also a challenge. The adjective appears usually after the noun, and, as in spoken language, they can be a part of a predicate. The signs for adverbs are identical with adjectives, and they can also be expressed by mimicry and incorporated into verbs, which are modified by them. The number of SL lexemes acting as pronouns is smaller than in spoken languages. According to the rules of linguistic economy, classifiers act as pronouns. Their infrequent usage is caused by the spatial quantities of sign languages and their syntax rules. The same applies to prepositions. They are also rarely represented and their appearance suggests being borrowed from Polish. The conjunctions are mostly skipped.

In the National Corpus of Polish there are 36 classes of lexemes. These classes, divided into flexemes with morphosyntactic markers, can be viewed in (Lewandowska-Tomaszczyk et al., 2012). On the other hand, only 16 POS are distinguished in PJM. This number stems directly from the nature of PJM, the examples of which were presented earlier. PJM is much simpler in terms of inflection.

POS in PJM	Assigned Polish POS
noun	noun, depreciative form, bound word
pronoun	non-3rd person pronoun, 3rd-person pronoun, pronoun "siebie"
verb	non-past form, future "być", l-participle, impersonal, imperative, infinitive, contemporary adv. participle, anterior adv. participle, gerund, active adj. participle, passive adj. participle, "winien", agglutinate "być"
adjective	adjective, ad-adjectival adjective, post-prepositional adjective, predicative adjective, predicative
adverb	adverb
preposition	preposition
main numeral	main numeral, collective numeral
coordinating conjunction	coordinating conjunction
subordinating conjunction	subordinating conjunction
punctuation	punctuation
past	-
particle-adverb	particle-adverb
abbreviation	abbreviation
interjection	interjection
alien	alien
unknown form	unknown form

Table 1: POS tags in PJM and their assigned counterparts in Polish.

4. Resources and Methods

To develop a statistical POS tagger we need:

- automatic annotation of glosses with features, i.e. possible POS tags
- a training corpus annotated with POS tags
- a sequence classifier

4.1. PJM Annotation With Possible POS

The annotation of glosses with potential POS tags is essential to limit output possibilities of classifiers and therefore increase tagging accuracy. An ideal solution would be a dictionary containing all PJM signs with their potential POS. However, such dictionary does not exist, so it has been developed automatically. Due to the fact that a gloss is usually created using the lemma of a Polish word, a dictionary for the Polish language can be used. PoliMorf (Woliński et al., 2012) is the morphological dictionary for Polish, consisting of more than 6.5 million word forms. Developing a PJM dictionary requires additional resources containing lemmas, which are represented in PJM by the same sign. 121 lemmas were assigned to 55 signs. Additionally, 847 multi-segment words are annotated with all of the possible POS tags of its segments. The extracted dictionary contains 315 thousand glosses and each gloss has 1.94 Polish POS tags on average. As only Polish POS are available in PoliMorf, it is necessary to assign them to PJM POS. In Table 4.1., we present how each of the 36 Polish POS is matched to one of the 16 POS chosen for PJM.

4.2. Annotated Corpus

There is no PJM corpus, especially annotated with POS. Therefore it has been developed using an existing Polish POS tagger and machine translation system.

4.2.1. Machine Translation System

In their previous works, the authors have developed a hybrid-dependency-based MT system that translates Polish sentences into PJM utterances represented by glosses. It is

a translator with manually created translation rules and statistical word ordering trained on 108 sentences. The system showed good translation quality in comparison to similar works (San-Segundo et al., 2012) reporting a 0.68 BiLingual Evaluation Understudy (BLEU) score. The system is awaiting publication with title: Hybrid dependency-based machine translation for the Polish Sign Language. Example sentences of training and test data are presented in the appendix.

It is necessary to add, that glosses, used as a representation of PJM utterances, do not cover all the information that is sent by a signing person. The use of signing space, mimicry, and other features specific for SL are ignored. This approach can be found in different works containing SL, e.g. (San-Segundo et al., 2012). Moreover, dialogue systems work usually by keyword spotting and representation using glosses will be sufficient. Often, the variations in signing and added articulators, such as mimicry, do not change the part-of-speech itself, although they change the sense of the sentence. For example [RZUCAĆ] (Eng. [THROW]) is a verb no matter if the direction of throwing is indicated or not. Glosses are basic and the most essential statements of the SL user, that can be enriched afterwards by signs associated with handshape, movement, etc.

4.2.2. Bilingual Parallel Corpus

In order to create a corpus annotated with PJM POS tags, we chose 34.5 thousand sentences from the 1 million-word subcorpus of the National Corpus of Polish Language (NKJP). It contains non-domain texts, mainly from journals, periodicals, and belles-lettres. The chosen sentences were shorter than 11 glosses to improve the quality of the translated sentences, due to the errors of the machine translation system.

4.2.3. Training Examples

During the translation, Polish sentences are tagged using the Concraft tagger (Waszczuk, 2012), which is necessary to find the dependencies between the words. The predicted tags were used as correct tags. It was also straightforward to align the new tags of the Polish words to PJM glosses because the MT system is transfer-based – it operates on tree nodes, keeping the information regarding the word and its feature transitions in proper nodes.

As showed in Chapter 3., PJM distinguishes less POS than Polish. At this stage, tag-transition rules are applied. A new tag PAST, attributed to glosses, was implemented to express the past tense. It is worth mentioning that the translation and alignment step determines the correct grammatical classes of those signs whose POS is not distinguished without the context of the sentence.

4.3. Classification

As a classifier, we use the Vowpal Wabbit (Langford et al., 2007) – an open-source learning system program. We train the model using the sequential algorithm "learning to search." It has similar accuracy to conditional random fields but has a better speed performance (III et al., 2014). We use 35 thousand examples for 6 passes as training data.

	Manual corpus	Automatic corpus
Number of sentences	108	1000
Average number of glosses in sentence	7.72	6.54
Ambiguous glosses	29.84%	19.46%

Table 2: Statistics of corpora used for testing.

POS tagger	Accuracy all	Accuracy ambiguous
baseline	83.21%	47.24%
presented tagger	93.85%	81.41%
SSL	78.7%	-
TaKIPI (only POS)	97.78%	91.54%

Table 3: Comparison of tagging accuracy of various taggers. Baseline and presented tagger are tested on manual corpus.

5. Evaluation

To evaluate the tagger, a corpus of 108 manually tagged, commonly used, and domain-free sentences was developed. The evaluation was also conducted using one thousand sentences from the NKJP subcorpus, automatically translated by the MT system, as described in Section 4.2.1. The statistics of the corpora are presented in Table 5.. The manual corpus is more complex: sentences are longer and ambiguous glosses are more numerous.

Table 5. presents the results for all words and ambiguous words separately. The ambiguous words have more than one possible POS assigned. As a baseline, a random choice from annotated possibilities was made. We compare the quality of the tagger with an automatically annotated corpus for Swedish Sign Language (SSL) (Östling et al., 2015) and the TaKIPI tagger (Piasecki, 2007) for Polish language.

Tests show that the possible POS tag annotation achieves 98.72% accuracy (the true answer is in the set of possible tags).

The second experiment shows the accuracy in the function of the number of training sentences. Figure 1 presents the accuracy of the tagger tested on a manually annotated corpus. Increasing the size of training data to more than 5 thousand sentences does not change the scores. The reason is that automatic data is not totally valid, and there is no information on how to correctly tag PJM utterances in special cases. However, Figure 2 presents the accuracy of automatically created test data and the score trend is increasing.

The tagger made 49 errors on the manual corpus. 20% of it could be resolved by a PJM dictionary. The remaining errors stem from a lack of correct training data or insufficient classifier power. The most occurring misclassification occurs with the gloss [PRACA] (Eng. [WORK]). In the manual corpus, it is equally represented by nouns and verbs, however in the training corpus it usually acts as a noun. This problem is also caused by the small dictionary of glosses which have more POS assigned to them.

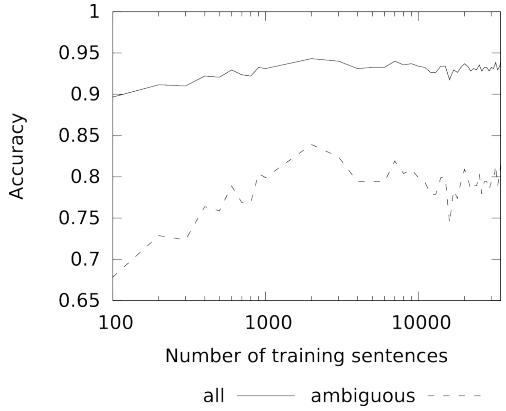

Figure 1: Accuracy of manually annotated test corpus in function of number of training sentences.

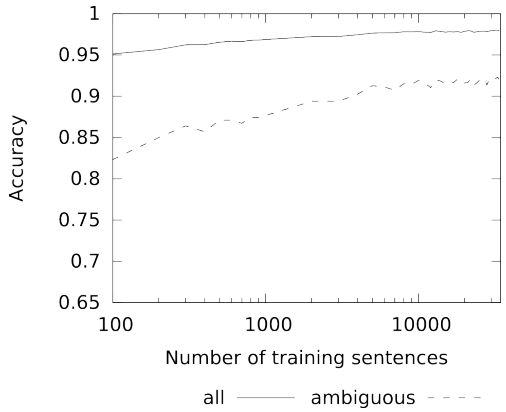

Figure 2: Accuracy of automatically annotated test corpus in function of number of training sentences.

6. Conclusions

The achieved results are promising. A POS tagger can be developed for a new language using an existing MT system and a POS tagger for the source language.

Using artificial corpus as an additional data is proved to improve results in many cases (Abdul-Rauf et al., 2016). Due to the lack of big corpus, creating a high-quality tool using artificial resources was a necessary attempt. Evaluation using real, annotated corpus has shown, that attempt succeeded.

The quality of the tagger can be improved in many ways:

- a better MT system
- an extended or manually-annotated dictionary of PJM-Polish (including multi-segment words and signs with many Polish words assigned)
- a manually annotated PJM corpus

In comparison to the difficulty of Polish language tagging, PJM has less POS tags, but it can not exploit its rich morphology.

Further work will be focused on developing a dependency parser for PJM and using POS tags for a statistical machine translation system.

7. Acknowledgment

This work was supported by the Polish National Centre for Research and Development – Applied Research Program under Grant PBS2/B3/21/2013 tilted *Virtual Sign Language Translator*.

8. Bibliographical References

Abdul-Rauf, S., Schwenk, H., Lambert, P., and Nawaz, M. (2016). Empirical use of information retrieval to build synthetic data for SMT domain adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24:745–754.

Acedański, S. (2010). A morphosyntactic Brill tagger for inflectional languages. In *Advances in Natural Language Processing*, pages 3–14. Springer.

Czajkowska-Kisil, M. (2014). *Opis gramatyczny Polskiego Języka Migowego (The grammatical description of Polish Sign Language)*. Ph.D. thesis, University of Warsaw, Faculty of Polish Studies.

Filipczak, J. (2014). *Czasownik i przestrzeń (Verb and spatial). Lingwistyka przestrzeni i ruchu. Komunikacja migowa a metody korpusowe (Linguistics of space and movement. Sign language communication and corpus methods)*. Zakład Graficzny Uniwersytetu Warszawskiego.

Główny Urząd Statystyczny. (2011). Stan zdrowia ludności Polski w 2009 r (The health status of the Polish population in 2009). *GUS, Warszawa*.

Hunger, B. (2006). Noun/verb pairs in Austrian sign language (ÖGS). *Sign Language & Linguistics*, 9(1-2):71–94.

III, H. D., Langford, J., and Ross, S. (2014). Efficient programmable learning to search. *CoRR*, abs/1406.1837.

Johnston, T. (2001). Nouns and verbs in Australian Sign Language: an open and shut case? *Journal of Deaf Studies and Deaf Education*, 6(4):235–257.

Kimmelman, V. (2009). Parts of speech in Russian Sign Language: The role of iconicity and economy. *Sign Language & Linguistics*, 12(2):161–186.

Langford, J., Li, L., and Strehl, A. (2007). Vowpal wabbit online learning project.

Lawson, A. (2007). The united nations convention on the rights of persons with disabilities: New era or false dawn? *Syracuse J. Int'l L. & Com.*, 34:563–715.

Barbara Lewandowska-Tomaszczyk, et al., editors. (2012). *Narodowy Korpus Języka Polskiego (National Corpus of Polish)*. Wydawnictwo Naukowe PWN.

Łozińska, S. (2015). *Czasownik w polskim języku migowym. Studium semantyczno-gramatyczne (Verb in Polish Sign Language. Study of semantic and grammatical features)*. Ph.D. thesis, University of Warsaw, Faculty of Polish Studies.

Östling, R., Börstell, C., and Wallin, L. (2015). Enriching the Swedish Sign Language Corpus with part of speech

tags using joint bayesian word alignment and annotation transfer. In *20th Nordic Conference on Computational Linguistics (NODALIDA 2015)*, pages 263–268. Linköping University Electronic Press.

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 11(1-2):151–167.

Pohl, A. and Ziółko, B. (2013). Using part of speech n-grams for improving automatic speech recognition of Polish. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 492–504. Springer.

Radziszewski, A. and Śniatowski, T. (2011). A memory-based tagger for Polish. In *Proceedings of the 5th Language & Technology Conference, Poznań*.

Radziszewski, A. (2013). A tiered CRF tagger for Polish. In *Intelligent tools for building a scientific information platform*, pages 215–230. Springer.

San-Segundo, R., Montero, J. M., Córdoba, R., Sama, V., Fernández, F., D'Haro, L., López-Ludeña, V., Sánchez, D., and García, A. (2012). Design, development and field evaluation of a Spanish into sign language translation system. *Pattern Analysis and Applications*, 15(2):203–224.

Schwager, W. and Zeshan, U. (2008). Word classes in sign languages criteria and classifications. *Studies in Language*, 32(3):509–545.

Marek Świdziński, editor. (2014). *Sytuacja osób głuchych w Polsce. Raport zespołu ds. g/Głuchych przy Rzeczniku Praw Obywatelskich (The situation of deaf people in Poland. The report of the committee for the deaf at Commisioner for Human Rights)*. Biuro Rzecznika Praw Obywatelskich.

Tkachman, O. and Sandler, W. (2013). The noun–verb distinction in two young sign languages. *Gesture*, 13(3):253–286.

Waszczuk, J. (2012). Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *COLING*, pages 2789–2804.

Woliński, M., Miłkowski, M., Ogrodniczuk, M., and Przepiórkowski, A. (2012). Polimorf: a (not so) new open morphological dictionary for Polish. In *LREC*, pages 860–864.

Appendix: MT input-output examples

Examples of the input-output of the MT system, which is at the basis of the study:

- Polish: Proszę przynieść jutro rachunki za gaz, wodę i prąd. (Eng. Please bring tomorrow bills for gas, water and electricity.)
  PJM: [JUTRO] [TY] [RACHUNEK] [GAZ] [WODA] [PRĄD] [PRZYNIEŚĆ] [PROSIĆ] [.] (Eng. [TOMORROW] [YOU] [BILL] [GAS] [WATER] [ELECTRICITY] [BRING] [PLEASE] [.])
- Polish: Ty pracujesz legalnie czy na czarno? (Eng. Do you work legally or illegally?)
  PJM: [TY] [PRACA] [BIAŁO] [CZY] [CZARNO] [?] (Eng. [YOU] [WORK] [WHITE] [OR] [BLACK] [?])
- Polish: Rachunek za telefon można zapłacić na poczcie. (Eng. Phone bill can be paid at the post office.)
  PJM: [RACHUNEK] [TELEFON] [POCZTA] [PŁACIĆ] [MÓC] [.] (Eng. [BILL] [PHONE] [POST OFFICE] [PAY] [CAN] [.])
- Polish: W pracy będę chodził na bezpłatny kurs masażu. (Eng. In the work I will attend free massage course.)
  PJM: [JA] [PRACA] [KURS] [MASAŻ] [BEZPŁATNY] [CHODZIĆ] [BĘDZIE] [.] (Eng. [I] [WORK] [COURSE] [MASSAGE] [FREE] [ATTEND] [WILL] [.])
- Polish: Ja interesuję się sportem, lubię oglądać mecze w telewizji. (Eng. I am interested in sports, I like to watch the matches on television.)
  PJM: [JA] [SPORT] [INTERESOWAĆ SIĘ] [,] [TELEWIZJA] [MECZ] [OGLĄDAĆ] [LUBIĆ] [.] (Eng. [I] [SPORT] [INTEREST] [,] [TELEVISION] [MATCH] [WATCH] [LIKE] [.])

Automated Spelling Correction for Dutch Internet Users with Intellectual Disabilities

Leen Sevens, Tom Vanallemeersch, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde
Centre for Computational Linguistics
KU Leuven, Belgium
firstname@ccl.kuleuven.be

Abstract

We present the first version of an automated spelling correction system for Dutch Internet users with Intellectual Disabilities (ID). The normalization of ill-formed messages is an important preprocessing step before any conventional Natural Language Processing (NLP) process can be applied. As such, we describe the effects of automated correction of Dutch ID text within the larger framework of a Text-to-Pictograph translation system. The present study consists of two main parts. First, we thoroughly analyze email messages that have been written by users with cognitive disabilities in order to gain insights on how to develop solutions that are specifically tailored to their needs. We then present a new, generally applicable approach toward context-sensitive spelling correction, based on character-level fuzzy matching techniques. The resulting system shows significant improvements, although further research is still needed.

Keywords: Automated Spelling Correction, Intellectual Disabilities, Pictograph Translation, Alternative and Augmentative Communication

1. Introduction

The Internet has influenced our daily lives in various ways. Being able to stay in touch with family and friends via email or social media websites strengthens the feeling of belonging to a community, even at distances. Therefore, not being able to access or use information technology is a major form of social exclusion. There is a dire need for digital communication interfaces that enable people with Intellectual Disabilities (ID) to contact one another.

We are developing a Text-to-Pictograph and Pictograph-to-Text translation system for the WAI-NOT[1] communication platform. WAI-NOT is a Flemish non-profit organization that gives people with severe communication disabilities the opportunity to familiarize themselves with computers, the Internet, and social media. Their safe website environment offers an email client that makes use of the pictograph translation solutions. The Text-to-Pictograph translation system (Vandeghinste et al., 2015; Sevens et al., 2015a) automatically augments written text with Beta[2] or Sclera[3] pictographs and is primarily conceived to improve the *comprehension* of textual content. The Pictograph-to-Text translation system (Sevens et al., 2015b) allows the user to insert a series of Beta or Sclera pictographs, automatically translating this image sequence into natural language text where possible, hereby facilitating the *construction* of textual content.

The Text-to-Pictograph translation system consists of various sub-processes. During the preprocessing phase, basic spelling correction (see section 5.1.) is applied, as some users have the ability to write short messages without having to rely on the pictograph selection menu. However, these messages often contain severe spelling errors. While it is important to encourage people with ID to write their own messages if they have the ability to do so, the resulting text may pose several problems. First, even if the receivers of the ill-formed messages are (to some extent) able to read written text, they might not be able to understand these messages because of all these mistakes. Secondly, as noted by Sproat et al. (2001), text normalization is recommended before applying a more conventional Natural Language Processing (NLP) process. The Text-to-Pictograph translation tool, which translates the email into pictographs for people who have reading difficulties, may retrieve wrong pictographs or no pictographs at all for erroneously written words. Vandeghinste et al. (2015) evaluated the Text-to-Pictograph translation system and showed that there is clearly room for further improvement in the automated spelling correction process, as the scores for the upper bound (manual spelling correction) were significantly better than the scores for the basic, automated spelling correction process (see Figure 1).

Figure 1: Example of an erroneous Text-to-Beta translation

We present the first version of an automated spelling corrector that is specifically tailored to users with ID. After a discussion of related work (section 2.), we thoroughly analyze tweets and messages sent with the WAI-NOT system and show that users with ID make more and different spelling mistakes than users who do not have cognitive disabilities (section 3.). We then proceed to describe the system architecture. On the one hand, the system consists of a variant generation and filtering step that is partially based on discovering phonetic similarities. On the other hand, we apply character-based fuzzy matching as a novel approach to

[1] http://www.wai-not.be/
[2] https://www.betasymbols.com/
[3] http://www.sclera.be/

context-sensitive spelling correction (section 4.). Our evaluations show that improvements over the baseline in the Text-to-Pictograph translation tool were made (section 5.). Finally, we conclude and describe future work (section 6.).

2. Related work

The rapid dissemination of electronic communication devices has triggered the emergence of new forms of written texts (Kobus et al., 2008). Microtext, or chatspeak-style text, such as tweets or text messages, is characterized by the use of abbreviations, misspellings, phonetic text, colloquial and ungrammatical language, lack of punctuation, and inconsistent capitalization, among other things (De Clercq et al., 2013). Several linguistic models and algorithms have been proposed to deal with errors. We will focus on three popular models for the correction of microtext in particular, as proposed by Kobus et al. (2008): the Noisy Channel or Spell Checking model, the Machine Translation model, and the Speech Recognition model.

The concept behind the *Noisy Channel* model, also called the *Spell Checking* model, is to consider a spelling error as a noisy signal that has been distorted somehow during transmission (Bassil and Alwani, 2012). The Noisy Channel model applies spelling correction on a word-per-word basis and is often limited to the correction of Out of Vocabulary (OOV) words. It relies on orthographic or phonemic surface similarity between two forms. Examples of the Noisy Channel approach for spelling correction are the rule-based system developed by de Neef and Fessard (2007), the system incorporating phonetic information developed by Toutanova and Moore (2002), and the Hidden Markov Model developed by Choudhury et al. (2007), which handles both graphemic variants and phonetic plays. Beaufort et al. (2010) note that the Noisy Channel model places excessive confidence in word boundaries.

The *Machine Translation* (MT) model considers the ill-formed text as the source language, and the correct text as the target language. Aw et al. (2006), for example, use phrase-based MT to tackle the spelling correction problem. It should be noted, though, that it is labor-intensive to construct an annotated corpus to cover ill-formed words and context-appropriate corrections (Han and Baldwin, 2011), especially since the lexical creativity in microtext is difficult to capture. Another issue is the fact that Statistical Machine Translation allows to handle many-to-many correspondences and applies methods to model the possible mismatch in word order (Kobus et al., 2008), while the normalization task is almost deterministic (Beaufort et al., 2010), with no change in word order. De Clercq et al. (2013) implement an MT-based approach and describe the first (and to our knowledge, only) proof-of-concept system for Dutch user-generated content normalization, but they do not consider users with ID.

The *Speech Recognition model* converts the input string into a phone lattice, followed by the creation of a word-based lattice using phoneme-to-grapheme rules, after which a language model is applied and a best-path algorithm is used (Beaufort et al., 2010). An example of this method is presented by Kobus et al. (2008). Han and Baldwin (2011) identify normalization candidates for an OOV word by decoding the pronunciation of all in-vocabulary words and retrieving all words that lie within a threshold character edit distance between the OOV word's pronunciation and the dictionary words' pronunciation.

Our spelling correction system can be considered as a combination of all three approaches, while also introducing new ideas. Although not only OOV words are considered, spelling variants are generated for individual tokens (Noisy Channel model). More specifically, these variants are generated (in the first place) by considering the ill-formed word as a result of phonetic confusion (Speech Recognition model). Finally, we match our new spelling hypotheses against a target language corpus of correctly written text (Machine Translation model). The system does not require large amounts of annotated data.

3. Error distribution: Comparison with tweets

Whenever microtext is considered in the literature, its description is often (if not always) limited to the analysis of SMS messages and tweets. Spelling correction for microtext is a young domain of research, due to the recent boom of social media websites, and its focus lies on users who do not necessarily have a cognitive disability. However, many people with cognitive disabilities resort to specialized communication platforms and apps, such as the WAI-NOT environment. The fact that the spelling correction tool possibly needs to deal with a completely new and different type of microtext should not be ignored. In order to verify this, we compared tweets written by people who supposedly do not have a cognitive disability with emails that were sent with the WAI-NOT email client.

	# OOV	# RWE	# Words	% Errors
WAI-NOT	481	183	8077	8.2%
Tweets	182	88	10964	2.5%

Table 1: Total amount of misspelled tokens. OOV = Out-of-Vocabulary tokens; RWE = Real-word errors

We selected a total of 1000 subsequent tweets from the Dutch Twitter feed, having excluded those messages that were not personal, such as news articles or advertisements. Additionally, a total of 1000 random WAI-NOT emails were selected after having thrown away 49 completely unreadable messages and 330 messages that consisted of pictographs only. We manually corrected all tweets and email messages, while analyzing the different types of errors that were made.[4]

Generally speaking (see Table 1), many more errors can be found in the WAI-NOT messages (8.2%) than in tweets (2.5%). Both OOV words and real-word errors were considered.

As shown in Table 2, the majority (52.1%) of spelling mistakes that are made by people with ID is caused by phonetic confusion, defined here as the orthographic approximation of a word's pronunciation (such as *wiekent* for *weekend*). Although this phenonemon can also be observed in tweets

[4]The corrected tweets and WAI-NOT emails are available on request. The emails may only be used for research purposes.

	Total # misspelled	# PW	% PW
WAI-NOT	664	346	52.1%
Tweets	270	95	35.2%

Table 2: Total amount of misspelled words that are a phonetic approximation of the correct word. PW = Phonetic words

(35.2%), Twitter users' phonetic spellings tend to be much more systematic. They usually concern deliberate mistakes in an attempt to mimic speech (such as the final *t* deletion in *da* or *nie* for *dat "that"* and *niet "not"*), or recurrent grammatical mistakes (such as *jou "you"* versus *jouw "your"* or *gebeurt "happens"* versus *gebeurd "happened"*). Han and Baldwin (2011) note that ill-formedness in regular messages is often intentional, whether due to the desire to save characters or keystrokes, due to the wish to belong to a social group, or due to convention. Phonetic mistakes in WAI-NOT messages are most likely undeliberate mistakes in an attempt to write a correct piece of text, and are therefore much more diverse. This idea is reinforced by the fact that a large part of the analyzed messages were addressed at teachers or caregivers, for whom one might do a deliberate effort.

	LD	# Words	Percentage
WAI-NOT			
	1	479	72.1%
	2	128	19.3%
	3	44	6.6%
	4	9	1.4%
	5	2	0.3%
	6	2	0.3%
Tweets			
	1	166	61.5%
	2	66	24.4%
	3	19	7%
	4	7	2.6%
	5	4	1.5%
	6	4	1.5%
	7	2	0.7%
	8	1	0.4%
	12	1	0.4%

Table 3: Overview of total amount of character operations required per erroneously spelled word. LD = Levenshtein distance

As an additional error measure, we counted the number of insertions, deletions, and substitutions needed to get from the original messages to their corrected counterparts (see Table 3). On the average, messages in WAI-NOT require 1.4 operations per erroneously spelled word, while tweets require 1.7 operations. This difference can be explained as follows. Relatively speaking, Twitter users are more likely to delete characters (75.6% of all required character operations are insertions) than WAI-NOT users (48.6%). This observation is most likely due to the 140-character limit for tweets or the wish to belong to a social group. Examples of deliberate abbreviations in tweets that require many character insertions are *wrschnlk* for *waarschijnlijk "probably"* and *mssch* for *misschien "maybe"*.

	# FL	# PN	# EN	# AB
WAI-NOT	10	0	6	0
Tweets	9	0	72	59

Table 4: Other factors that should be taken into consideration. FL = Flooding; PN = Phonetic numbers; EN = English words; AB = Abbreviations

There are other problems related to spelling errors that may need correction (see Table 4). Flooding, the constant repetition of one character, which occurs when emphasis is given by the user (such as *nooooooo* or *cooooool*), can be found in both genres, while we did not encounter any examples of numbers encoding phonetic values (such as *m8* for *mate* in English). English words were hardly used in the Dutch WAI-NOT messages (with the exception of *I love you*). Abbreviations (such as *m.b.t. "w.r.t."* for *met betrekking tot "with respect to"*) did not occur in these messages at all. Therefore, as long as our system focuses on users with ID, it should not be dealing with foreign language detection or abbreviation solving.

From this analysis, it can be concluded that text written by people with ID is indeed a different kind of microtext. Not only does it contain more errors and phonetic approximations, common abbreviations are lacking, and the users barely use any English words for which a Dutch alternative is available.

Spelling errors made by children who are still learning how to spell and people with Alzheimer's disease might be very similar to text written by people with ID. [5] This hypothesis will have to be tested.

4. System architecture

We describe our prototype version of a spelling corrector that is specifically tailored to Dutch text written by people with ID (see Figure 2). In the first phase, the input text to be corrected undergoes a number of preprocessing steps (section 4.1.). Next, spelling variants are generated for all OOV tokens and infrequent real words (section 4.2.); a trigram language model is used to narrow down the final amount of possibilities. The third step consists of using a character-based fuzzy matching technique for finding the best combination of spelling variants and, additionally, performing new character substitutions when a strong context match is found (section 4.3.). Finally, we describe the parameter tuning process (section 4.4.).

4.1. Preprocessing

The input text undergoes a number of preprocessing steps. The *word builder* module (Vandeghinste, 2002) takes every two adjacent tokens and checks whether they can be put together in order to form a real word. The word builder parameters (see section 4.4.) set different threshold frequencies for the (non-)acceptance of the newly created compound word.

The rule-based *tokenizer* splits the punctuation signs from the words, as the variant generation module works on the token level. Given that the hyphen/dash and the apostrophe

[5] We thank one of the anonymous reviewers for this suggestion.

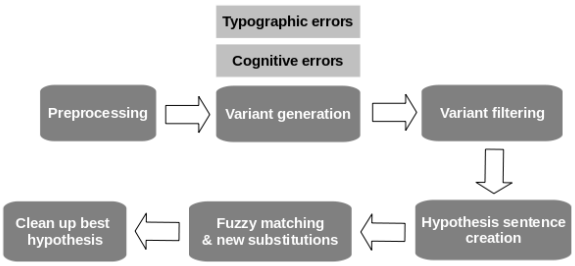

Figure 2: System architecture

often belong to the word, they are not dealt with by the tokenization process.

Although most messages sent by the users only contain one sentence, *sentence detection* is applied. Segmentation is based on full stops.

In the next step, upper-case letters are converted to lower-case letters. Names keep their capital first letter, so they will not be involved in the spelling correction process, as long as the name can be found in a database of first names. [6]

The constant repetition of one character or *flooding* is tackled by reducing any repeated sequence of characters to two characters.

Finally, we created a very small dictionary containing popular greetings (such as *hey*) for tokens that will have to be left out of the correction process.

4.2. Variant generation and filtering

Spelling variants are generated for all OOV tokens and infrequent real words according to a reference corpus. Our variant generation process focuses on phonetic, i.e. cognitive errors (section 4.2.1.). If no variants are generated, the system checks for basic typographic errors (section 4.2.2.). The final amount of variants is narrowed down by a trigram language lookup before proceeding to the next step (section 4.2.3.).

4.2.1. Generating variants for cognitive errors

Cognitive errors occur when the writer does not know how to spell a word, and often rely on the identical pronunciation of words (Toutanova and Moore, 2002). As shown by the error distribution (see Table 2), phonetic confusion causes the majority of spelling errors that are made by the target group.

Building conversion rules The approach described in this section is partially inspired by the finite-state framework for normalizing SMS messages developed by Beaufort et al. (2010).

First, we manually correct 1000 sentences written by WAI-NOT users.

We then align the uncorrected and corrected sentences on the character level, using Levenshtein Distance Alignment. [7] This metric (Levenshtein, 1966) computes the edit distance of two strings by measuring the minimum number of operations (substitutions, insertions, deletions) required to transform one string into the other. Delimiters, such as commas and spaces, are also aligned. Missing characters on either side of the alignment are indicated by inserting a hyphen (-).

In the next step, we create token pairs. A token pair is retrieved when the same delimiter is found at the same location in both the source/uncorrected and target/corrected language character string. From these token pairs, we extract all possible character 4-grams on the source language side and the characters they align with on the target language side. This process is repeated for three, two, and one character(s).[8]

Having obtained all character 4-gram, trigram, bigram, and unigram alignments, probabilities are estimated: For every character n-gram on the source side, we calculate the likelihood of obtaining a particular character sequence on the target side. The sequence obtained is usually identical, but sometimes different. For example, the character trigram "int" on the source/uncorrected language side remained "int" in 91% of the cases, but had been corrected into "ind" in the remainder of the sentences.

For the construction of our final rule set, we retain only those cases where we observe a 1% to 100% probability of changing a particular character n-gram into a different n-gram. The idea behind this is that rarely occurring alternations might actually have a typographic rather than a phonetic origin. By contrast, more commonly occurring mistakes are most likely due to phonetic confusion. This idea is also reflected in the final version of our inventory. For instance, the written sequences "int" and "ind", "pra" and "praa", "orie" and "orry" can indeed be pronounced the same way.

This inventory of commonly appearing alternations allows us to build a system of character rewrite rules, in which character 4-gram rules overrule trigram rules, trigram rules overrule bigram rules, and so on.

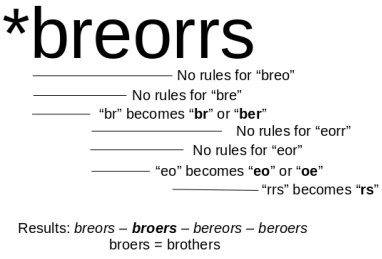

Figure 3: Example of how the conversion rules are applied

Applying conversion rules For every non-word and every real word that has a frequency lower than the *real word minimum frequency threshold* (see section 4.4.) in our frequency list,[9] the conversion rules are applied (see Figure 3).

[6] http://www.quietaffiliate.com/free-first-name-and-last-name-databases-csv-and-sql/

[7] http://rosettacode.org/wiki/Levenshtein_distance/Alignment

[8] In future work, we will evaluate whether five- or six-character pairs make the variant generation process more robust.

[9] The frequency list contains roughly eighty million words of Belgian Dutch newspaper text.

A four-character window slides over the token, starting with the first four characters of the token, and checks if a 4-gram rule can be found for that particular sequence. If a rule is found, all the conversion outputs (including the original sequence) are stored and the system will proceed to check the next four characters of the token. If no rule is found, the system backs off to the first three characters of the token and attempts to find a trigram rule for that sequence. If, even at the one-character level, no rules could be found, the original character is retained and the system proceeds to find rules for the next four-character sequence.

In the end, all conversion outputs are concatenated and both non-words and real words may have been formed. If a real word is formed with a frequency higher than the *variant frequency* (see section 4.4.), it will be retained as a variant for that token.

Note that our approach, although phonetically similar variants are generated, does not yet decode the pronunciation of words into phonemes.

4.2.2. Generating variants for typographic errors

Typographic errors are mostly related to the keyboard (Toutanova and Moore, 2002). If in the previous step no variants were generated for a non-word or a word that has a frequency lower than the *real word minimum frequency threshold* (see section 4.4.), we apply basic typographic error correction principles. We generate variants based on five different operations.

The first operation is the word splitting module. This is an insertion module for one space character. The system checks whether the erroneous or infrequent token can be split into two parts at any position. Frequency thresholds are determined by parameters (see section 4.4.).

The next operations are one-character deletion, insertion, substitution, or adjacent transposition at every position of the token. If a real word is formed with a frequency higher than the *variant frequency* (see section 4.4.), it will be retained as a variant for that token. If the original token was a real word, then that word will always be retained as one of the variants.

4.2.3. Filtering the variants

It is possible that, at this point, the system has generated multiple variants for a single erroneous word or infrequent real word. Especially when considering very short words, many real-word alternatives can often be created. The filter module narrows down the total amount of possibilities before proceeding to the next step. A trigram language model trained on a very large corpus (a combination of the Dutch part of Europarl, Corpus Gesproken Nederlands, Cross-Language Evaluation Forum, DGT-Translation Memory, and Wikipedia) is used to check whether the variant ever occurs within its context in the language model. As the token's direct context may also contain variants, all possible trigrams are checked until a match is found. If a match is found, the variant will be retained. If no trigram matches are found for any of the variants because the context does not provide enough information, all variants are retained.

4.3. Character-based fuzzy matching

The combination of all variants described above leads to the creation of a number of potentially correct sentences. Each one of these sentences is a hypothesis, one of which will receive the highest score through fuzzy matching. Fuzzy matching techniques, which have been developed in the context of translation memories (databases with source sentences and their translations used by professional translators), allow to find strings in a corpus that approximately (rather than exactly) match a string. We are applying this technique to a monolingual corpus. In the development of the spelling correction tool, we explored the new possibility of applying fuzzy matching techniques at the character level.

Each one of the hypotheses is split into individual characters. The space is replaced by a dummy character, the % sign, and should also be recognized as a character. As our corpus, we use the Spoken Dutch Corpus (Corpus Gesproken Nederlands, (Oostdijk et al., 2002)) since spoken language better reflects the language used in user-generated content (De Clercq et al., 2013). This corpus is also split into individual characters. During the fuzzy matching process, we use a filter called *approximate query coverage* (Vanallemeersch and Vandeghinste, 2015). Its purpose is to select candidate sentences in a corpus which are likely to reach a minimal matching threshold when submitting them to a fuzzy matching metric, in order to increase the speed of matching. Candidate sentences share one or more n-grams of a minimal length with the input hypothesis, and share enough n-grams with the input hypothesis to cover the latter sufficiently (according to some threshold). In our spelling correction model, the unigram is one single character. A very efficient search for sentences sharing n-grams with the input hypothesis can be done by means of a suffix array (Manber and Myers, 1993).[10]

A hypothesis that shares many and long character n-grams with candidate sentences from the corpus has a bigger likelihood of becoming the winning hypothesis than one that shares only few and short n-grams.[11]

The context-sensitivity of the fuzzy matching method allows us to deal with additional spelling errors, even if the correct variant had not been generated in the variant generation phase. If a high-scoring corpus match is found for two strings of characters and there is a gap of maximum three characters between those strings in both the corpus and the original hypothesis, those characters will be replaced in the hypothesis. For example, the hypothesis *kan je dat misschien nog aan jou moeder vragen "maybe you can ask your mother"* contains a common spelling mistake in Dutch. *Jou* is a personal pronoun, while *jouw* is a possessive pronoun. *Jouw* would be correct here. However, no variants were generated for *jou* in the variant generation process, as it is a highly frequent word. One of the matching strings in the corpus is *nu moeten we het nog aan jouw moeder vragen "now we still need to ask your mother"*. Looking at this

[10] We used the SALM toolkit (Zhang and Vogel, 2006) for building and consulting suffix arrays.

[11] For sake of brevity, we refer to Vanallemeersch and Vandeghinste (2015).

sentence and the hypothesis, there is a character overlap between *% n o g % a a n % j o u* and *% m o e d e r % v r a g e n*. The system finds the character *w* in the corpus, surrounded by the two substrings (a one-character gap). This character is inserted in the original sentence, hereby fixing the spelling mistake. We will include the maximal gap width as one of the parameters in future work.
The winning hypothesis is cleaned up. The spaces between the characters are removed, the % signs are converted into spaces and the first letter is capitalized.

4.4. System parameters

The system contains a number of parameters, which were tuned.
There are two *word builder penalties*. The first penalty concerns the frequency of the separate parts of the (potentially in-vocabulary) compound token. If both parts are real words and their frequency is high enough (post-tuning value: 120), they won't have to pass through the word builder module. The other penalty is the minimum frequency required to accept a newly built real word (post-tuning value: 210).
Similarly, there are two *word splitter penalties*. The first penalty concerns the minimum frequency of the token. If the frequency of this token is high enough (post-tuning value: 1760), it will not have to pass through the word splitter module. However, if the frequency is not high enough or if the token turns out to be a non-word, the system will attempt to split the token into two real word parts. The second penalty sets a minimum frequency for those two words (post-tuning value: 1680). If the frequency is high enough, the original word will be split.
The *real word minimum frequency threshold* determines how common a correctly spelled word should be in order to avoid going through the spelling variant generation process (post-tuning value: 100).
When real word variants are generated for a token, they need to have a minimum frequency, the *variant frequency*, in order to be accepted as a variant (post-tuning value: 220).
There are also three fuzzy matching penalties. The *n-gram penalty* decides on the minimum amount of contiguous characters that should occur as a sequence in the corpus sentence (post-tuning value: 8). The *minimum score penalty* sets the minimum matching score needed to retrieve a corpus sentence (post-tuning value: 0.2). Finally, the *highest frequency threshold* decides that, if a certain n-gram has a very high frequency, the fuzzy matching system will ignore it for fuzzy matching, for reasons of speed (post-tuning value: 100).
We created a tuning corpus by manually correcting 200 new WAI-NOT messages. We used the local hill climber algorithm as described in Vandeghinste et al. (2015), which varies the parameter values when running the spelling corrector script on the test set. The BLEU metric (Papineni et al., 2002) was used as an indicator of relative improvement. BLEU is a precision-oriented metric which compares the system output to one or more reference translations, by counting how many n-grams overlap, and correcting for brevity. We ran five trials of a local hill climbing algorithm. We did this until BLEU converged onto a fixed score after several thousands of iterations. Each trial was run with random initialization values, and varied the values between certain boundaries in order to cover different areas of the search space. From these trials, we took the best scoring parameter values.

	BLEU	NIST	WER	# CO
No corrector	0.64	8.62	12.37	699
Old corrector	0.62	8.07	19.51	816
New corrector	0.84	10.49	7.57	238

Table 5: Automated evaluations on 300 email messages. CO = Number of character operations

	Old	New
# Justified corrections of erroneous words	41	145
# Unjustified corrections of erroneous words	74	29
# Non-corrected erroneous words (# Real)	157 (76)	98 (70)
# Inappropriate changes to correct words	16	0

Table 6: Analysis of how the systems deal with erroneous words

5. Evaluation

We present the results of our evaluations. Section 5.1. evaluates the system on an unseen test set of WAI-NOT messages and compares its corrections with the corrections made by the system that was originally developed for the Text-to-Pictograph translation tool. Section 5.2. evaluates the system within the context of the Text-to-Pictograph translation pipeline.

5.1. Intrinsic evaluation

After having filtered unreadable messages and messages that consisted of pictographs only, we took 300 random emails from the WAI-NOT corpus and manually corrected them to the best of our ability. Our baseline is the original set of uncorrected messages. We also show the result of applying the spelling correction system that was originally developed for the Text-to-Pictograph translation tool. The original system applies one-character substitutions, deletions, and insertions to generate a list of variants and selects the most frequent variant according to the frequency list (see section 4.2.). This context-insensitive approach is compared to the output generated by the new system.
Table 5 shows the word-based BLEU, NIST (Doddington, 2002), and Word Error Rate (WER) scores. NIST is similar to BLEU, but gives less credit to high frequency non-informative n-grams. WER counts the number of words that are incorrect with respect to the reference translation(s) and is very well suited for the evaluation of NLP tasks where the input and output strings are closely related. We also automatically calculated the amount of character operations needed in order to get to the reference correction.
As shown in Table 5, the original spell checker does more things wrong than right. However, significant improvements were made using the new spell checker.
Table 6 presents an analysis of how both the original and the new system deal with erroneous words in the email messages. The new system is able to detect more erroneous forms (as the original system was limited to OOV

		With proper names		Without proper names	
Condition	**Precision**	**Recall**	**F-Score**	**Recall**	**F-Score**
Sclera					
Baseline	89.2%	86.2%	87.7%	85.2%	87.2%
New system	92.6%	89.1%	90.8%	88.2%	90.3%
Rel.improv.	*3.7%*	*3.3%*	*3.5%*	*3.6%*	*3.6%*
Beta					
Baseline	85.9%	89.5%	87.6%	88.7%	87.3%
New system	89.8%	91.5%	90.6%	90.8%	90.3%
Rel. improv.	*4.5%*	*2.3%*	*3.4%*	*2.4%*	*3.5%*

Table 7: Manual evaluation of the Text2Picto translation engine

errors) and finds the appropriate correction for 83.3% of these words, while the old system only manages to correct 35.6% of the detected words. 71.4% of the unretrieved words in the new system are highly frequent real words. Many of these real-word errors can be contributed to grammatical confusion, such as the difference between *jou* and *jouw* (see section 4.3.). These errors lead us into the domain of grammar correction and are currently beyond the scope of our work.

The old system corrects some words that should not have been corrected in the first place. These erroneous corrections mostly concern common greetings and proper names that are not included in our list of first names and for which a low-frequency variant is generated. These problems are solved in the new system by the introduction of the small greetings dictionary and the fact that low-frequency variants will not be proposed for an unknown name.

Comparing our system with other systems is difficult, as they do not consider text written by users with ID (and most tools focus on English text). De Clercq et al. (2013), who created the first and only normalization tool for Dutch microtext, admit that words requiring different types of operations are difficult for their system, while our approach allows for multiple (phonetic) substitutions within a single word. De Clercq et al. showed that their system is best at resolving smaller words requiring only one or two insertions, while especially phonetic problems turned out to be an important obstacle for them. As our system is made for users with ID, phonetic alternations are the core of the variant generation process.

5.2. Extrinsic evaluation

We also manually evaluated the effects of the spelling correction system within the larger context of the Text-to-Pictograph translation tool. The baseline, which uses the old spelling corrector, is the system as described by Vandeghinste et al. (2015). The new system implements the spelling corrector as presented in this paper. We used the same systematic and objective approach to manual evaluation as Vandeghinste et al. (2015).

The evaluation set of 50 Dutch messages that have been sent with the WAI-NOT email system consists of 84 sentences (980 words). These were all translated into a sequence of Sclera or Beta pictographs using the Text-to-Pictograph translation tool. [12]

[12]Our gold standards are made available on request.

We have performed a manual annotation with one judge, who removed untranslated words that were considered not to contribute to the content. This allowed calculating the recall. For each of the translated words, she judged whether the pictograph generated was the correct pictograph, in order to calculate precision. Results are presented in Table 7. As proper names occur frequently in e-mail messages, we have calculated recall and F-score with and without proper names, in the latter case removing all proper names from the output. In the case where proper names are included, they are not converted into pictographs. Precision remains the same in both cases. In the WAI-NOT environment, proper names occurring in the contact lists of the users are converted into the pictures attached to these profiles, resulting in more personalized messages.

Figure 4: Example of a correct Text-to-Beta translation

An increase in precision and recall was obtained for both the Beta and the Sclera condition. Examples of erroneous words that previously could not be translated into pictographs are *grapeg* for *grappig "funny"*, *ikhoop* for *ik hoop "I hope"* and *heeeeel* for *heel "very"*. Examples of erroneous words that previously led to an erroneous pictograph translation were *wiekent* for *weekend "weekend"*, which was corrected into *wieken "wings"* (and translated into a pictograph showing a bird's wings, see Figure 1), and *moelijke* for *moeilijke "difficult"*, which was corrected into *mogelijke "possible"* (and translated into a pictograph showing the verb "can"). The new spelling corrector has managed to tackle these issues.

6. Conclusion and future work

We described the first version of an automated spelling corrector for Dutch text written by people with ID. The system can be extended to other languages, provided some corrected data is available in order to infer new phonetic rules for the variant generation step. Nevertheless, the current approach does not require massive amounts of training data.

The results show that the system already improves over the baseline, but there is ample room for enhancement.

In the first place, the variant generation process is not yet able to correct tokens in which both elements of phonetic confusion and typographic errors are present. While these are currently two completely unrelated steps within the variant generation process, the ideal scenario would be to find an efficient way to combine them without overgenerating. Additionally, phone lattices should be introduced in order to go deeper than purely orthographic variation patterns.

For the fuzzy matching step, we will add more and/or different corpora to the corpus and evaluate their influence on the system's performance. These corpora should be exempt from spelling errors and share as many characteristics with informal text or oral conversations as possible. The corpora should contain plenty of first-person and second-person forms.

Finally, we should consider performing a grammar check during the spelling correction process in order to detect real-word errors that are left out of the variant generation process because of their high frequency.

7. Acknowledgements

We would like to thank IWT and the European Commission's Competitiveness and Innovation Programme for funding Leen Sevens' doctoral research and the Able-To-Include project, which allows further development and valorisation of the pictograph translation tools.

We also thank the people from WAI-NOT for their valuable feedback and the integration of our tools on their website.

8. Bibliography

Aw, A., Zhang, M., Xiao, J., and Su, J. (2006). A phrase-based statistical model for SMS text normalization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 33–40, Sydney, Australia. Association for Computational Linguistics.

Bassil, Y. and Alwani, M. (2012). Context-sensitive Spelling Correction Using Google Web 1T 5-Gram Information. *Computer and Information Science*, 5(3).

Beaufort, R., Roekhaut, S., Cougnon, L.-A., and Fairon, C. (2010). A Hybrid Rule/Model-based Finite-state Framework for Normalizing SMS Messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 770–779, Uppsala, Sweden. Association for Computational Linguistics.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., and Basu, A. (2007). Investigation and Modeling of the Structure of Texting Language. *International Journal of Document Analysis and Retrieval: Special Issue on Analytics of Noisy Text*, 10:157–174.

De Clercq, O., Schulz, S., Desmet, B., Lefever, E., and Hoste, V. (2013). Normalization of Dutch User-Generated Content. In *Proceedings of the 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, pages 179–188, Hissar, Bulgaria.

de Neef, E. G. and Fessard, S. (2007). Evaluation d'un système de transcription de SMS. In *Actes du 26e Colloque international Lexique Grammaire*, Bonifacio, France.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Diego, California, USA.

Han, B. and Baldwin, T. (2011). Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (ACL-HLT 2011)*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Kobus, C., Yvon, F., and Damnati, G. (2008). Normalizing SMS: Are Two Metaphors Better Than One? In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 (COLING '08)*, pages 441–448, Manchester, UK. Association for Computational Linguistics.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10.

Manber, U. and Myers, G. (1993). Suffix Arrays: A New Method for On-line String Searches. *SIAM Journal on Computing*, 22:935–948.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, PA, USA. Association for Computational Linguistics.

Sevens, L., Vandeghinste, V., Schuurman, I., and Van Eynde, F. (2015a). Extending a Dutch Text-to-Pictograph Converter to English and Spanish. In *Proceedings of 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2015)*, Dresden, Germany.

Sevens, L., Vandeghinste, V., Schuurman, I., and Van Eynde, F. (2015b). Natural Language Generation from Pictographs. In *Proceedings of 15th European Workshop on Natural Language Generation (ENLG 2015)*, pages 71–75, Brighton, UK. Association for Computational Linguistics.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., and Richards, C. (2001). Normalization of Non-standard Words. *Computer Speech and Language*, 15(3):287–333.

Toutanova, K. and Moore, R. C. (2002). Pronunciation Modeling for Improved Spelling Correction. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 144–151, Philadelphia, PA, USA. Association for Computational Linguistics.

Vanallemeersch, T. and Vandeghinste, V. (2015). Assessing Linguistically Aware Fuzzy Matching in Translation Memories. In *Proceedings of the 18th Annual Confer-*

ence of the European Association for Machine Translation (EAMT 2015), Antalya, Turkey.

Vandeghinste, V., Schuurman, I., Sevens, L., and Van Eynde, F. (2015). Translating Text into Pictographs. *Natural Language Engineering*, pages 1–28.

Vandeghinste, V. (2002). Lexicon Optimization: Maximizing Lexical Coverage in Speech Recognition through Automated Compounding. In *Proceedings of the Third International Confererence on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.

9. Language Resource references

Oostdijk, N., Goedertier, W., Eynde, F. V., Boves, L., Martens, J.-P., Moortgat, M., and Baayen, H. (2002). Experiences from the Spoken Dutch Corpus Project. In *the Third International Confererence on Language Resources and Evaluation (LREC 2002)*, pages 340–347, Las Palmas, Spain.

Zhang, Y. and Vogel, S. (2006). Suffix Array and Its Applications in Empirical Natural Language Processing. Technical report, Carnegie Mellon University, Pittsburgh, PA, USA.

Predicting Reading Difficulty for Readers with Autism Spectrum Disorder

Victoria Yaneva*, Richard Evans*, and Irina Temnikova**
*Research Institute in Information and Language Processing, University of Wolverhampton, UK
**Qatar Computing Research Institute, HBKU, Doha, Qatar
v.yaneva@wlv.ac.uk, r.j.evans@wlv.ac.uk, itemnikova@qf.org.qa

Abstract

People with autism experience various reading comprehension difficulties, which is one explanation for the early school dropout, reduced academic achievement and lower levels of employment in this population. To overcome this issue, content developers who want to make their textbooks, websites or social media accessible to people with autism (and thus for every other user) but who are not necessarily experts in autism, can benefit from tools which are easy to use, which can assess the accessibility of their content, and which are sensitive to the difficulties that autistic people might have when processing texts/websites. In this paper we present a preliminary machine learning readability model for English developed specifically for the needs of adults with autism. We evaluate the model on the ASD corpus, which has been developed specifically for this task and is, so far, the only corpus for which readability for people with autism has been evaluated. The results show that out model outperforms the baseline, which is the widely-used Flesch-Kincaid Grade Level formula.

Keywords: readability, accessibility, autism, automatic text classification

1. Introduction

This paper focuses on the development and evaluation of the first readability model derived by machine learning that is developed specifically for the needs of people with high-functioning autism. Autism Spectrum Disorder (ASD) is a congenital lifelong condition of neural origin, which affects the ability of a person to communicate and interact socially (American Psychiatric Association, 2013).

1.1. Autism Spectrum Disorder

Some people with autism who are less able may remain non-verbal or may develop intellectual disability. People at the higher ends of the spectrum are referred to as high-functioning and are able to produce and comprehend language well, with the exception of certain linguistic constructions and a relative inability to use context and obtain a coherent representation of the text meaning (Happé and Frith, 2006; Frith and Snowling, 1983). At the lexico-semantic level, areas of particular difficulty may include long and unfamiliar words, abstract words and polysemous words, with some autistic people showing deficits in the ability to use context in order to disambiguate homographs (Happé, 1997). At the syntactic level, difficulties include the processing of long sentences containing many clauses, complex punctuation, negation and passive voice, among others (O'Connor and Klein, 2004; Martos et al., 2013). Finally, at the discourse level, readers with autism have been shown to have difficulties grasping the gist of the content of a text as a whole, and difficulties understanding irony, sarcasm, metaphor and authors' intentions (Whyte et al., 2014).

Currently, 1 in 100 people are diagnosed with autism in the UK (Brugha et al., 2012), and it is believed that there are two undiagnosed cases for each one diagnosed (Baron-Cohen et al., 2009). Autism prevalence is expected to increase even more due to recent broadening of the diagnostic criteria and increasing understanding of the characteristics of autism, especially within females. Deficits in reading comprehension are indicated to be one of the reasons for reduced academic achievement and increased school dropout within this population (Brugha et al., 2007).

1.2. Autism and Social Inclusion

Enabling content developers of textbooks, websites, webpages and social media to make their content autism-accessible has the potential to enhance the independence and wellbeing of people with autism, as well as to reduce the resources needed for staff members to support autistic service users in finding relevant information about job accessibility, benefits, disability rights, healthcare, etc. The aim of the readability model presented in this paper is to provide autistic individuals, their tutors, and their carers with an easy way to filter information and to find texts to read that are accessible. The readability model will also provide content developers of websites, textbooks, newspapers, and other media, with an inexpensive, quick, and reliable tool to test the accessibility of their material.

While reading comprehension deficits affect school performance and Web searching behavior, they become an even greater barrier when it comes to using social media such as Twitter, Facebook, WhatsApp, Pinterest, etc. In these environments users need to quickly comprehend written text while chatting and are also exposed to a lot of visual content, which has been shown to affect concentration and comprehension in people with autism (Yaneva et al., 2015). At the same time, social media and the Web are particularly important to people with disabilities because these channels empower them to build an identity in which their disability is not at the forefront, a situation quite different from that in face-to-face communication (especially in the cases of motor, visual or hearing impairments). Communication via social media and the Web allows people with a wide range of disabilities to connect with other people without the complexities of real-world social interactions which have been shown to be especially relevant for those with autism (Bosseler and Massaro, 2003; Putnam and Chong, 2008).

Evidence for the demand of people with autism for accessible and safe social media includes the development of

platforms such as the UK-wide *Autism Connect*[1], in which autistic users can connect to each other in a moderated environment. Accessibility features in these social media include both their simplified design and also their provision of easy-to-read explanations of how to use and navigate the platform. One example of such explanation is the following:

Account - an account is a record of your details. Every user has an account that they have to log in to. The account remembers the things you do and the things that other people have said and done in reply to you [2].

Such accessibility features implemented in disability-friendly social media show how crucial accessible writing is for this population of users.

1.3. Aim of This Study

Two ways in which the demand for accessible writing is currently addressed in English are the *Plain English campaign*[3] and the *Easy-to-read campaign* (Tronbacke, 1997), in which writers follow a set of guidelines to make their text easy to comprehend. In cases where this content is targeted particularly to people with cognitive disabilities, the common practice is to evaluate its complexity via consultations with focus groups of target users, which can be time-consuming and expensive and may also require more than one round of rewriting and evaluation. We aim to address this problem by developing an autism-specific readability assessment model, which can evaluate the accessibility of text content before it has been brought to a focus group for evaluation. The model can also be applied in cases where such groups are not available or the text content is too large to be properly evaluated by humans.
Improving accessibility for users with a certain type of disability may also be of benefit to people with other conditions. This is why, in addition to developing a classifier to distinguish between easy and difficult texts for people with autism, we evaluate the generalizability of this model on a dataset of easy and difficult texts evaluated by people with Mild Intellectual Disability (MID). The main contributions of this research are as follows:

- Development and evaluation of a readability model specifically for people with high-functioning autism
- Development of the ASD corpus, which is a set of reading passages, the complexity of which has been assessed by autistic adults through reading comprehension experiments
- Investigation of the model generalizability on the LocalNews corpus (Feng et al., 2009), containing texts whose complexity has been assessed by readers with mild intellectual disability

[1] https://www.autism-connect.org.uk/
[2] https://www.autism-connect.org.uk/index.php/site/siteuse
[3] http://www.plainenglish.co.uk/

To the best of our knowledge, this is the first research to propose a machine-learning based readability classifier for people with autism and is the first study to evaluate an autism-specific readability metric on text passages assessed by autistic users. Furthermore, this classifier is especially relevant to the assessment of Web text content, as the sets of texts used in both training and evaluation were obtained from Web sources.
The rest of this paper is structured as follows. Section 2 discusses related work on readability assessment. Section 3 describes the corpora used for the development of the classifier, including the user-evaluated text passages whose readability was measured in experiments with the participation of autistic readers, and Section 4 presents the linguistic features, specifically matched to the reading difficulties of this population. The training and evaluation of the classifier are presented in Section 5, while Section 6 discusses the implications of this research to the field of accessibility research. The main conclusions and avenues for future work are summarized in Section 7.

2. Related Work

2.1. Readability Assessment

Readability has been defined as the ease of comprehension because of the style of writing (Harris and Hodges, 1995). Other definitions such as the ones by (Pikulski, 1995) add that readability is a construct that takes into account the relationship between specific reader populations, specific texts, and the purpose of reading. Investigations into what makes a text readable and the endeavor to find formal expressions by which to measure it, namely the readability formulae, date as far back as the end of the 19th century (Dubay, 2004) and gained a lot of popularity during the '40s and '50s of the 20th century with the growth of the publishing business. Readability formulae are equations which typically exploit surface features of the text such as word length and sentence length, aiming to predict the difficulty of a text. The most popular readability formulae are the Flesch Reading Ease formula (Flesch, 1948), Flesch-Kincaid Grade Level (Kincaid et al., 1975), Army's Readability Index (ARI) (Senter and Smith, 1967), the Fog Index (Gunning, 1952), the Simple Measure of Gobbledygook (SMOG) (McLaughlin, 1969), etc.
Readability formulae have been criticized for not taking into account features related to the background knowledge and cultural bias of the reader, the way ideas are organized and connected within the text, and the amount of memory and cognitive load imposed by the text on the reader (Benjamin, 2012; Siddharthan, 2006; Dubay, 2004). For instance, while word length has been shown to correlate closely with the lexical difficulty of texts as perceived by their readers (Gunning, 1952), it has been pointed out that this measure does not take into account how abstract or concrete the words of the text are or whether they are truly familiar to readers of a certain age and background. To address these drawbacks, cognitive scientists have developed cognitively-motivated features, which were proposed on the basis of human rankings and which aim to account for the familiarity and age of acquisition of common words, as well as their levels of abstractness, concreteness, imaga-

bility and meaningfulness (Coltheart, 1981). The majority of these and other cognitively-based lexical features have been computed for a total of 98 538 words and are contained in the MRC Psycholinguistic Database (Coltheart, 1981). These and other cognitively-motivated features such as features of cohesion are implemented in the readability assessment tool Coh-Metrix (McNamara et al., 2010).

Finally, advances in the fields of Natural Language Processing and Artificial Intelligence enable both faster computation of existing statistical features and the development of new NLP-enhanced features which can be used in more complex methods of assessment based on machine learning. This makes large-scale readability assessment feasible and allows customization of the assessment models to specific text content and readership. Examples of this are the unigram models, which have been found particularly suitable for assessment of Web content (Si and Callan, 2001), where the presence of links, email addresses and other elements biases the traditional formulae. Readability assessment for people with different types of cognitive disability has also been investigated and is discussed in the next Section 2.2.

2.2. Readability Assessment for People with Cognitive Disabilities

Individuals with mild intellectual disability have been found to have smaller working memory capacity, resulting in difficulty remembering within- and between-sentence relations (Jansche et al., 2010). Specific readability features developed to capture the characteristics of this particular reader population include entity density (counts of entities such as persons, locations and organisations per sentence) and lexical chains (synonymy or hyponymy relations between nouns) (Jansche et al., 2010; Feng, 2009; Huenerfauth et al., 2009). Evidence from eye-tracking experiments and comprehension questions conducted with Spanish readers with dyslexia, suggests that lexical features such as word length or word frequency are more relevant to people with dyslexia, who do not experience difficulties integrating information from the text but instead struggle with decoding particular letter and syllable combinations (Rello et al., 2012a; Rello et al., 2012b).

Due to the lack of corpora whose reading difficulty levels have been evaluated by people with autism, most readability research for this population has so far been focusing on texts simplified by experts using features matched to reflect the reading difficulties of people with autism (Martos et al., 2013; Štajner et al., 2014; Štajner et al., 2012). User-evaluated texts were used for the first time in an earlier study, where the discriminatory power of a number of features was evaluated on a preliminary dataset of 16 texts considered easy or difficult to comprehend by people with autism, based on reading comprehension experiments (Yaneva and Evans, 2015). The results indicated that 6 features had a high discriminatory power:

1. the number of words per sentence,
2. the number of metaphors per text,
3. the average number of words occurring before the main verb in a sentence,
4. the similarity of syntactic structures of adjacent sentences,
5. the Flesch-Kincaid Grade Level, and
6. the Automated Readability Index.

The current experiment builds upon this work by (1) expanding the set of user-evaluated texts and (2) optimizing combinations of features to distinguish between two classes of difficulty by means of a machine learning algorithm. The process of evaluating the reading passages with people with autism, and the rest of the corpora used for building and evaluating the readability classifier are presented in Section 3.

3. Corpora

The main problem when discussing corpora with respect to training readability classifiers for people with cognitive disabilities is that there is a lack of corpora large enough to be used as a training set. In previous research on readability for people with mild cognitive disability, this issue was addressed by training the classifier on a general corpus with 5 readability levels (The Weekly Reader) and then evaluating it on the LocalNews corpus, a small set of 11 difficult and 11 easy user-evaluated texts (Feng et al., 2009). We propose a similar set-up in which our classifier is trained on the WeeBit corpus (Vajjala and Meurers, 2012), which is a comparatively large corpus consisting of material for schoolchildren of different ages (Section 3.1). After that we evaluate the generalizability of the model on a smaller set of 27 text passages whose difficulty was assessed by adult readers with autism (Section 3.2.1) and on the LocalNews corpus (Feng et al., 2009) (Section 3.2.2), which contains 11 original and 11 simplified versions of newspaper articles whose complexity has been evaluated on readers with mild intellectual disability.

3.1. Training and Intrinsic Evaluation

The WeeBit corpus (Vajjala and Meurers, 2012) comprises two sub-corpora, The Weekly Reader[4] and BBC-BiteSize[5], obtained from educational websites of the same names. The Weekly Reader is an educational web-newspaper with articles from the domains of fiction, news and science intended for children of ages 7-8 (Level 2), 8-9 (Level 3), 9-10 (Level 4) and 9-12 (Senior level). BBC-BiteSize contains articles at 4 levels corresponding to educational key stages (KS) for children between ages 5-7 (KS1), 7-11 (KS2), 11-14 (KS3) and 14-16 (GCSE). After removing audio files and non-textual information (including all of KS1, as it consists mostly of images), the combined WeeBit corpus comprises 5 readability levels corresponding to the Weekly Reader's Level 2, Level 3 and Level 4 and BBC-BiteSize KS4 and GCSE levels. The corpus contains 615 documents per level with average document length of 23.4 sentences at the lowest level and 27.8 sentences at the highest level.

As the primary purpose of our work is to build a readability classifier for people with autism, we normalized the

[4]http://www.weeklyreader.com/
[5]http://www.bbc.co.uk/education

WeeBit corpus to include texts of only two readability levels: Easy and Difficult, to match the format of the corpus evaluated by people with autism. Thus, texts in the WeeBit corpus with class labels BitGCSE and BitKS3 (age 11-16) were mapped to Difficult and those with class labels WRLevel2 and WRLevel3 (age 9 -11) were mapped to Easy. Instances representing texts of class label Weekly Reader Level4 were filtered from the dataset, as the intended readership of this class (people aged 9-12) overlaps with that of Weekly Reader Level3 (9-10), BitKS2 (7-11), and BitKS3 (11-14).

3.2. Extrinsic Evaluation

3.2.1. ASD Corpus: Developing Reading Passages Evaluated by People with Autism

This section presents the design and procedure for the evaluation of the text complexity of reading passages by people with autism. 27 texts from various domains were evaluated by 26 different people with autism (texts 1-16 by 20 people and texts 17-27 by 18 people).

Design: The participants were asked to read text passages and answer three multiple choice questions (MCQs) per passage. Evaluation of the difficulty of the texts is then based on their answers to the questions and their reading time scores.

Text passages: The text set included a total of 27 text passages which vary in difficulty and were obtained from the Web covering miscellaneous domains and registers (Table 1). The size of the text set is small because the length of each text and the number of texts presented to each participant was selected with a view to avoid fatigue and to comply with ethical considerations. Table 1 summarises some of the characteristics of the texts included in this study. The Flesch-Kincaid Grade Level (FKGL) is proportional to text difficulty. Conversely, Flesch Reading Ease (FRE) score, which is expressed on a scale from 0 to 100, is inversely proportional to text difficulty.

Participants: The texts presented in this study were evaluated in two consecutive sessions by two groups of participants. Texts 1-16 were evaluated by Group 1, consisting of 20 adult participants (7 female, 13 male). Texts 17-27 were evaluated by Group 2, consisting of 18 adult participants (11 male and 7 female). All participants had a confirmed diagnosis of autism and were recruited through 4 local charity organisations. None of the 26 participants had other conditions affecting reading (e.g. dyslexia, intellectual disability, aphasia etc.). Mean age (m) for Group 1 in years was m = 30.75, with standard deviation SD = 8.23, while years spent in education, as a factor influencing reading skills, were m = 15.31, with SD = 2.9. For Group 2, mean age in years was m = 36.83 , SD = 10.8 and years spent in education were m = 16, SD = 3.33. All participants were native speakers of English.

Text classification results: The numbers of correct and incorrect answers provided by each participant to the questions for each text were recorded, as was the reading time measured in seconds. First, each reading time was divided by the number of words in the text in order to obtain raw reading time score. After that an answering score was obtained by counting the number of correct answers each participant had given to the 3 questions for each text. Thus, if a participant had answered 2 out of 3 questions correctly for Text 1, then Text 1 has an answering score of 2 for this participant. Finally, to capture the relation between reading time and correctness of the answers, each answer score was divided by the raw reading time for the same participant in order to obtain one single score per text. This was done because answering score is proportional to comprehension level (the more correct answers, the easier the text), while reading time is inversely proportional to comprehension level: the longer a participant reads a text, the more difficult that text is for the participants. Thus, texts were classified based on one general index for each participant for each text.

	Genre	Words	FKGL	Flesch
T1	Easy-read	77	8.16	60.11
T2	Easy-read	96	6.73	67.33
T3	Easy-read	74	2.71	92.54
T4	Easy-read	178	5.52	75.33
T5	Easy-read	77	5.79	70.67
T6	Easy-read	121	1.75	95.00
T7	Easy-read	58	6.63	68.16
T8	Educational	163	4.93	79.548
T9	Educational	178	4.671	80.22
T10	Educational	206	7.577	65.437
T11	Educational	189	9.276	56.758
T12	Newspaper	226	11.983	40.658
T13	Newspaper	160	8.866	59.82
T14	Newspaper	163	8.765	66.657
T15	Newspaper	185	14.678	45.34
T16	Newspaper	188	9.823	58.298
T17	General	108	4.243	82.305
T18	General	141	4.561	79.108
T19	Newspaper	166	10.344	57.859
T20	Educational	209	6.087	70.124
T21	Educational	151	5.783	60.258
T22	Educational	158	6.102	57.2013
T23	Newspaper	198	13.204	46.481
T24	General	147	11.035	51.965
T25	Encyclopedic	101	8.229	55.011
T26	Encyclopedic	100	2.943	94.15
T27	Encyclopedic	113	6.963	67.304

Table 1: Characteristics of the texts included in the experiment

A Shapiro-Wilk test showed that the general text scores are non-normally distributed. A Friedman test was performed, confirming that there were significant differences between scores obtained for different texts ($\chi^2(16) = 55.258$, $p < 0.000$). After that a Wilcoxon Signed Rank test with Holm-Bonferroni correction was used to determine where the differences in text scores are and on this basis the texts were divided into two groups of "Easy" texts (texts 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 17, 18, 24, 25, 26 and 27) and "Difficult" texts (texts 12, 13, 14, 15, 16, 19, 20, 21, 22 and 23). A Friedman test was applied to each group individually, indicating that there were no statistically significant

differences between the answer scores to the texts in each group (Easy texts: $\chi^2(10) = 15.046$, $p < 0.130$; difficult texts: $\chi^2(5) = 9.676$, $p < 0.085$). A Wilcoxon Signed Rank test confirmed that there was a statistically significant difference between the two groups of Easy and Difficult texts (z = -5. 104, $p < 0.000$).

3.2.2. LocalNews corpus and readers with mild intellectual disability

The LocalNews corpus (Feng et al., 2009) consists of 11 original and 11 simplified news stories and is, to the best of our knowledge, the only other resource in English, for which text complexity has been evaluated by people with cognitive disabilities. The articles were first manually simplified by humans, a process in which long and complex sentences were split and important information contained in complex prepositional phrases was integrated in separate sentences. Lexical simplification included the substitution of rare words with more frequent ones and deletion of sentences and phrases not closely related to the meaning of the text. The texts were then evaluated by 19 adults with mild intellectual disability, showing significant differences between their comprehension scores for the two classes of documents (Feng et al., 2009).

4. Features

A total of 43 features were evaluated in the WeeBit, ASD, and LocalNews corpora. These features are grouped in 5 categories, as presented below.

Lexico-semantic: This group includes surface lexical features such as *Syllables in long words* and *Average word length in syllables*, and semantic features such as *Number of polysemous words* and *Polysemous type ratio*. Lexical diversity is measured through *Type-token ratio, Vocabulary variation* (word types divided by common words not in the text) and *Number of numerical expressions*. Statistical measures include *Numbers of infrequent words*, as well as *Total number of words* and *Dolch-Fry Index*, which evaluates the proportion of words in the text that appear in the *Fry 1000 Instant Word List* (Fry, 2004) or the *Dolch Word List* (Dolch, 1948)

Syntactic: Here were included surface features such as *Long sentences* (proportion of sentences in the text that contain more than 15 words), *Words per sentence* (total words in input file / total sentences in input file), *Average Sentence Length, Total number of sentences* and *Paragraph index* (10 * total paragraphs / total words). Also, features quantifying the number of punctuation marks indicating syntactic complexity were evaluated: *Number of Semicolons/suspension points, Number of Unusual punctuation marks* and *Comma index* (10 * total commas in input file / total words in input file). The cognitive load imposed in syntactic processing by the presence of non-canonical syntactic constructions, verb forms, and modifiers was measured through features such as *Number of passive verbs, Agentless passive density, Negations* and *Negation density*.

Features of cohesion: Cohesion is a property of the text which reflects the ease with which different components are integrated into a whole. As discussed in Section 1, this is especially problematic for readers with autism. We evaluated several features indicating referential and discourse cohesion *Number of illative conjunctions, Comparative conjunctions, Adversative conjunctions, Pronouns* and *Definite descriptions*. These features were computed as in (McNamara et al., 2010)

Cognitively-motivated features: This class of features was obtained through human rankings as explained in Section 2. People with autism have been shown to sometimes find it difficult to form mental representations of word referents if the words are too abstract or unfamiliar (Martos et al., 2013). The source for these features for our classifier were the word lists in the MRC Psycholinguistic database (Coltheart, 1981), where each word has an assigned score as described in Section 2. These features included *Absolute Average Word Frequency, Age Of Acquisition, Imagability, Concreteness* and *Familiarity*. These indices apply only to those words which were present in the MRC database lists, as opposed to all words in the texts, which is why they are referred to as "found only" in Table 2. The number of personal words in a text is hypothesised to improve ease of comprehension (Freyhoff and Van Der Veken, 1998), which is why evaluation of the number of first and second person pronominal references were included as features in the classification model.

Readability formulae: This list included popular formulae such as ARI (Smith et al., 1989), Coleman-Liau (Coleman, 1971), Fog Index (Gunning, 1952), Lix (Anderson, 1983), SMOG Reading Ease (McLaughlin, 1969), Flesch Reading Ease (Flesch, 1948), Flesch Kincaid Grade Level (Kincaid et al., 1975) and FIRST Readability Index (Jordanova et al., 2013). The latter is given by the formula:

$$95.43 - (0.076 \times CI) + (0.201 \times PI) - (0.067 \times SI) - (0.073 \times SLI) - (35.202 \times TTR) - (1.060 \times VV) + (0.778 \times DFI)$$

Where CI is Comma Index, PI is Paragraph Index, SI is Syllable Index, SLI is Sentence Length Index, TTR is Type Token Ratio, VV is Vocabulary Variation, and DFI is Dolch-Fry Index. It was developed specifically for people with autism in the EC-funded FIRST project by professional in mental healthcare.

5. Training and Evaluation Results

The partial decision tree (PART) classifier distributed in Weka (Frank and Witten, 1998) was used to derive the decision lists presented in Tables 2 and 3. This partial decision tree served as the text classifier in our experiments.[6] Of the classifiers distributed with *Weka*, PART had best performance in testing. The decision list consists of 14 rules. Of the 43 features tested, 28 are directly exploited by this automatically learned rule set.

The learned classification model classifies a text as *difficult* if evaluation of the features presented in Section 4 reveals that it meets all of the conditions in one or more of the sets presented in Table 2. Similarly, the model classifies a text as *easy* if evaluation of the features presented in Section 4

[6]PART is an iterative learning procedure which works by building a partial C4.5 decision tree (Quinlan, 1993) in each iteration and making the "best" leaf into a rule for inclusion in the model.

reveals that it meets all of the conditions in one or more of the sets presented in Table 3.

Set	Feature	Value
1	Long sentences	> 2
	Age of acquisition found only	> 6.04
	Illative conjunctions	> 1
	Pronoun2Incidence	> 0
	Average sentence length	> 10.97
2	Long sentences	> 4
	Age of acquisition found only	> 5.8
	Pronouns	> 11
3	Age of acquisition found only	> 6.51
	Possible senses	> 1844
	Lix	> 27.1
4	Age of acquisition found only	> 6.34
	Spanish readability index	> 67.876001
	ARI	> 7.9
5	Paragraph index	> 0.565217
	AgeOfAcquisition	> 5.51
	Syllable long words	≤ 0.705882
6	AgeOfAcquisitionFoundOnly	> 6.4
	AgeOfAcquisitionFoundOnly	> 6.73
7	ImagabilityFoundOnly	≤ 395.18
	ConcretenessFoundOnly	≤ 362.94
8	AverageSentenceLength	> 11.27
	FamiliarityFoundOnly	≤ 582.53
9	Illative conjunctions	≤ 9

Table 2: Conditions characterising *difficult* texts

Set	Feature	Value
1	AgeOfAcquisitionFoundOnly	≤ 6.51
	Polysemous type ratio	> 0.609442
	AverageSentenceLength	≤ 16.23
	Long sentences	≤ 5
	Fog	≤ 9
	NegationDensity	≤ 10.13
2	Pronoun2Incidence	≤ 12.5
	AverageWordFrequencyAbs	> 359091.82
	Passive verbs	≤ 4
	Average sentence length	≤ 17.16
	Infrequent words	≤ 116
	Adversative conjunctions	≤ 0
3	Polysemous type ratio	> 0.632075
	FleschKincaidGradeLevel	≤ 10.37
	Comma index	> 0.167131
	Illative conjunctions	≤ 6
4	Long sentences	≤ 3
	Fog	≤ 11.7

Table 3: Conditions characterising *easy* texts

We evaluated the classifier with respect to its ability to label input texts as either *easy* or *difficult* for people with ASD. The test data consisted of the three corpora presented in Section 3. Table 4 displays the f_1-scores achieved by the classifier when processing these texts. The WeeBit corpus was exploited as training data. The f_1-scores achieved by the model in classifying texts from this corpus were obtained via ten-fold cross validation.

The table includes statistics on the accuracy of three different versions of the classifier derived from different feature sets (Column *Feature Selection*). The first version (*All*) exploits all 43 features presented in Section 4. The second version (*FKGL*) exploits just one feature, *Flesch-Kincaid Grade Level*. The third version exploits only the 28 features that are used to condition the rules in the sets derived by the PART classifier (*PART*).

Feature Selection	f_1-score		
	WeeBit	Local News	ASD Corpus
All	0.989	1	0.89
FKGL	0.894	0.829	0.654
Features exploited by PART rulesets	0.990	0.725	0.748

Table 4: Evaluation results of the text classifier for the three collections

Table 4 reveals that *PART* is more accurate than the other models in its classification of texts from the WeeBit corpus, but less accurate when classifying texts of the other two categories. Given that we seek to optmimise the classification of texts in *ASD Corpus*, the classifier exploiting the full set of 43 features is preferred in this context. *All* is more accurate than both *FKGL* and *PART* over texts of both *Local News* and *ASD Corpus* categories.

6. Discussion

The results presented in Section 5 show that the classifier trained on the WeeBit corpus outperforms the widely used Flesch-Kincaid Grade Level (FKGL) formula by achieving an f_1 score of 0.89 when classifying texts, compared to f_1 score of 0.654 for FKGL. There are two interesting observations which could be made based on the results from this study.

The baseline model containing all 43 features performed better than the model including only those features which were retained by the features selection algorithm in PART (PART feature set). In fact, when evaluating by 10 fold cross-validation of the WeeBit corpus, use of the PART feature set achieves slightly better performance. However, the model using only these features does not generalise well to the other text collections. We are not certain of their role in the classification process, but the features in the baseline model which were not included in the PART feature set appear to help the classifier to generalise better.

Readers will note that when classifying texts from the LocalNews corpus when exploiting all features, the classifier worked with perfect accuracy. It should be noted that the number of texts in this set is too small to be considered truly representative of those sought by readers with mild intellectual disability. Classifying a single text incorrectly would reduce f_1 to 0.94. Another reason is the fact that the differences between Easy and Difficult documents in the LocalNews texts have been artificially introduced by manual simplification, in which sentence length and word length have been deliberately shortened. As a result, all formulae and classifiers have an advantage when distinguishing between the two classes. This raises an important issue about the kinds of data used to measure the external validity of readability models. In the best case scenario, this data should consist of documents "in their own

right" rather than texts which are modified versions of other texts. This observation gives additional credit to the result obtained over the ASD corpus, in which Easy texts were not derived from Difficult ones. It would be interesting to test whether original and simplified versions of documents would make a suitable training set for readability classifiers for people with cognitive disabilities, where the simplification has been done with respect to the particular difficulties of the target population. Further, it would be interesting to investigate whether a classifier trained on this type of user-specific data would outperform other classifiers trained on larger scale but generic data.

It is important to note that the findings of this paper and the classifications of the texts from the ASD-corpus are relevant to the population of adults with high-functioning autism and are not necessarily applicable to adults at the lower ends of the spectrum, children, or people with cognitive disabilities other than autism.

7. Conclusions and Future Works

This paper presented work towards the development of a machine learning-based classifier which distinguishes between two levels of difficulty of texts for adults with high-functioning autism. First, the ASD corpus was created containing 27 texts classified as easy or difficult through a reading comprehension experiment involving autistic adults. Then a classifier was trained on the WeeBit corpus containing graded educational materials for children between ages 7-16. The generalizability of the model was tested on the ASD corpus and the LocalNews corpus (evaluated on people with mild intellectual disability), where the presented classifier outperformed the widely-used Flesch-Kincaid Grade Level formula (Kincaid et al., 1975) for both datasets.

Future work involves developing a more fine-grained model to distinguish between 3 levels of difficulty suitable for adults with high-functioning autism, as well as adults with autism and comorbid mild intellectual disability. Another future challenge is the development of a tool to distinguish between easy and difficult sentences for this population, thus optimising future text simplification decisions.

8. Acknowledgements

The authors are indebted to all participants, who took part in the reading comprehension experiments, as well as to Dr. Georgiana Marsic for her valuable help with the extraction of some of the features.

9. Bibliographical References

American Psychiatric Association, . (2013). Diagnostic and Statistical Manual of Mental Disorders (5th ed.).

Anderson, J. (1983). Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., Matthews, F. E., and Brayne, C. (2009). Prevalence of autism-spectrum conditions: Uk school-based population study. *The British Journal of Psychiatry*, 194(6):500–509.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:1–26.

Bosseler, A. and Massaro, D. W. (2003). Development and evaluation of computer-animated tutor for vocabulary and language learning in children with autism. *Journal of Autism and Developmental Disorders*, 33(6):553–567.

Brugha, T., McManus, S., Meltzer, H., Smith, J., Scoth, F. J., and Purdon, S. (2007). Autism spectrum disorders in adults living in households throughout England. Report from the Adult Psychiatric Morbidity Survey 2007. Technical report, The National Health Service Information Centre for Health and Social Care, London.

Brugha, T. S., Cooper, S. A., and McManus, S. (2012). Estimating the Prevalence of Autism Spectrum Conditions in Adults: Extending the 2007 Adult Psychiatric Morbidity Survey. Technical report, NHS, The Health and Social Care Information Centre., London.

Coleman, E. B., (1971). *Developing a technology of written instruction: some determiners of the complexity of prose*. Teachers College Press, Columbia University, New York.

Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Dolch, E. W. (1948). *Problems in Reading*. The Garrard Press, Champain, IL.

Dubay, W. H. (2004). *The Principles of Readability*. Impact Information.

Feng, L., Elhadad, N., and Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 229–237.

Feng, L. (2009). Automatic readability assessment for people with intellectual disabilities. *SIGACCESS Access. Comput.*, (93):84–91, January.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.

Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. In J. Shavlik, editor, *Fifteenth International Conference on Machine Learning*, pages 144–151. Morgan Kaufmann.

Freyhoff, G., H. G. K. L. T. B. and Van Der Veken, K. (1998). Make it Simple. European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability. Technical report, ILSMH European Association.

Frith, U. and Snowling, M. (1983). Reading for meaning and reading for sound in autistic and dyslexic children. *Journal of Developmental Psychology*, 1:329–342.

Fry, E. (2004). *1000 Instant Words: The Most Common Words for Teaching Reading, Writing and Spelling*. Teacher Created Resources.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill, New York.

Happé, F. and Frith, U. (2006). The weak coherence account: Detail focused cognitive style in autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 36:5–25.

Happé, F. (1997). Central coherence and theory of mind in autism: Reading homographs in context. *British Journal of Developmental Psychology*, 15:1–12.

Harris, T. L. and Hodges, R. E. (1995). *The Literacy Dictionary: The Vocabulary of Reading and Writing*. International Reading Association.

Huenerfauth, M., Feng, L., and Elhadad, N. (2009). Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '09, pages 3–10, New York, NY, USA. ACM.

Jansche, M., Feng, L., and Huenerfauth, M. (2010). Reading difficulty in adults with intellectual disabilities: Analysis with a hierarchical latent trait model. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '10, pages 277–278, New York, NY, USA. ACM.

Jordanova, V., Evans, R., and Pashoja, A. C. (2013). First project - benchmark report (result of piloting task). Central and Northwest London NHS Foundation Trust. London, UK.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Technical report, CNTECHTRA Research Branch Report.

Martos, J., Freire, S., González, A., Gil, D., Evans, R., Jordanova, V., Cerga, A., Shishkova, A., and Orasan, C. (2013). FIRST Deliverable - User preferences: Updated. Technical Report D2.2, Deletrea, Madrid, Spain.

McLaughlin, H. G. (1969). SMOG grading - a new readability formula. *Journal of Reading*, pages 639–646, May.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., and Graesser, A. C. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion, May.

O'Connor, I. M. and Klein, P. D. (2004). Exploration of strategies for facilitating the reading comprehension of high-functioning students with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 34:2:115–127.

Pikulski, J. J. (1995). Re a d a b i l i t y.

Putnam, C. and Chong, L. (2008). Software and technologies designed for people with autism: What do users want? In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '08, pages 3–10, New York, NY, USA. ACM.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rello, L., Baeza-yates, R., Dempere-marco, L., and Saggion, H. (2012a). Frequent Words Improve Readability and Shorter Words Improve Understandability for People with Dyslexia. (1):22–24.

Rello, L., Bayarri, C., and Gorriz, A. (2012b). What is wrong with this word? dyseggxia: A game for children with dyslexia. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12, pages 219–220, New York, NY, USA. ACM.

Senter, R. J. and Smith, E. A. (1967). Automated Readability Index. Technical Report AMRL-TR-6620, Wright-Patterson Air Force Base.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 574–576, New York, NY, USA. ACM.

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:1:77–109.

Smith, D. R., Stenner, A. J., Horabin, I., and Malbert Smith, I. (1989). The lexile scale in theory and practice: Final report. Technical report, MetaMetrics (ERIC Document Reproduction Service No. ED307577)., Washington, DC:.

Tronbacke, B. (1997). Guidelines for Easy-to-Read Materials. Technical report, IFLA, The Hague.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Stroudsburg, PA, USA. Association for Computational Linguistics.

Štajner, S., Evans, R., Orasan, C., and Mitkov, R. (2012). What can readability measures really tell us about text complexity? In Luz Rello et al., editors, *Proceedings of the LREC'12 Workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Štajner, S., Mitkov, R., and Pastor, G. C., (2014). *Simple or not simple? A readability question*. Springer-Verlag, Berlin.

Whyte, E. M., Nelson, K. E., and Scherf, K. S. (2014). Idiom, syntax, and advanced theory of mind abilities in children with autism spectrum disorders. *Journal of Speech, Language, and Hearing Research*, 57:120–130.

Yaneva, V. and Evans, R. (2015). Six good predictors of autistic text comprehension. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 697–706, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Yaneva, V., Temnikova, I., and Mitkov, R. (2015). Accessible texts for autism: An eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, pages 49–57, New York, NY, USA. ACM.

SimplexEduReading: Simplification of Natural Language for Reading Comprehension Improvement in Education

Estela Saquete, Ruben Izquierdo, Sonia Vázquez
University of Alicante, Vrije Universiteit Amsterdam, University of Alicante
Alicante. Spain, Amsterdam. The Netherlands, Alicante. Spain
stela@dlsi.ua.es, ruben.izquierdobevia@vu.nl, svazquez@dlsi.ua.es

Abstract

The main aim of this paper is presenting a system, known as SimplexEduReading, capable of transforming educational natural language texts in Spanish into simpler and enriched texts in order to improve the reading comprehension process. The goal is to help people with comprehension problems, for instance, deaf people or people who are learning a language. For each source of difficulty our system provides different transformation processes using Natural Language Procesing techniques such as: text summarization, preserving the original meaning, and text enrichment with very simple additional information. In order to reinforce understanding we add extra information to the original texts: 1) name entities detection, adding information about them, for example, related images, information from Wikipedia or synonyms; 2) temporal expressions detection and resolution, adding a chronological timeline of the events; 3) complex sentences simplification, dividing them into simpler ones; 4) words definitions in the original text; and 5) main topics detection in order to easily provide a context of the text to the reader.

Keywords: Text simplification, Reading comprehension, Name Entity Recognition, Temporal Expressions Resolution

1. Introduction

According to the PISA 2000 report, reading comprehension is defined as "the ability to understand, use and think about information from written texts, with the aim of achieving personal goals, developing the knowledge and the personal potential, and taking efficient part in the society". For this reason, the reading comprehension has currently become one of the main important research issues in the psychology and education field (Olivé, 2009). Different research works have determined that the skills and conditions required to a properly reading comprehension are many and very complex.

This paper presents a system (SimplexEduReading) capable of transforming educational texts in Spanish into very easy reading texts by using different Natural Language Processing (NLP) techniques, in order to support people with problems in reading comprehension. The process of simplification and enrichment of texts consists of the automatically detection of those specific linguistic features of input texts: a) the reduction and removal of obstacles, but preserving in all cases the original meaning of the text, and b) the enrichment of texts using different tools.

2. Background

Low levels of ability in reading comprehension is an increasing problem in this society, as it was presented in the previously mentioned PISA report, published in 2006. Moreover, the problem of reading comprehension becomes even harder for deaf people. This problem has been studied during years, not only from a lexical perspective, but also from a syntactic point of view (King and Quigley, 1985) (Berent, 1996) (LaSasso and Davey, 1987) (Paul and Gustafson, 1991). Previous studies have detected the following linguistic barriers that hearing impaired people find in reading comprehension:

- Ambiguity problems: there are a lot of words that can have multiple meanings, and have a different sense depending on the context in which they appear. All of these polysemous words can lead to multiple problems for reading comprehension.
- Limited vocabulary: this type of readers focuses fundamentally on common words, using very specific nouns and familiar verbs. Most of the times they have problems recognizing name entities and contextualizing them.
- Complex sentences: difficulties in the interpretation of complex syntactic structures, that are different from the basic syntactic structures like noun-verb-noun and subject-verb-object. Therefore, more complex structures like transitive active sentences, passive sentences or subordination increase the problem in reading comprehension.
- Temporal reasoning: problems in locating events in the temporal timeline. A text has usually different temporal points, and it goes temporally back and forward, implying the interpretation of temporal signals and temporal expressions for the whole comprehension.

Usually, most of these studies are focused in English, and only some of them deal with Spanish texts, not only at a lexical level (Mies, 1992), but also at a syntactic level (Stockseth, 2002). Besides, there are also some works in the educational field (Alegría and Leybaert, 1985) (Asensio and Carretero, 1989) (Mora, 1989). Within the Natural Language Processing research area, there are several works related to the extraction of sign languages from written and spoken texts with automatic and semi-automatic approaches (Parton, 2005) (C. Wu, 2004) (Duchnowsku et al., 2000). Additionally, the MAS project (Manchón, 2001) must be pointed out. The aim of this project is to check the

effects of using a multimedia tool to improve the reading comprehension using sign languages.

3. Main objectives

The main objective of the whole research is the design, development and evaluation of a system that is able to transform Spanish texts into texts that are easier to understand, and we are going to focus on hearing impaired people. In this paper we focus on a system proposal that would help these people to improve their reading comprehension of the texts.

The transformation issue implies: a) the automatic detection of specific linguistic features of the input texts that may interfere in the reading comprehension together with the automatic decrease and/or removal of these barriers, taking into account that the original meaning of the text has to be preserved, and b) the enrichment of these texts using different tools like WordNet[1], Simple Wikipedia[2], Wiktionary[3] or Google Images[4]. Natural Language Processing techniques will be applied to locate and eliminate these barriers, transforming them into much simpler elements or enriching the elements with additional information, thus facilitating the understanding process. Such language barriers are derived from complex structures, ambiguity in terms, lack of context and problems with timelines, so the tool would generate supporting material by means of images, definitions of proper nouns extracted from online encyclopaedias, timelines and resolution of temporal expressions to concrete dates and setting the context and topics of the original texts.

Specifically, the main goals in this research are:

1. Analyse with potential users and their assistants the linguistic features that difficult reading comprenhension. A list of linguistic barriers will be defined together and the possible alternatives or supporting materials that could help these users in comprenhension issues.
2. To analyze in depth the existing different approaches in NLP to treat each barrier that difficult text comprehension.
3. To develop the necessary tools to solve these problems, evaluating each tool independently.
4. To integrate all the necessary tools in a system that, giving an input text, is able to obtain an easy reading text, with the supporting material previously mentioned, as well as the reduction of language barriers of the text.
5. Defining a set of measures that are able to determine not only how the tool performs but also the improvement in reading comprehension that the users obtain after using the tool. Therefore, an intrinsic and extrinsic evaluation will be defined. This evaluation will be defined in further works.

[1]http://wndomains.fbk.eu/
[2]http://simple.wikipedia.org
[3]http://es.wiktionary.org
[4]http://images.google.es

As it was aforementioned, the existing technological tools in this field are mainly oriented to help reading comprehension using the sign language (Parton, 2005) (C. Wu, 2004) (Duchnowsku et al., 2000). However, our proposal aims at helping users by simplifying the text, preserving also its meaning. In this manner, not only the lexical comprehension is facilitated, but also the syntactic and semantic comprehension of the text.

The final system will provide a considerable improvement for hearing impaired people, due to the fact that it will increase their reading comprehension and therefore, it will allow them to widen their information and cultural horizons. Moreover, this help is very appropriate in the educational area, especially with deaf students or people who are learning new languages. Nowadays, teachers and professionals are helping these people by manually performing this simplification task in order to make easier the reading comprehension for these students.

4. Architecture of SimplexEduReading

The architecture of the final system is shown in Figure 1. The system would integrate different NLP tools and open resources in order to transform the texts into simpler ones and enrich them with additional information.

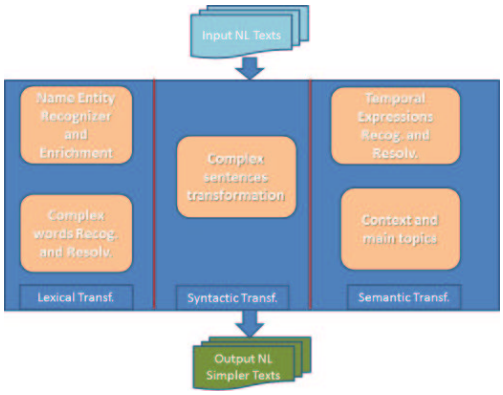

Figure 1: Architecture of SimplexEduReading

As it can be seen in the architecture, we distinguish in the system three main parts: the lexical transformations, that include the recognition of Name Entities with the enrichment of them using Wikipedia and images, and the detection of complex words whose definition will be provided. Secondly, the syntactic transformations, that affected the transformation of complex syntactic structures to simpler ones. And finally, the semantic transformations, that include, the temporal expressions recognition and resolution, and the detection of main topics of the text.

5. Implemented modules

At this moment, part of the proposed system has been developed. In this step, the prototype system has been developed for Spanish. The interface is very simple and intuitive. The user can upload a document and the system requires

that the date of the input document is indicated in order to resolve the temporal expressions appearing in the text.
The modules that are part of the transformation of the text at this moment are: 1) Name Entity Recognition and Enrichment, 2) Complex Words Recognition and Resolution, 3) Temporal Expressions Recognition and Resolution and 4) Context and Main Topics Detection

5.1. Name Entity Recognition and Enrichment

In order to recognize and resolve Name Entities in the text, the web services provided by the OpeNER project[5] have been used. OpeNER is a project funded by the European Commission under the FP7 (7th Framework Program). OpeNER main goal is to provide a set of ready to use tools to perform some natural language processing tasks. Specifically, in this system we have used three of them: a) the tokenizer, b) the POStagger and c)the name entity recognizer. Once the entities are recognized, the system will mark them in blue colour font and it will allow the user to click them, if he or she wants to obtain additional information related to the Named Entity. This makes the information easy to access. For instance, Figure 2 shows an example of a text, where the Named Entities "Max Weber" and "Europa" are recognized.

Figure 2: NE in the tool

As shown in Figure 2, by clicking the entity, the system opens a pop-up window, where the entity is explained. This information is automatically obtained from Wikipedia[6]. By clicking the button next to the entity, another pop-up window appears with a set of images related to the entity and automatically extracted from Google Images[7].

5.2. Complex Words Recognition and Resolution

At this moment, due to the fact that complex words are not being detected, all the words in the text could be clicked in order to obtain a definition from an online dictionary. In our system, we used Wiktionary.org[8]. This tool is a collaborative project to produce a free-content multilingual dictionary. It aims to describe all words of all languages using definitions and descriptions. The information is presented following the same procedure as for the other types of entities (pop-up window).

Figure 3: Common entities in the tool

In the example shown in Figure 3 the word "Carisma" (Charisma) has been clicked and the dictionary was automatically invoked, presenting the different definitions of the word. In further work, the complex words will be dectected by using specific dictionaries of unfrequent words, and own dictionaries created from experts in the area.

5.3. Temporal Expressions Recognition and Resolution

Temporal entities in the text are automatically recognized and resolved using a tool called TERSEO (Saquete et al., 2005). TERSEO system is a tool that performs the recognition and resolution of temporal expression in texts using a knowledge database that was manually created for Spanish and it was automatically extended to other languages like English and Italian. TERSEO system obtained at TERN2004 competition[9] an F1 measure of 86% in recognition and 70% in resolution of the time expressions.
Given an input text, the system analyzes the text with a part-of-speech tagger. The system takes this information and a temporal expression grammar, and it is able to recognize temporal expressions. After that, the expressions are resolved using the information stored in a resolution rule database. Finally, using the specific dates and periods obtained, the events are ordered in a timeline sequence. Our proposed system integrates TERSEO, indicating in a red colour link when a temporal expression appears in the text. If the user clicks on the expression, a pop-up window appears with the exact date or period of dates that the expression is referring to.
For instance, the temporal expression "1984" may be found by calling to TERSEO system and marked in the text. By clicking in the temporal expression,a pop-up window is obtained with the exact date or period of dates the expression is referring to. As further work, a graphical timeline of the events in the text will be provided to the user for an easy interpretation of the information.

5.4. Main Topics Detection

Topics provide a crucial piece of information to understand what a text is about. These topics contribute to frame the

[5]http://www.opener-project.org/
[6]http://es.wikipedia.org
[7]http://images.google.es
[8]http://es.wiktionary.org/

[9]The most important competition of temporal processing systems working with TIDES annotation

meaning of the text for the reader understanding, by supplying a high level, yet useful, information. We decided to use WordNet Domains labels as our possible list of topics. WordNet Domains (WND) contains around 200 labels organized in a hierarchical structure. All WordNet synsets have been annotated semi–automatically with these WND labels. Some examples of WND labels are: ECONOMY, ARCHITECTURE or RELIGION. We believe that the information provided by these labels is highly informative to capture the meaning of a text, and they are specially interesting considering the goal of this paper.

We conducted a simple but effective heuristic to extract the WND labels for a given text: from the monosemous words (words with a single meaning according to WordNet), the WND labels are extracted and accumulated over the whole text. In this manner a frequency is obtained for every WDN label. Only those reaching a threshold frequency (that can be set by the user) are returned. We translated the WND labels from its original language (English) to Spanish using the Google translation service as support for the posterior manual checking of this translation.

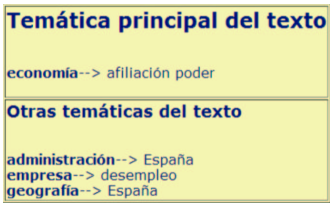

Figure 4: Topic Detection in the tool

As shown in Figure 4, the most important topics of the text are shown together with the words of the text that denote each topic. Less important topics are also shown in case they could be relevant to the reader.

6. Conclusions

In this paper we have presented a system, called SimplexEduReading, that processes educational texts in natural language with the purpose of: a) detecting and removing language barriers that difficult the reading comprehension process of deaf people, and b) adding supporting information to help in the reading comprehension.

The system at this moment is able to: 1) recognizing named entities or proper nouns providing extra information of them, for example, related images and information extracted from Wikipedia; 2) recognizing and resolving temporal expressions in the text, giving the exact date or period of dates that the expression is referring to; 3) providing definitions for common entities in the text extract from an online dictionary; and 4) providing the context of the text by a set of important topics extracted from the domains of the words.

As future work, the system will be completed in all the modules and it will be evaluated by real deaf university students in order to determine the improvement in reading comprehension for these students. A questionnarie with comprehension questions related to the texts will be proposed to the students to measure the improvement compared to the same texts without the enrichment.

7. Acknowledgements

This paper has been supported by the University of Alicante with the project GRE11-21 and by the Spanish government, Ministerio de Economía y Competitividad con número de referencia TIN2012-31224.

8. Bibliographical References

Alegría, J. and Leybaert, J. (1985). Adquisición de la lectura en el niño sordo: un enfoque psicolingüístico. *Investigación y Logopedia*.

Asensio, M. and Carretero, M. (1989). La lectura en los niños sordos. *Cuadernos de pedagogía*, 174.

Berent, G. (1996). The acquisition of English Syntax by Deaf Learners. *Handbook of Second Language Acquisition*, pages 469–506.

C. Wu, Y. C. y. C. G. (2004). Text generation from Taiwanese Sign Language using PST-based language model for augmentative communication. *IEEE Trans Neural Syst. Rehabil Eng.*, 12(4).

Duchnowsku, P., Lum, D., Krause, J., Sexton, M., Bratakos, M., and Braida, L. (2000). Development of speechreading supplements based on automatic speech recognition. *IEEE Trans Biomed Eng.*, 47(4):487–496.

King, C. and Quigley, S. (1985). Reading and Deafness. *Collegue Hill Press*.

LaSasso, C. and Davey, B. (1987). The relationship between lexical knowledge and reading comprehension for prelingually, profoundly hearing impaired students. *The Volta Review*, 89:211–220.

Manchón, A. F. (2001). La comprensión lectora en personas sordas adultas y el acceso a la Universidad. *ISAAC 2001: Odisea de la Comunicación. Segundas Jornadas sobre comunicación Aumentativa y Alternativa*.

Mies, B. (1992). El léxico en la comprensión lectora: Estudio de un grupo de alumnos sordos adolescentes. *Rev. Logop., Fon., Audiol.*, XII.

Mora, J. A. F. (1989). La lectura en el corriculum escolar del niño sordo. *Rev. Logop Fon Audiol*, IX.

Olivé, M. (2009). La lectura: una necesidad para la inclusión social y la democracia. *Separata del Manifiesto PIAPAS. Confederación española de familias de personas sordas*.

Parton, B. (2005). Sign language recognition and translation: a multidisciplined approach from the field of the artificial intelligence. *J. Deaf Stud Deaf Educ.*

Paul, P. and Gustafson, G. (1991). Hearing-impaired students' comprenhension of hig-frequency multi-meaning words. *Remedial and special education*, 12:52–62.

Saquete, E., Muñoz-Guillena, R., and Martínez-Barco, P. (2005). Event ordering using TERSEO system. *Data and knowledge Engineering Journal*, 58.

Stockseth, D. R. (2002). Comprensión de la sintaxis española por lectores sordos chilenos. *Revista Signos*, 35.