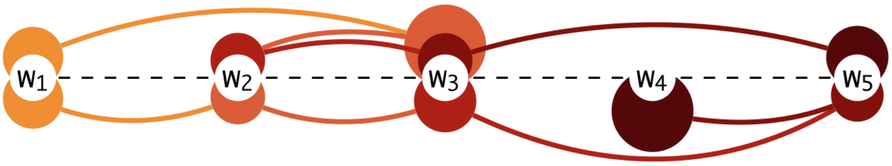
GLIMPSED

Improving natural language processing with gaze data

Klerke, Sigrid

SIGRID KLERKE

GLIMPSED



GLIMPSED

IMPROVING NATURAL LANGUAGE PROCESSING WITH GAZE DATA

SIGRID KLERKE

August 2016

Sigrid Klerke: *Glimpsed,* Improving natural language processing with gaze data, Copenhagen, August 2016

SUPERVISOR:
Anders Søgaard, Prof., Dr. Phil., PhD

SUPERVISOR:
Mads Poulsen, Assoc. Prof., PhD

AFFILIATION:
Centre for Language Technology,
Department of Nordic Research
Faculty of Humanities
University of Copenhagen

SUBMITTED:
August 2016

Til Vibe og Irene

ABSTRACT

This thesis addresses the problem of detecting challenging text by exploring whether recordings of readers' eye movements can be leveraged for learning what parts of texts obstruct readers, and investigates how this information can help improve NLP applications.

The problem of detecting and handling errors and deviations that make text unnecessarily difficult to read is becoming increasingly important to address, as the use of language technologies for improving information accessibility and communication efficiency grows.

It is necessary to address this problem regardless of whether the challenging points were introduced by human writers or Natural Language Processing (NLP) systems, and solving the problem can benefit both human readers and downstream NLP systems.

In addition, changes in language use and new norms can arise faster than they can be formally described or included in corpora. This poses a challenge to keeping systems that rely on such resources updated. Relying on readers' gaze reactions for determining what is an error instead can help alleviate this problem.

In the thesis, four independent studies target the tasks of automatic text simplification, machine translation, sentence compression and lexical complexity detection. The empirical investigation presents evidence that it is possible to obtain and make use of information about text complexity from readers' gaze behaviour.

The results presented and discussed herein contribute to the field of Natural Language Processing (NLP) by identifying important potentials and limitations of several different approaches to using gaze data in NLP.

RESUMÉ

Denne afhandling addresserer problemet at detektere svære tekstpassager ved at undersøge om optagelser af læseres øjenbevægelser kan udnyttes til at lære hvilke dele af tekster der forstyrrer læsningen, og studerer hvordan denne informationskilde kan udnyttes til at forbedre sprogteknologier.

Udfordringen med at håndtere fejl og sproglige afvigelser som gør tekster unødigt svære at læse bliver mere presserende at addressere i takt med at sprogteknologiske værktøjer i stigende grad anvendes til at gøre information tilgængeligt og effektivisere kommunikation.

Det er nødvendigt at addressere denne udfordring uanset om den svære tekst skyldes menneskelige forfattere eller sprogteknologiske værktøjer og løsninger på problemet kan gavne både menneskelige læsere og eventuelle videre processeringssystemer.

Ydermere opstår nye sproglige normer og ordformer hurtigere end de kan beskrives i corpora, hvilket er en fortsat udfordring for sprogprocesseringssystemer der afhænger af den type resurser. Ved at lade læseres øjenbevægelser afgører hvad der tæller som fejl kan dette problem afhjælpes.

Afhandlingens fire uafhængige undersøgelser behandler delopgaverne automatisk tekstsimplificering, maskinoversættelse, sætningskomprimering og detektering af særligt svære ord. Det empiriske arbejde viser at det er muligt at opdage og udnytte indikatorer på obstruerende tekst direkte fra læseres øjenbevægelser.

Tilsammen bidrager præsentationen og diskussionen af disse resultater til det sprogteknologiske felt ved at identificere væsentlige muligheder og begrænsninger ved en række tilgange til at bruge øjenbevægelser i natursprogsprocessering.

PUBLICATIONS

This thesis presents my work as a PhD student at the University of Copenhagen (UCPH). The work has been carried out, written and presented as articles in collaboration with supervisors and colleagues at UCPH and abroad. The articles have been reformatted for inclusion in the thesis but are otherwise identical to the published versions.

Klerke, Sigrid, Héctor Martínez Alonso, and Anders Søgaard (2015). "Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences." In: *Proceedings of NODALIDA 2015*.

Klerke, Sigrid, Sheila Castilho, Maria Barrett, and Anders Søgaard (2015). "Reading metrics for estimating task efficiency with MT output." In: *Proceedings of Workshop on Cognitive Aspects of Computational Language Learning, EMNLP 2015*.

Klerke, Sigrid, Yoav Goldberg, and Anders Søgaard (2016). "Improving sentence compression by learning to predict gaze." In: *Proceedings of NAACL-HLT 2016 (Best short paper)*.

Klerke, Sigrid, Alexandra Uitdenbogerd, Falk Scholer, and Timothy Baldwin (2016). "Predicting lexical complexity from user gaze." Submitted for publication.

ACKNOWLEDGMENTS

This work was not completed in an isolated void, and the people who have been sourrounding me—be it for a single enlightening discussion or for the entire stretch and beyond—deserve credit for making this possible at all. To all of you: Thanks for taking the time.

Maria, Barbara, Bjørn and especially Héctor deserve special thanks for the hard work and countless coffee breaks each put into making my work better at the finishing line. During the same time, Vibe, Irene, Thorstein, Mathilde, Morten, and Linda showed infinite patience and went out of their way for me. Thanks, guys! Many more friends and family have done the same in the full course of this project, and I am grateful for that.

The dinner club requires a cheer: Through everything, whether Tuesday, Wednesday or Thursday, dinner club night is something as rare as certain to be a pleasure.

The writing group similarly gets a special thanks simply for being around—a quality which is invaluable in my book.

Through this journey, I have had the immeasurable fortune to have been sourrounded by the Coastal research group which have provided inspiration, support, tireless feedback and endless encouragement, fun and consolation. Thank you, friends!

For three months early in 2015 I was incredibly warmly welcomed as a visitor at the department of Information and Computing at the University of Melbourne. If I could change one thing, I would not have left my many new friends there again so soon.

Humbly refraining from trying to list everyone, this serves as a thank you to the many, many competent and caring people who have helped and supported me at the Centre for Language Technology, the Department for Nordic Studies and the Faculty of Humanities, University of Copenhagen.

Supervisors get the final declaration of respect and gratitude; Anders, Mads and Tim, as well as AJ and Héctor, thank you! I had not been able to get to this point without the spheres of freedom and support you created.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

NLP	Natural Language Processing
ML	Machine Learning
MT	Machine Translation
PoS	Part-of-Speech
RNN	Recurrent Neural Network
ATS	Automatic Text Simplification
LSTM	Long-Short-Term Memory (cell in RNN)
AOI	Area of Interest

Part 1

OVERVIEW

1 INTRODUCTION

1.1 MOTIVATION

Eye movements are the most salient behavioural trace of the reading process. With the current maturity of eye tracking technologies, this rich source of information about a text, obtained through a reader's interaction with it, deserves close investigation.

Consider the amount of resources we as computational linguists readily expend to obtain annotations that reflect some aspect of readers' experience and interpretation of texts; in this light the potential of gaze as annotation is still curiously underdeveloped.

In this thesis, I present investigations of novel ways of using gaze recording for informing evaluation[1] and development of NLP systems' ability to handle what I will broadly refer to as *text complexity*.

In Chapter 2, text complexity is operationalised in the context of each of the concrete tasks and study designs included in the four studies in Part II. The studies make up the core of this work.

In the thesis I attempt to bridge the notably diverging perspectives of psycholinguistics and NLP on whether text complexity ultimately pertains to traits inherent in the text or to readers' reactive cognitive processes.

Gaze recording data has been widely used in psycholinguistic research, providing rich insights into the largely automatic and unconscious cognitive processing that readers exert to derive meaning from text.

While the research presented in this thesis relies heavily on the body of knowledge that psycholinguistic studies have built, the research questions tackled herein are primarily concerned with the development of language technologies and thus will

[1] *Evaluation* here is used in the narrow sense of the machine learning term, denoting concrete, often standard procedures taken to quantify the performance of a model.

only selectively engage the lines of academic debate that have shaped the current psycholinguistic conceptualisation of the interplay of cognition and eye movements.

Before laying out the argument for exploring uses of gaze recordings in the context of NLP, a few supporting, yet orthogonal perspectives are presented here, as they have influenced the project from its inception, even though they do not fall within the scope of the current research.

Most importantly, the ubiquity of video cameras built into laptops, tablets and mobile phones has inspired a large leap towards making eye tracking technology relatively easily available, with lower prices and improved usability. This development ultimately provides a potential for utilizing readers' gaze for the development of individually tuned real time reactive language technologies sensitive to eye gestures. For instance, tutoring systems or voice-controlled virtual assistants are examples of applications that might benefit from gaze feedback. However, while the technology is mature, we still know very little about what structures in language are reliably reflected in gaze data outside psycholinguistic laboratories and their notoriously controlled stimuli.

The second important advantage of working with gaze data, is the rich detail awarded by rapid sampling frequencies compared to manually obtained measures of behaviour such as comprehension questionnaires and task performance.

Although rich data in general increases the need for preprocessing and thereby increases the risk of introducing ill-informed reductions, the richness of the data source provides a powerful platform. In particular, rich details allow for tailoring the data representation to the level of abstraction needed for fitting a definition of some construct of interest. This work explores several representational levels of abstraction which are presented in Chapter 3.

Note that fundamentally, the choice of level of abstraction necessarily disregards all information outside the chosen level, regardless of whether that was an intended effect. Consequently, the choice of level of abstraction, i. e. the data representation, inherently carries a risk of contributing to an idiosyncratic and overly narrow perspective on a data source. A narrowing that is further exaggerated by the pervasive requirement to be able to compare results across studies. In turn, idiosyncratic data representations may hinder or slow down the exchange of ideas and data between fields of research.

Importantly, I believe that this is where the NLP-community is particularly well-equipped to join the exploratory expedition, bringing to the table an ingrained tradition of tinkering—Intermittently labelled as data-promiscuity and theory-agnosticism, depending on who gets to apply the label.

1.2 RESEARCH QUESTION

In a broad perspective, the aim of this thesis is to identify ways to improve language technologies and the evaluation of them by enlisting knowledge and practices from cognitive psychology.

Concretely, the central research question concerns whether information about text complexity at sentence- and word-level can be obtained from readers' gaze behaviour, and how this information can be employed for improving applications in NLP.

The work presented in this thesis thus primarily builds on research from the fields of NLP and psycholinguistics. While these two fields share a common focus on human language, the disciplines pursue very different research goals; one is concerned with providing technologies that can handle human language in ways meaningful to human users while the other attempts to learn how humans go about processing language.

Necessarily, specific research questions lead to specific, tailored techniques for experimental and theoretical validation and innovation. While such specialized methodological and theoretical developments may be relevant to share across disciplines, any method, measure or model needs to be validated and appropriated to fit the different research goals and constraints of a new context.

Within the part of NLP concerned with processing text outside controlled language domains, linguistic concepts and datasets for processing natural languages are regularly exploited for new applications. In contrast, the knowledge of language processing that is more directly based on cognitive psychology, has remained largely unexplored.

The aim of this thesis is to contribute to the process of adapting behavioural data derived from cognitive processes during reading for improving NLP systems.

The challenge lies with the above mentioned difference in research focus. Concretely, existing measures and theoretical conceptualisation of cognitive-behavioural data were designed to address questions about *how the human cognitive apparatus* *functions*, typically by dissecting statistically perceptible effects of controlled external stimuli into distinct measurable response variables as evidence towards a specific hypothesised underlying cognitive organisation.

In contrast, the central questions within NLP research are more commonly addressing the engineering problems of *how to automate a process which would otherwise require human cognitive processing*, and further, how to automate it well enough to be able to apply the proposed automated solution to data of unknown variability, popularly referred to as operating "in the wild".

Drawing on an example from computational psycholinguistics, the following passage describes how core NLP tools, syntactic parsing and language modelling, have underpinned testable competing models of human sentence processing (Demberg, Keller, and Koller, 2013; Frank and Bod, 2011; Keller, 2010; Mitchell et al., 2010).

In these experiments, statistics derived from language models and syntactic parsers, tools which were originally developed to exploit computational, statistical and mathematical principles for automatically generating sentences and replicating syntax-annotations, are instead supplied as predictors of reading times. The goal of the experiments is to measure the predictive power of the system-derived statistics as *the amount of observed variance in reading times explained* after controlling for other known correlates. Comparison of the competing models' respective predictive power then serves to test concrete hypotheses about human cognition. Specifically, they test whether human reading must be sensitive to abstract hierarchical syntactic structures or may plausibly get by relying only on more primitive co-occurrence patterns of words or word classes.

Importantly, the parsers and language models of those studies were adapted for psycholinguistic experiments. This was achieved by adjusting which features and computational resources the models had access, effectively serving to mimic the known limits of humans' physiological, attentional and computational resources.

The point of the above example is that when adapting language technologies for experiments in cognitive psychology, the technologies must first be constrained to rely only on plausibly human-like input and computational resources.

In contrast to the above example of how core NLP tools can be re-purposed to emulate cognition, the work presented in this thesis takes a parallel but opposite route. Here, I seek to identify where psycholinguistic insights and data sources are likely to be able to alleviate current limitations of NLP tools.

1.3 CONTRIBUTION

This thesis contributes empirical evidence of direct influence of text complexity on readers' gaze behaviour, and approaches to taking advantage of this influence for improving NLP systems. For this purpose, a range of strategies for representing gaze behaviour is presented along with assessments of their individual applicability to specific NLP tasks.

The primary contribution is to show that goals and constraints of NLP system evaluation and development can be successfully met when using gaze data representations optimized for detail retention and Machine Learning (ML) architectures while setting aside the goals of the psycholinguistic literature.

Specifically, Chapters 4 and 5 support the claim that gaze measures can distinguish types of complexity specific to the NLP tasks of Automatic Text Simplification (ATS) and Machine Translation (MT), while Chapter 6 finds that gaze data can improve a sentence compression system, and Chapter 7 show that lexical complexity is detectable from individual fixations as represented by their immediately surrounding gaze behaviour.

These results demonstrate a potential for developing evaluation procedures based on functional cognitive feedback in NLP tasks where formal solutions can not be derived.

To encourage further research and promote comparability and replicability, the three datasets of gaze recordings collected as part of this work are made available.[2]

1.4 STRUCTURE OF THE THESIS

The remaining chapters of Part I introduce the concepts *text complexity* (Chapter 2) and *gaze data representation* (Chapter 3) in terms of the particular tasks and methods employed in the four empirical studies reported in the research papers making

[2] Study 1: bitbucket.org/klerkes/study1/, Study 2: bitbucket.org/klerkes/study2/, Study 4: bitbucket.org/klerkes/study4/

up Part II. The introductory chapters and empirical studies together form the basis for the general conclusion and perspectives drawn in Part 8.

2 TEXT COMPLEXITY

This section introduces the concept of *text complexity*. The concept is operationalized within the concrete contexts of the empirical work presented as articles in Part II.

The concept of text complexity has no comprehensive, precise, stable, or otherwise widely agreed-upon definition, theoretically or empirically, even as it has severe, widely agreed-upon impact on large groups of human readers. This is the background for choosing a narrow notion of text complexity, i. e. as tied to the empirical work rather than presenting a comprehensive overview.

In particular there is no established mapping between the two main perspectives encountered in this work, namely the view of text complexity as pertaining to features of the text, as held within the NLP-community, and that of text complexity as pertaining to features of the reader's cognition, as held within cognitive psychology.

This lack of alignment is not a case of directly opposing or necessarily conflicting views, but one of disjoint goals: In NLP, where the goal concerns optimizing text, the focus is on aspects of the text that can be optimized. Similarly, in psycholinguistics, where the goal is to test hypotheses about human cognition, this is reflected in an operationalisation tuned to measures sensitive to variation in cognitive processes.

The former of these views thus tends to posit that all texts have an inherent complexity level, disregarding any systematic individual variation in human reading behaviour. The latter view, in contrast, builds on the assumption that the systematic variation in cognitive responses can be reliably provoked when salient text traits are sufficiently controlled.

The current chapter takes its outset in the NLP-oriented perspective on text complexity as something pertaining to text traits—tied to the operational unit of a given system. In contrast, the cognition-oriented view plays a more prominent role in Chapter 3 on how readers' gaze may be represented to capture those human-centered aspects of text complexity. To accommodate this focus on human cognition, that chapter is structured around the concrete gaze representation strategies employed in the empirical work presented in Part II.

Text complexity is thus mapped out in the following by describing how it relates to each of the specific tasks tackled in this work, namely lexical complexity detection, sentence compression, text simplification, summarization and machine translation.

The first focus of this chapter is to the low-level, versatile task of sentence compression. This task can in principle be considered a special instance of both of the two tasks of simplification and summarization which are described second and third. The fourth task, MT, necessarily requires a translation step which will not be considered directly in this work. However, as a generative system, MT must address textual aspects that directly influence text complexity. The chapter ends with a description of the task of lexical complexity detection which is not a standard NLP task, but is the task that closest addresses the reasearch problem of this thesis.

2.1 SENTENCE COMPRESSION

Sentence compression as a task is concerned with providing a shortened sentence which conveys the most important information content of an original, longer sentence. This occurs for example in subtitling, summaries and abstracts (Knight and Marcu, 2002) and is thus a relevant, if not necessary, subtask of three of the following tasks described, namely simplification, summarization and translation.

Here, compression is used in Studies 1 and 3 in the context of simplification and summarization, but note that the motivation for doing automatic sentence compression ultimately depends on the higher-level task it is solving.

Under the broad definition given above, the space of possible solutions is infinite. Approaches to tackling the full range of permissible transformations including lexical substitutions, paraphrasing, sentence splitting and other syntax operations have been proposed (Cohn and Lapata, 2008; Heilman and Smith, 2010; Rello, Pielot, et al., 2013; Woodsend and Lapata, 2011a, i.a.).

Here however, the focus is on the restricted task of selecting an ordered subset of tokens from the original input i. e. sentence compression by word deletion (McDonald, 2006).

Sentence compression as deletion initially put primary weight on ensuring grammatical output by relying on syntax trees and formulating objectives and scoring functions in terms of tree-operations (Cohn and Lapata, 2009; Filippova and Strube, 2008; Knight and Marcu, 2002; Riezler et al., 2003).

By using different model architectures and objectives later approaches have allowed an increasingly distant reliance on syntax annotations. Thus, by formulating the problem as a sequence labelling task where each token is considered along with features extracted from syntax annotations, it is left for the learning algorithm to discover whether there is statistical evidence for relying on those syntax-informed features (Elming et al., 2013; Filippova, Alfonseca, et al., 2015; McDonald, 2006). This is the approach taken in Study 1.

The move away from syntax-based constraints does not entail that the field cares less about producing grammatical output, but instead reflects two important challenges related to the limitations of syntactic parsers.

The first challenge is that parsers' error-rates grow with sentence length as the syntactic variation grow in terms of available lexical and syntactic choices. Consequently, the observed statistical evidence for individual syntactic patterns grows sparser (McDonald and Nivre, 2007).

The second challenge is that parser performance declines in languages and even domains outside the richly resourced area of English news text. This effect limits the benefit of relying on parsers for ensuring grammatical output (Tsarfaty et al., 2013).

In recent work by Filippova, Alfonseca, et al., (2015) and in Study 3 the sequence-to-sequence model for compressing sentences is implemented as a Recurrent Neural Network (RNN) with Long-Short-Term Memory (LSTM) cells. These fully connected neural network models with hidden states allow for new model designs (Goldberg, 2015). In Study 3, the model uses gold-standard syntactic information as a regularizing mechanism on the hidden layers during training. This approach effectively removes the reliance on automatic parsers. In fact, in the study by Filippova, Alfonseca, et al., (2015), no significant difference was found between a model entirely without syntax-derived features and the one trained with syntax information.

In these models, words are represented by dense word-embedding vectors reflecting their distributional semantic relations, and the RNNs are tasked with learning what role context should play.

2.1.1 *Text complexity in sentence compression*

Text complexity in all of the above mentioned models most saliently relates to how frequently a given input-output pair is observed in the data. This frequency is the upper limit on how strong statistical evidence may be obtained for a decision to delete or keep a word in a given context.

What empirically gets to be considered text complexity in these models is therefore a function of what information is represented. That is, what units are exposed to the decision function and the co-occurrence patterns of these units; e.g. sub-trees, sentence chunks, sub-strings of words such as stems, suffixes and affixes, PoS tags, and corpus-derived information about the word's usage, such as embedding vectors.

Whichever traits or aspects of texts are most prominently represented in model and feature design will be what gets to implicitly define text complexity as *anything but the most commonly observed patterns* of those textual aspects.

For instance, syntax-based representations thus reflect rare syntactic patterns as complex; lexical features tend to reflect word-frequency as a the primary driver of text complexity. If both lexical and syntactic features are represented, combinations of them can form patterns that, dependent on the frequency with which the patterns are observed, may be considered concrete templates for complex text. Ideally, such a combined representation of the syntactic and lexical level would allow a model to recognize the most common proverbs as less complex than either their possibly antiquated words or outmoded syntactic realisation would indicate by themselves.

Finally, the role of text complexity is shaped by the datasets used for training and testing compression models; in supervised learning, exemplary parallel data of input texts and target output texts defines the concrete learning examples that must be generalized. The model therefore learns to replicate as much as possible of this observed *compression behaviour*. This is achieved by assigning the most deciding power to the patterns in the input text representation that best predict which deletions were performed in the exemplary compressions, regardless of whether these compressions served to reduce text complexity.

In other words, the most direct influence on whether the compression output is actually less complex than the input, by any measure, is the data that the model saw during training; if the exemplary compressions were less complex than the inputs—and the model learned to replicate them well—the model will arguably have learned to reduce text complexity.

In the following sections on simplification and summarization it is further discussed which characteristics of the data available at training time that may affect the complexity of the output of a trained model.

2.2 AUTOMATIC TEXT SIMPLIFICATION

In the field of NLP, no shared definition of the task of text simplification exists. However, as in the case of sentence compression, a number of modelling approaches and datasets form some clusters of similar conceptualisations and definitions of the task of automatically simplifying texts.

While human editors may choose to simplify a text at any level from lexical choice to argument structure in order to meet a target reader's needs, it is commonly held that automated systems can still provide useful, simplified output. The aim is then to mimic a limited subset of edit-types, such as by doing either paraphrasing, sentence splitting, lexical substitution or word deletion.

Because the individual edit-types are observable in isolation in human produced simplifications (Amancio and Specia, 2014) it is assumed both that each edit-type must be helpful by itself and that systems performing specific edits could be applied in succession to closer mimic the varied operations present in professional simplifications.

Motivations for Automatic Text Simplification (ATS) vary widely. For example one motivation is a democratic aspiration to make information available to as many readers as possible. A different motivation focuses on servicing particular groups of weak readers who face specific challenges. A third motivation, however, targets NLP pipelines, alleviating steep performance drops in downstream systems by reducing upstream text complexity (Siddharthan, 2014).

In order to illustrate how the task of text simplification is delimited in practice, the following describes the task parameters that influence the conceptualisation of text complexity and how they shape the outcome.

2.2.1 *Text complexity in simplification*

In particular, the choices of data, model and evaluation design define the target conceptualisation of *simplified text*. Notably, the influence of those choices are present whether left implicit or explicitly addressed by the researcher.

DATASET As mentioned in the context of compression models, in supervised learning designs the observable difference between the source and target texts ideally should embody the researcher's definition of text simplification as closely as possible, since this limits what the model can learn.

Focusing on simplification datasets, they may reflect more or less consistent adherence to rules and heuristics depending on how the exemplary training output was obtained: If it was created by humans; what task formulation was given, or if it was automatically assembled; what algorithm was used. Such rules and heuristics can in turn reflect beliefs about how text traits affect the readers' cognition, e. g. when an editor takes children's vocabulary into account by using word lists annotated for age-of-acquisition when simplifying into children's texts.

Several commonly used datasets are sentence-aligned parallel corpora of original source texts and simpler target texts of encyclopaedic resources such as the Britannica corpus (Barzilay and Elhadad, 2003) and the SimpleWiki corpus (Coster and Kauchak, 2011) and, as is typical in NLP, newswire (François and Miltsakaki, 2012; Klerke and Søgaard, 2012; Rauzy and Blache, 2012; Xu, Callison-Burch, and Napoles, 2015, i.a.). In these corpora, a source version and target version of the texts are paired and sentences with substantial overlap are aligned. When the corpus is harvested from existing resources, the source-target pairing is based on topic and the complex/simple relation is assumed by virtue of the texts being originally edited for different target populations, i. e. skilled readers vs. children, foreign language learners or people with impaired reading skills.

An example of how datasets contribute to the implicit definition of text complexity in an ATS model is seen in Study 1 of this thesis, where the encyclopaedic SimpleWiki was found to contain almost no sentence pairs that fit the definition for sentence compression by deletion. In contrast, the newswire-based

Danish simplification corpus used in the reported experiments, displayed enough of these pairs for training a sentence compression model to perform simplification by deletion.

This difference between corpora stems from their design. The articles in English Wikipedia and Simple English Wikipedia which form the two sides of a sentence pair in SimpleWiki have most often been written independently (Amancio and Specia, 2014; Coster and Kauchak, 2011; Woodsend and Lapata, 2011b; Xu, Callison-Burch, and Napoles, 2015). In contrast, the article pairs in the Danish simplification corpus were produced by direct editing of the source newswire with the aim of providing a concurrent online news outlet to a broad group of less-skilled readers of Danish (Klerke and Søgaard, 2012), resulting in a large overlap between source and target texts.

MODEL Both the choice of model and of how the text is represented in the model limits what simplification operations can be learned.

For example, it is unlikely for a model to learn a strategy of keeping the subject and main verb of sentences if it does not have explicit access to syntactic function annotations but instead relies on patterns in alternative available features as a proxy. For instance, such a model may learn to retain individual words that happen to exclusively occur in one of these functions—such as pronouns—but then fail to retain words that only intermittently appear as subject or main verb.

Additionally, the choice of model architecture defines what patterns may be recognized and operated on. For example, by definition the compression model described above where words in a sequence receive a binary tag each indicating whether they should be deleted, does not perform paraphrasing or re-ordering operations regardless of how many of these operations are present in the training data.

An alternative modelling approach casts simplification as a translation problem (Coster and Kauchak, 2011; Filippova and Strube, 2008; Klerke and Søgaard, 2012; Specia, 2011). In this approach however, central heuristics of the translation models can shape the implied definition of text simplification in an inadvisable way. Examples of such heuristics include when the model is rewarded for keeping the output length similar to the input length or penalised for leaving words untranslated.

As was the case with sentence compression, ATS models have also been implemented as rule-based manipulation of syntax-trees and lexical substitution based on word lists (Heilman and Smith, 2010; Siddharthan, 2006; Woodsend and Lapata, 2011a; Zhu, Bernhard, and Gurevych, 2010). As the compression models based on syntax, these approaches are vulnerable to the availability and reliability of parsers and other domain- and language-specific resources.

EVALUATION Finally, any system evaluation delimits which of a model's choices are counted as positive or negative instances of text simplification. This happens through the test set, through the choice of evaluation metric and through any instructions provided to human evaluators. However, automatic evaluation is often largely dictated already by the dataset, representation and model choices.

Whichever approach is taken, the concrete implementation guides the selection of automated measures for model learning and evaluation. To give an example, an approach based on classification, such as the compression model which tags words to delete as described above, would favour the class-wise F1-score which balances precision and recall. Alternatively, raw accuracy can be meaningful with classifications, whereas BLEU (Papineni et al., 2002) or Rouge (Lin, 2004), based on n-gram overlap, are preferred for models adapted from MT or summarization.

Note however, that in any case, given the one-to-one parallel datasets, the automatic evaluation will suffer from the bias of only having one reference solution and potentially falsely penalise good alternative solutions.

While traditional readability metrics such as Flesch-Kincaid grade level (Flesch, 1948) or LIX (Bjornsson, 1983) are sometimes reported, these metrics *apriori* assumes grammaticality, fluency and adequacy by design and therefore may be uninformative in the context of machine-edited text.

More commonly, some combination of sentence fluency, grammaticality, appropriateness and readability are evaluated subjectively on a Likert-scale by human judges. This subjective evaluation practice is designed to exploit readers' intuitions about language and thereby sidestep the problem of providing a definition of fluency, grammaticality, appropriateness or readability. While popular, this approach suffers from the important weaknesses of being vulnerable to random and antagonistic annotators (Hovy et al., 2013) as well as cognitive biases (Sackett, 1979). The relevant biases include both the observer effect and compliance bias where participants are biased to adapt their judgements to comply with the experimenters' own biased expectations.

Moreover, in the case of subjective readability estimation at the level of individual sentences or words, the subjective evaluation approach suffers under the inability of humans to consciously monitor their own automatic cognitive processes.

In addition to these general approaches to evaluating system output, there is a persistent awareness within the field of ATS of the target users' particular profiles; even relatively low error rates can be prohibitively obstructing to struggling readers. There is an active and general discussion within the fields of ATS and readability assessment on whether the evaluation methodologies available are reflecting aspects of relevance to particular users and their teachers (François and Miltsakaki, 2012; Rello, Baeza-Yates, et al., 2013; Siddharthan and Katsos, 2012; Vajjala and Meurers, 2013). This awareness of the need for reliable automatic simplification and readability assessment is the concrete basis for the core problem in this thesis; that this area of research could be substantially advanced by achieving a better sensitivity to text complexity in models and evaluation.

In conclusion, text complexity in ATS is explicitly and implicitly defined in distinct design choices; through the choice of data, defining complexity as the observable changes, through the choice of data representation and model constraints, defining the scope of permissible and reachable simplification operations, and through the choice of evaluation criteria, defining which operations are rewarded via metrics and rating dimensions.

2.3 SUMMARIZATION

The goal of summarization is to provide a shortened version of long texts to skilled readers. Systems compete to reach an optimal trade-off between the density of relevant information and text coherence. Summarization systems are either extractive at the sentence or word level, where models may rely directly on sentence compression, or abstractive where models may incorporate paraphrasing and sentence splitting to save space. Additional strategies like named entity recognition, anaphora resolution and discourse modelling may also be leveraged to optimize the coherence of the output summary.

Evaluation of summarization systems include comparison to reference summaries with the Rouge family of metrics (Lin, 2004) and subjective evaluation on dimensions such as informativeness and coherence.

2.3.1 *Text complexity in summarisation*

The most prominent differences in how text complexity is handled in summarization compared to ATS stems from the target user group; skilled readers. This assumption about the skill level of the audience makes it a viable strategy to defer as much as possible of the responsibility for making sense of the summary to the reader.

Consequently, the role of text complexity is at most an indirect one; When a piece of a summary becomes too complex to make sense off, e. g. from a critical lack of coherence, the particular text piece flips from making a positive to a negative contribution towards the objective of supplying a maximally informative summary within a limited summary length.

This leads to a conceptualisation of text complexity primarily as a threshold phenomenon; disregarding subtle variations, but keenly enforcing the border between acceptable and unintelligible.

The task of summarization is peripheral to the empirical work presented in this thesis and connected only through the sentence compression experiments in Study 3 which are based on sentence compression datasets intended for summarization. This section therefore primarily serves to highlight how the difference of motivation compared to ATS form a different conceptualisation of text complexity.

2.4 STATISTICAL MACHINE TRANSLATION

The goal of Machine Translation (MT) is, widely scoped, to provide information in a target language from input in a different source language. Unlike the previous tasks, MT require systems to generate entirely new text in the target language from the source language input.

At the core of statistical MT is the reliance on statistical relationships between co-occurences of words both between the source and target languages—as captured through word aligments—as well as within the target language—from language modelling.

Wherever machine learning is applied and rely on co-occurrence patterns, the quality of the output drops when the observed input deviates from commonly observed input; simply because the statistical evidence available for how to handle rare events per definition is sparse relative to that for commonly observed events.

Input text length is another important influence on output quality of MT systems, as every increase in input length exponentially increases the size of the search space for candidate solutions.

2.4.1 *Text complexity in machine translation*

Given the above arguments and the previously discussed extent to which text complexity correlates with word frequency and sentence length, it is unsurprising that MT systems in general perform worse on more complex input text compared to when the input is simpler.

For output text, notwithstanding the challenges mentioned earlier on using MT models for ATS, the underlying statistical machinery inherently has a tendency to favour slightly less complex output text. This happens entirely as an effect of preferring solutions with the strongest possible statistical support, i. e. the most frequently observed patterns of text.

Still, as was the case in all of the above tasks, a model learning from data will be biased by design towards an output level of text complexity similar to that of the exemplary output provided during training and the corpus underlying any built-in language model.

The experiments presented in Study 2 use and compare translations produced by both expert translators and a generic statistical MT system. Because this thesis presents no MT system implementation, this section focuses solely on how text complexity in the input and output text is handled while details about how translation is automated are omitted.

2.5 LEXICAL COMPLEXITY DETECTION

The task of detecting complex words in context—as introduced in Study 4—is not a standard NLP task. It is designed specifically with the purpose of testing the hypothesis that *gaze behaviour* reflects text traits associated with complex text systematically and in near-real-time. This section highlights how the task formulation operationalizes text complexity.[1]

In the absence of a comprehensive operational definition of text complexity, this task narrowly identifies complex words by proxy of a single or a few salient text traits. These proxy-traits of text complexity are word frequency, word length, spelling deviation and genre-specific visual marking.

The criteria for choosing these proxy-traits of complex text are that they are generally associated with readability, are expected to affect gaze behaviour and can be unambiguously annotated automatically.

Note that because these traits are all easy to annotate automatically, they are already also easy to detect and target in NLP applications. For this reason, the task as formulated here has no motivating application outside this work.

2.5.1 *Rationale for the task of lexical complexity detection*

The goal is to test the hypothesis that readers react systematically when encountering complex text.

In order to bypass the *hard problem* of annotating the cases of complex text that are arguably difficult to identify automatically—or even agree upon in the absence of a comprehensive definition—the strategy employed here is to start from a superficially narrow definition of text complexity, choosing only the *easy cases* of complex words.

The underlying assumptions warranting this reduction of the problem is that eye movement strategies in reading are

1. sufficiently few and general that the gaze reaction resulting from encountering a complex word may be the same regardless of whether the complex word belongs to the narrow class of easily annotated lexical complexity, and

[1] The task is described in detail in Chapter 7.

2. sufficiently distinct that gaze reactions to the target words are discernible from reactions to words with irrelevant traits (e. g. reactions to words with emotional content), and

3. sufficiently strongly expressed to be reliably detected by the gaze recording equipment.

Under these assumptions, being able to detect obvious cases of lexical complexity directly from readers' systematic automatic gaze response patterns is the first step towards being able to identify the hard cases of complex text. This work thus represents a first step towards a comprehensive definition of text complexity in NLP as one pertaining to the reaction of the user.

3 GAZE REPRESENTATION

All of the empirical work in this thesis relies on gaze data. The proposed methods for using this data source constitute the main contribution of this thesis.

The goal of this chapter is to bring together the individual approaches to representing gaze data as presented in the four empirical studies in order to provide a unified, coherent perspective on this aspect of the work.

Each of the sections of this chapter aligns the gaze representation of a study from Part II to the levels of gaze data representation illustrated schematically in Figure 1, except for the minimal primer on eye movements in reading following immediately below.

3.1 EYE MOVEMENTS IN READING

This short introduction uses numbers and facts from the comprehensive textbooks by Rayner, Pollatsek, et al., (2012) and Holmqvist et al., (2011).

We say readers scan a text as their gaze moves over it. More precisely, the scanning happens in tiny steps, so-called *fixations*. The steps do not follow the word order of the text exactly; sometimes words are skipped and sometimes previously scanned parts of the text are re-fixated. A sequence of fixations constitutes a *scanpath*, illustrated both as level 3 in Figure 1 and in Figure 2.

The text rendered in sharp focus in the fovea of the eye's retina typically spans five to nine characters, asymmetrically favouring the text ahead.

A *saccade* is the brief moment when the eyes are moving between fixations. During the saccade the reader is effectively blind. In this timespan the eye muscles adjust the eyeballs to the new fixation target. The target is selected in advance and encoded into what is termed the eye motor program as part of the signal to execute the saccade. Eye motor precision tends to drop with the distance between fixations why short remedial saccades frequently occur after long saccades.

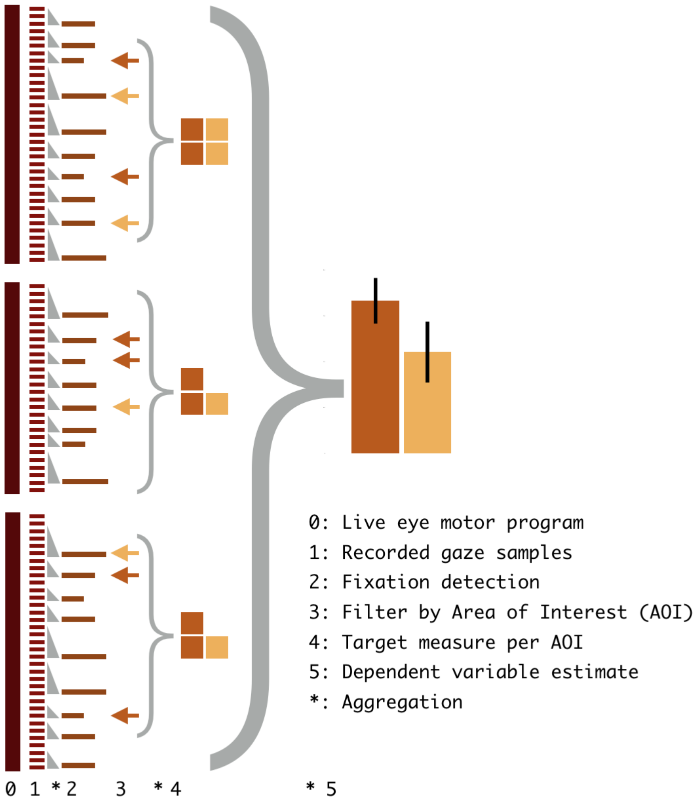

0: Live eye motor program
1: Recorded gaze samples
2: Fixation detection
3: Filter by Area of Interest (AOI)
4: Target measure per AOI
5: Dependent variable estimate
*: Aggregation
0 1 * 2 3 * 4 * 5

Figure 1: Schematic illustration of levels of gaze data representations of three independent readings. Grey indicates aggregating operations, yellow and orange (at level 3–5) are representations informed by experimental choices (e. g. linguistic theory).

Fixations can in principle last for any amount of time. It has been shown that only about 50–60 ms are necessary for sampling the visual impression of the part of a text in focus. Fixations in reading usually last three to six times that duration and vary systematically with a number of textual aspects. Aspects of the currently fixated word have been shown to be most influential. This immediate relationship between textual features and fixation behaviour is taken as evidence of a close eye-mind link in reading.

From a high-level perspective, eye movements in reading can be considered a measurable by-product of the specialised cognitive process of skilled reading. During uninterrupted reading this process controls the two main parameters of the eye motor program; the timing and guiding of fixations. The eye movement program execution is a continuous process as illustrated in Figure 1, level 0.

An eye tracking device captures snapshots of this process (Figure 1, level 1). It applies a geometric calculation based on a recorded picture of the eyes. Hereby the coordinates on the stimuli display that fall directly in the line of sight are recovered and recorded as a raw sample at a fixed frequency practically in real-time while the image of the eye is discarded.

A fixation detection algorithm (level 2) is applied to cluster raw recorded samples of coordinates in space and time into fixations and saccades based on their euclidean distance. This procedure forms a sequence of distinct, ordered fixation coordinates and durations which together constitute the scanpath. Fixation detection algorithms exploit basic physiological and cognitive constraints to reconstruct, with high accuracy, the gaze event sequence as it was executed (Holmqvist et al., 2011).

The levels 3 and 4 in Figure 1 depend ultimately on stimulus design. The definition of what concrete areas on the stimulus display form meaningful units under one's hypothesis, i. e. the choice of Areas of Interest (AOIs) along with which measures to use, is thus determined as part of the experimental design. This means that these two levels are also the levels where linguistic theories are usually applied.

At level 5 the analysis design determines how the aggregation of individual recordings into an estimate of some dependent variable of interest is performed. This is also typically the level where hypothesis testing (concerned with the theories applied at the two previous levels) is performed. As such, this level is directly shaped by the choice of statistical testing paradigm.

Below it is further described how—at the expense of the traditional hypothesis-testing framework and the familiarity and comparability afforded by the higher, aggregating levels of gaze representation—NLP system evaluation and development can benefit from keeping the richness of detail that is present at the lower levels.

3.2 NLP SYSTEM AS EXPERIMENTAL CONDITION

The ubiquitous and direct influence that other branches of linguistics have had on NLP development (see for instance Bender, 2013) stands in sharp contrast to the lack of influence psycholinguistics have had on NLP.

In psycholinguistics, the study of eye movements in reading is largely concerned with effect sizes recorded on a small set of well-established gaze measures. Effects are measured by comparing distinct experimental conditions. Most often comparisons are between small closed sets of conditions. Effects are often sought replicated in series of studies with small variations. This methodology is used to incrementally build support for increasingly comprehensive models of cognition. In addition, this course of development has led to devising ingenious experimental procedures for teasing apart ambiguous evidence in corner-cases. However, NLP systems only sporadically incorporate insights from this rich body of knowledge.[1]

The only salient point of departure taken to leverage psycholinguistic insights in the context of NLP research is seen within system evaluation, such as in MT in the work of Doherty, Kenny, and Way, (2012), Specia, (2011), and Stymne, Danielsson, et al., (2012). In those approaches to system evaluation, NLP systems plug conveniently into the methodology of measuring and comparing effect sizes by simply considering each system a separate experimental condition. By providing groups of human subjects with a task to perform on a set of system outputs, researchers can estimate and compare systems' effects on the human subjects' behaviour.

In Study 1 a sentence simplification system is evaluated both automatically, subjectively and by gaze recording. This study follows the approach outlined above to using gaze for system evaluation. The experimental conditions correspond to the kinds of text that readers were presented with, namely the *original source input*, *expert-simplified target output* and post hoc split categories of syntactically *acceptable* and *unacceptable system outputs*.

The representation of gaze behaviour also follows the psycholinguistic literature. The study assesses estimated measures based on aggregates of the duration and number of fixations per sentence and per word and of the proportion of regressing fixations. That is, the gaze representation and analysis proceeds through all five levels illustrated in Figure 1.

[1] Only five papers were accepted for the (so far only) workshop on Eyetracking and Natural Language Processing (Carl, Bhattacharya, and Choudhary, 2012), several of which propose uses of NLP for psycholinguistic research while none propose the opposite.

This choice of how the gaze data is represented and analysed also shapes the results. In Study 1 the gaze measures show significant differences in readers' reactions to the kinds of sentences, even when comparing to the sample of system output with unacceptable syntax with a small sample size of just 27 sentences.

A further encouraging result is that the two conditions found to produce the most similar gaze behaviour are the expert-simplified sentences and the sample of system outputs with acceptable syntax. Because the text simplification system was trained to mimic precisely the expert-generated sample, this result supports the hypothesis that gaze data does distinguish between variations in text in a way that is meaningful for NLP system development.

A notable limitation is that the interpretation of these results is complicated by the lack of control over confounding factors such as sentence length. In addition, the need to set up stratified Latin squared stimuli presentation to ensure that every participant only sees the same sentence in one version turns the evaluation procedure into a complex experiment by itself. Moreover, the stratification severely limits the statistical power of the evaluation experiment.

Yet, the results of Study 1 indicate a potential for readers' gaze behaviour to help distinguish between text of varying complexity; whether the complexity was introduced by the ATS system or was present already in the human writers' original texts.

3.3 SCALING TO MORE SYSTEMS

Building on the encouraging results with the experimental design in Study 1, Study 2 scales up the number of comparisons with five versions of each text. This experiment thus aspires to test whether differences in translation quality are also distinctly reflected in readers' gaze. Furthermore, the study investigates how the variation in gaze behaviour produced by differences in translation quality compares to the gaze behaviour variation stemming from reading in different languages.

The study uses logic puzzles which participants are asked to read, solve and rate.

Similar to Study 1, the evaluation methods used for comparing the five types of text that constitute the experimental conditions are automatic evaluation, subjective quality judgements and gaze measures. In addition, Study 2 includes *task efficiency* which takes both puzzle solving time and success into account.

This metric serves as a measure of text usability inspired by Castilho et al., (2014) and Doherty and O'Brien, (2014). The intuition is that the best text is that which will allow participants both to solve more tasks and to solve them faster. Arguably, making readers able to make the most efficient use of a text is a shared underlying goal in much of NLP. On this account, efficiency is considered the *true* evaluation in Study 2.

The results indicate that efficiency, the usability metric, correlates significantly only with gaze measures, while neither subjective assessments or the automatic BLEU scores are found to reflect this measure of usability.

In parallel with the previous study, the results of Study 2 support the hypothesis that gaze behaviour is adequately expressive for NLP system evaluation. However, with the increased complexity of the comparison, the challenges of setting up a valid experimental design, the problem of weakened statistical power, and the challenges of interpreting the results also become substantially more difficult.

Thus, an important finding resulting from the added complexity of comparing more systems, is that the method of plugging NLP system evaluation into the psycholinguistic experimental framework does not scale well.

3.4 A LOWER-LEVEL REPRESENTATION

In contrast to the above described approach to gaze representation, Study 3 represents gaze corresponding to level 4 of Figure 1; using a single measure registering gaze behaviour per word per reading, but without aggregating the individual readings further.

The study draws inspiration from the successful use of aggregated gaze data for PoS-tagging and dependency parsing in the work by Barrett and Søgaard, (2015a,b), as well as the recent powerful developments in sequence-to-sequence deep neural network models in NLP (Goldberg, 2015, i.a.).

Study 3 uses a neural network model trained on multiple tasks, including gaze prediction, to do sentence compression. Thus, instead of using gaze data for system evaluation, the goal is to make information about eye movements available to the NLP-system at training time. This setup provides a test of the hypothesis that gaze behaviour captures information about the text that can somehow boost task performance.

The experimental design of this study conveniently permits the use of data from an existing large corpus of gaze recordings with the only requirement that the vocabulary is substantially overlapping between the training tasks.

The idea is to force the neural network model not to overfit the training data from the compression task by using gaze prediction as an auxiliary task during training. Moreover, including all readings of each individual sentence as independent training instances means that the gaze data itself is hard to overfit because the same sentence is likely to be observed with a unique sequence of gaze behaviour for each participant. Since the auxiliary task is to predict the gaze representation, this means that the same input sequence of words appear 10 times in the entire dataset, potentially with a different label sequence each time.

Under this design of the auxiliary task of gaze prediction, the model only benefits from making strong predictions for words that are looked at in much the same way by several readers.

The study presents results using two different measures, an *early* and a *late* measure. These measures split the gaze data into two distinct representations, respectively catching time spent on first encounter with each word and time spent re-fixating words.

The intuition behind filtering on this dimension of the gaze behaviour is that the latter, while sparse, is considered to be less predictable. This decrease in predictability happens because the late measure is more likely to reflect cognitive processes that are not as specifically tied to local parts of the text, such as resolution of ambiguities and integration with previous information. These are also processes that are expected to depend more on subjects' background knowledge and preferences. In contrast, the early measure reflects the immediate, automatic processing (Rayner, Pollatsek, et al., 2012).

The model architecture incorporating gaze prediction as a regularising device, was found to be successful at improving sentence compression by deletion. This held true over a selection of different compression datasets, which indicates this architecture to be a robust approach to learning deep RNN compression models from much smaller datasets than what has previously been viable.

Representing gaze as arbitrary relations between parallel sequences of words and labels is akin to a general tool; it simply plugs into the multi-task learning setup to provide a regularising signal. Whether the representation will support learning a given task of interest beyond the regularisation is an open question.

Thus, the limitation of this approach is that it does not directly provide insights into whether the regularising power from simultaneously predicting gaze behaviour is somehow allowing the network to identify text complexity; only that it facilitates learning to delete words for sentence compression.

So while this is a simple and scalable use of gaze data for improving NLP systems, a substantial amount of trial-and-error with different tasks and datasets is needed to form a reliable intuition about when and how information from gaze prediction may benefit various NLP tasks.

3.5 REPRESENTING SCANPATH SNIPPETS

The gaze data in Study 4 is represented in two ways. While the first representation is used for calculating effects for comparison with existing datasets (fully aggregated, corresponding to level 5), the second representation discards both of the two higher levels of aggregation (corresponding to level 3 in Figure 1).

This representation, the scanpath, is recognized as a close trace of the eye motor program (Holmqvist et al., 2011).

The primary challenge of using scanpaths to represent reading stems from the need to compare the structures in a meaningful, reproducible way. Formally, that problem can be cast as a special instance of graph matching or string matching.

The challenge is easily illustrated by the two stylised scanpaths in Figure 2. Deciding which pair of saccades are counted as similar between the two depicted paths will depend entirely on which descriptive parameters one chooses to focus on; i. e. choosing saccade starting point, ending point, direction, distance travelled, fixation duration or cardinality or any combination of these may affect the answer.

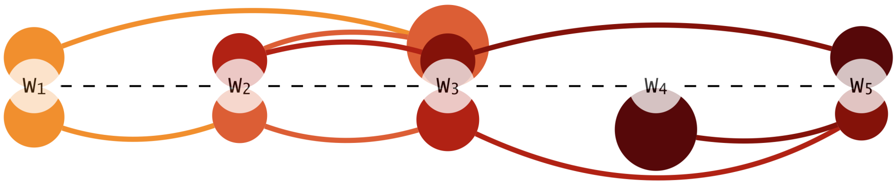


Figure 2: Two stylised scanpaths on a sequence of words. Saccades are represented by connecting lines and fixations by circles with duration encoded by diameter and order of execution encoded by hue.

Note that determining the role of each of these descriptors within a model of cognition is a daunting task, logistically as well as in terms of theoretical work required.

Von der Malsburg and Vasishth, (2011) review previous approaches to using scanpaths within psycholinguistics. While other metrics have mainly relied on versions of the Levenshtein edit distance (Levenshtein, 1966), Von der Malsburg and Vasishth, (2011) proceed to propose Scasim, a scalar metric of pairwise scanpath similarity. The metric first produces a vector based on an elaborate calculation designed to capture the similarity per pair of fixations in a fixation-aligned pair of scanpaths. To arrive at a scalar value, the vector's dimensions are summed, resulting in a lower score for more similar pairs of paths and for paths with fewer fixations.

However, while it is likely that the Scasim or another scanpath similarity metric would be sensitive to text complexity, the assumptions built into the per-fixation similarity calculations and the inescapable reliance on scanpath alignment makes these measures prohibitively inflexible as basis for representing scanpaths in NLP systems.

Instead, the representation of scanpaths in Study 4 exploits the fact that ML models in general benefit from training on large datasets with rich feature vectors. In addition, these models are robust to the noise that aggregate measures are designed to reduce.

The main experiments in Study 4 therefore represent each fixation as one datapoint. This means that one full scanpath corresponds to a sequence of datapoints. Each datapoint is given a binary label denoting whether the fixation landed on a complex word.

From the perspective of a traditional psycholinguistic experimental framework, these labels correspond to the experimental conditions; complex word or not. However, from an ML perspective, this label design can conveniently be modelled as a sequence tagging problem.

In order to learn to predict the labels, each datapoint must be represented by a feature vector (a set of predictors). By forgoing on the ability to say what role each feature plays and instead rely on the model's ability to distinguish noise from patterns, a learned sequence tagging model can answer whether the feature vector presents information that is sufficiently sensitive and reliable to recreate the labels.

Recall that each datapoint in Study 4 is one fixation. The representation of each fixation is a fixed-length feature vector of information supposed to describe the current fixation in terms of the gaze data only.

These requirements fit nicely with a choice of representing information drawn from a sliding window over the scanpath around the current fixation. The information amount to 128 features detailing both distances and durations as well as how far along the word sequence the fixation has travelled from the start of the text.

The window on the scanpath that describes the current fixation covers the previous fixation and the two following ones, forming a [-1,+2] interval. This span is—given the close eye-mind link in reading—-the most likely span of eye movements to be directly influenced by the word that the current fixation is targeting. This representation can be described as a sequence of scanpath snippets or sub-paths.

The results of the experiments show that the model trained to predict lexical complexity from the scanpath snippets is in fact able to learn to recognize patterns in the gaze representation that primarily occur when complex words are fixated.

Thus, the problem of defining what should count as similar scanpaths in the context of lexical complexity has been reduced to defining a sufficiently rich and reliable feature representation.

The primary limitation of this approach lies in translating predictions obtained per fixation back to a representation of the word sequence. While words that are skipped during reading receive no predictions, words that are re-fixated receives several, which may need to be adjudicated.

A benefit of representing gaze data by scanpath snippets for learning a predictive model is that it can flexibly be adapted for learning gaze reaction patterns to other constructs of interest. This can be achieved by replacing the labelling strategy. Hereby, the described approach can help facilitate an increased exploration of avenues for using gaze reaction patterns in NLP.

Part II

ARTICLES

4 STUDY 1: LOOKING HARD: EYE TRACKING FOR DETECTING GRAMMATICALITY OF AUTOMATICALLY COMPRESSED SENTENCES

ABSTRACT

Natural language processing (NLP) tools are often developed with the intention of easing human processing, a goal which is hard to measure. Eye movements in reading are known to reflect aspects of the cognitive processing of text (Rayner and Pollatsek, 2013). We explore how eye movements reflect aspects of reading that are of relevance to NLP system evaluation and development. This becomes increasingly relevant as eye tracking is becoming available in consumer products. In this paper we present an analysis of the differences between reading automatic sentence compressions and manually simplified newswire using eye-tracking experiments and readers' evaluations. We show that both manual simplification and automatic sentence compression provide texts that are easier to process than standard newswire, and that the main source of difficulty in processing machine-compressed text is ungrammaticality. Especially the proportion of regressions to previously read text is found to be sensitive to the differences in human- and computer-induced complexity. This finding is relevant for evaluation of automatic summarization, simplification and translation systems designed with the intention of facilitating human reading.

4.1 INTRODUCTION

Intuitively, the readability of a text should reflect the effort that a reader must put into recognizing the meaning encoded in the text. As a concept, readability thus integrates both content and form.

Sentence-level readability assessment is desirable from a computational point of view because smaller operational units allow systems to take rich information into account with each decision. This computer-centric approach is in contrast to traditional human-centric readability metrics which are explicitly constructed for use at text level (cf. Bjornsson, (1983) and Flesch, (1948)) and are by their own definitions unsuitable for automatic application (cf. Benjamin, (2012) for an evaluation of readability-formula usability).

The standard approach to assessing text readability in natural language processing (NLP) is to ask readers to judge the quality of the output in terms of comprehensibility, grammaticality and meaning preservation (cf. Siddharthan and Katsos, (2012)). An alternative is to use existing text collections categorized by readability level for learning models of distinct categories of readability e.g. age or grade levels (Schwarm and Ostendorf, 2005; Vajjala and Meurers, 2014).

In this paper we seek to establish whether readers share an intuitive conceptualization of the readability of single sentences, and to what extent this conceptualization is reflected in their reading behavior. We research this by comparing subjective sentence-level readability judgments to recordings of readers' eye movements and by testing to what extent these measures co-vary across sentences of varying length and complexity. These analyses enable us to evaluate whether sentence-level simplification operations can be meaningfully and directly assessed using eye tracking, which would be of relevance to both manual and automated simplification efforts.

4.1.1 *Automatic Simplification by Compression*

Amancio and Specia, (2014) found that more than one fourth of the transformations observed in sentence pairs from Wikipedia and Single English Wikipedia were compressions. To obtain automatically simplified sentences we therefore train a sentence-compression model.

With inspiration from McDonald, (2006), we train a sentence compression system on a corpus of parallel sentences of manually expert-simplified and original newswire text where all simplifications are compressions. The system is described in detail in section 4.2.

Sentence compression works by simply dropping parts of a sentence and outputting the shorter sentence with less information content and simpler syntax. This approach allows us to control a number of variables, and in particular, it guarantees that each expert simplification and each system output are true subsets of the original input, providing three highly comparable versions of each sentence. Further the system serves as a proof of concept that a relatively small amount of task-specific data can be sufficient for this task.

Sentence compression is, in addition, an important step in several downstream NLP tasks, including summarization (Knight and Marcu, 2000) and machine translation (Stymne, Tiedemann, et al., 2013).

Below, we present the automatic simplification setup, including the parallel data, features and model selection and details on how we select the data for the eye-tracking experiment. The following section details the eye movement recording and subjective evaluation setup. Section 4.4 presents our results followed by a discussion and our conclusions.

4.2 AUTOMATIC SIMPLIFICATION SETUP

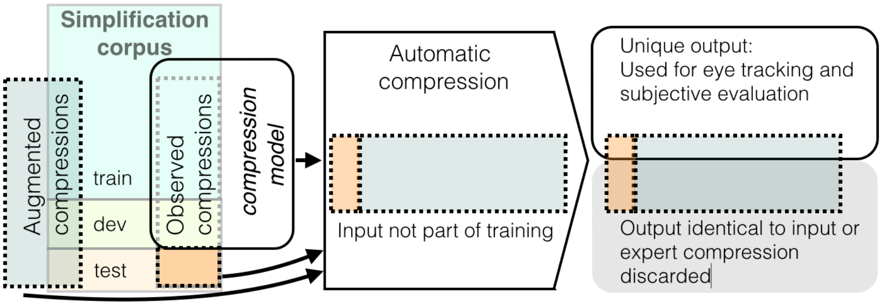

Simplification corpus
Augmented compressions
train
dev
test
Observed compressions
compression model
Automatic compression
Input not part of training
Unique output:
Used for eye tracking and subjective evaluation
Output identical to input or expert compression discarded

Figure 3: We extract observed compressions from the simplification corpus and train an automatic compression model. For the eye tracking and subjective evaluation we run the model on data that was not used for training. We only keep automatic compressions that are different from both the input and the expert compression. Augmented compressions are similar to compressions, but in addition they display one lexical substitution. We augment these by substituting the original synonym back in the expert simplification, thereby making it a compression.

4.2.1 *Training and Evaluation Corpus*

For the sentence compression training and evaluation data we extracted a subset of ordinary and simplified newswire texts from the Danish DSim corpus (Klerke and Søgaard, 2012). In Figure 3 we give a schematic overview of how the data for our experiments was obtained.

For model development and selection we extracted all pairs of original and simplified sentences under the following criteria:

1. No sentence pair differs by more than 150 characters excluding punctuation.
2. The simplified sentence must be a strict subset of the original and contain a minimum of four tokens.
3. The original sentence must have at least one additional token compared to the simplified sentence and this difference must be non-punctuation and of minimum three characters' length.

This results in a corpus of 2,332 sentence pairs, close to 4% of the DSim corpus. Descriptive statistics of this corpus are shown in Table 1.

We followed the train-dev-test split of the DSim corpus forming a training set of 1,973 sentence pairs, a development set of 239 pairs, and a test set of 118 pairs.[1]

For our experiment with eye tracking and subjective evaluation we created a similar dataset, denoted "augmented compressions" in Figure 3, from sentence pairs displaying similar compressions and in addition exactly one lexical substitution. We augmented these pairs by simply changing the synonym back to the original word choice, resulting in a valid compression. We obtained an automatically compressed version of these sentences from the trained model[2]. This results in a corpus of sentence triples consisting of an **original**, an **expert** simplification and a **system** generated version. In some cases the system output was identical to either the original input or to the expert simplification. We therefore selected the evaluation data to

[1] The corpus was PoS-tagged and parsed using the Bohnet parser (Bohnet, 2010) trained on the Danish Dependency Treebank (Kromann, 2003) with Universal PoS-tags (Petrov, Das, and McDonald, 2011).

[2] Note that this dataset did not contribute to training, tuning or choosing the model.

include only sentence triples where all three versions were in fact different from one another resulting in 140 sentence triples, i.e. 420 individual stimuli. On average the system deleted 15 tokens per sentence while the experts average around 12 token deletions per sentence.

	Orig. newswire		Exp. compressions		Difference
	Chars	Tokens	Chars	Tokens	% deleted tokens
Total	288,226	46,088	133,715	21,303	53.8%
Mean	123.6	19.8	57.3	9.1	51.0%
Std	43.2	7.1	24.5	4.0	18.2%
Range	24 – 291	5 – 45	15 – 178	4 – 33	4.4% – 86.2%

Table 1: Statistics on the full specialized corpus, 2.332 sentence pairs in total. Except for the row "Total", statistics are per sentence. "Difference Tokens" report the average, standard deviation and range of the proportional change in number of tokens per sentence.

4.2.2 *Compression Model and Features*

The compression model is a conditional random field (CRF) model trained to make a sequence of categorical decisions, in each determining whether the current word should be left out of the compression output while taking into account the previous decision. We used CRF++ (Lafferty, McCallum, and Pereira, 2001) trained with default parameter settings.

Below, we describe the features we implemented. The features focus on surface form, PoS-tags, dependencies and word frequency information. Our initial choice of features is based on the comparisons in Feng et al., (2010) and Falkenjack and Jönsson, (2014), who both find that parsing features are useful while the gain from adding features beyond shallow features and dependencies is limited. In the CRF++ feature template we specified each feature to include a window of up to +/- 2 tokens. In addition we included all pairwise combinations of features and the bigram feature option which adds the model's previous decision as a feature for the current token.

Shallow FORM, POS, CASE: This group consists of the lowercase word form, universal PoS-tag and the original case of the word.

Length W_LENGTH, S_LENGTH: This group registers the word length (characters) and sentence length (tokens).

Position PLACE, NEG_PLACE, REL_TENTH, THIRDS: This group records the token indices from both the beginning and end of the sentence, as well as each token's relative position measured in tenths and in thirds of the sentence length.

Morphology BIGRAM, TRIGRAM, FOURGRAM: The group records the final two, three and four characters of each token for all tokens of at least four, five and six characters' length, respectively.

Dependencies DEP_HEAD, DEP_LABEL: These two features capture the index of the head of the token and the dependency label of this dependency relation.

Vocabulary OOV, FREQ_3, FREQ_5, FREQ_10PS, FREQ_10EXP: This feature group records a range of frequency-counts[3]. The first feature records out-of-vocabulary words, the remaining features assign the token to one of 3, 5 or 10 bins according to it's frequency.[4] In the 10-bin cases "Pseudo tenths" (PS) assigns the token to one of 10 bins each representing an equal number of word forms[5], while "Exponential" splits the vocabulary into 10 bins containing a decreasing number of word forms as the contained word form frequencies rise exponentially.

4.2.3 *Feature Selection*

We tested five types of feature selection on the development set of the corpus, namely single best feature, single best feature group, add-one, and feature-wise and group-wise feature ablation. On the development set the single best feature was POS alone, the single best feature group was the Shallow group alone, while the add-one-approach returned the combination of the three features FORM, PLACE and FREQ_10PS, and single feature ablation returned all individual features minus FREQ_10EXP, OOV, REL_TENTHS, and group-wise ablation favored all groups minus the Vocabulary and Shallow groups. Of these, the last model, chosen with group-wise feature ablation, obtained the best F1-score on the test set. We use this

[3] We used the Danish reference corpus KorpusDK (Asmussen, 2001) concatenated with the training part of the DSim corpus

[4] 3 bins: in 1K most frequent tokens (mft), 5K mft or outside 5K mft. 5 bins: in 100 mft, 500 mft, 1K mft, 5K mft or outside 5K mft.

[5] Three large bins were assigned word forms occurring 1, 2 and 3 times respectively while the remaining word forms were sorted in seven bins of equal number of word forms

model, which include the feature groups Length, Position, Morphology and Dependencies, to generate system output for the subsequent experiments.

4.3 HUMAN EVALUATION

The experiment described in the following section consisted of an eye tracking part and a subjective evaluation part. The eye tracking part of the experiment was carried out first and was followed by the subjective evaluation part, which was carried out by email invitation to an online survey.

We recruited 24 students aged 20 to 36 with Danish as first language, 6 male and 18 female. All had normal or corrected-to-normal vision. None of the participants had been diagnosed with dyslexia. A total of 20 participants completed the evaluation task. The experiment was a balanced and randomized Latin-square design. This design ensured that each participant saw only one version from each sentence-triple from one half of the dataset while being eye-tracked. Afterwards participants were asked to assign relative ranks between all three versions in each sentence-triple in the half of the dataset which they had not previously seen. In total, each version of each sentence was read by four participants in the eye-tracking experiment and ranked by 9-11 other participants.

In the subjective evaluation task participants had to produce a strict ordering by readability of all three versions of each sentence, with the rank '1' designating the most readable sentence. Presentation order was fully randomized.

4.3.1 *Eye Tracking Design*

The stimuli were presented on a screen with 1080 x 1920 resolution, and eye movements were recorded with a Tobii X120 binocular eye tracker at 60hz. We used the IV-T fixation filter with standard parameter settings (Olsen, 2012). The eye tracker was calibrated to each participant.

Each stimulus was presented on one screen with left, top and right margins of 300 px and 1-6 lines per slide[6]. The font vas Verdana, size 60px and line spacing was 0.8em[7].

Participants were given written instructions and three demo trials before they were left alone to complete the experiment. All participants completed 72 trials in three blocks, with the option to take a short break between blocks. Each trial consisted of a fixation screen visible for 1.5 seconds, followed by stimulus onset. The participants were instructed to try to notice if each sentence was comprehensible and to press a key to proceed to the following trial as soon as they had finished reading.

This setup only encourages but does not require participants to read for comprehension. Through data inspection and informal questions after the experiment, we ascertained that all participants were in fact reading and trying to decide which sentences were comprehensible.

4.3.2 *Eye-movement Measures*

Eye movements in reading can be divided into fixations and saccades. Saccades are rapid eye movements between fixations, and fixations are brief periods of relatively stationary eye positions where information can be obtained from an area covering the central 1-2 degrees of the visual field. Because reading is largely sequential, we can observe regressions, which denote episodes of re-reading, that is, fixations directed at text which is located earlier in the text than the furthest fixated word (Holmqvist et al., 2011).

In our analyses we include the measures of eye movements described below. All measures are calculated per single sentence reading and averaged over all four individual readings of each version of each sentence.

Fixation count (Fix), the average total number of fixations per sentence. This measure is expected to vary with sentence length, with more text requiring more fixations.

Total duration (ms), the average time spent reading the entire sentence. This measure is expected to increase with sentence length and with sentence complexity.

[6] After recording, sentences with seven lines were discarded due to data quality loss at the lower edge of the screen

[7] Following Rauzy and Blache, (2012) who show that the viewing patterns with large text sizes are comparable to smaller text sizes and can be detected with this type of eye tracker.

Fixations per word (Fix/w), the average number of fixations per word. This measure is sensitive to the number of saccades relative to the sentence length and is expected to reflect the reader's confusion as more fixations are needed to collect additional information. It should also be expected to be sensitive to high amounts of long words.

Reading time per word (ms/w), the average time spent per word. This measure increases with slower paced reading, regardless of the number of fixations. Reading time is considered a measure of processing cost and is influenced by both lexical and syntactic complexity.

Proportion regressions (%-regr), the proportion of fixations spent on parts of the text that were already passed once. This measure is typically 10-15% in full paragraphs, and is expected to increase with sentence complexity. (Rayner, Chace, et al., 2006)

We include the sentence length as number of words (n-words) in our analyses for comparison because sentence length can influence the reading strategy (Holmqvist et al., 2011).

Longer sentences will typically have a more complex syntax than short sentences due to the number of entities that need to be integrated into both the syntactic and mental representation of the sentence. However, unfamiliar or even erroneous words and syntax can add processing difficulties as well, leaving the reader to guess parts of the intended message. We consider all these cases under the term *complexity* as they are all likely to appear in automatically processed text. This is a natural consequence of the fact that statistical language processing tools are typically not able to distinguish between extremely rare, but admissible text use and text that would be judged as invalid by a reader.

4.4 RESULTS

We first analyze the correlation of the subjective evaluations followed by analyses that compare eye movement measures, subjective rankings and sentence version.

4.4.1 *Ranking*

First we test whether the subjective rankings are similar between subjects. We estimate agreement with Kendall's τ_B association statistic, which is a pairwise correlation coefficient appropriate for comparing rank orderings. The range of τ_B is $[-1, 1]$ where -1 indicates perfect disagreement, i.e. one ranking is the precise opposite order of the other, 1 indicates perfect agreement and 0 indicates no association, that is, the order of two elements in one ranking is equally likely to be the same and the opposite in the other ranking. The odds-ratio of a pair of elements being ranked concordantly is $(1 + \tau_B)/(1 - \tau_B)$. The metric τ_B compares pairs of rankings, and we therefore calculate the average over all pairs of participants' agreement on each ranking task. We use the one-tailed one-sample student's t-test to test whether the average agreement between all 91 unique pairs of annotators is significantly different from 0. If the rankings are awarded based on a shared understanding and perception of readability, we expect the average agreement to be positive.

We find that the average τ_B is $0.311(p < 0.0001)$. This corresponds to a concordance odds-ratio of 1.90 which means that it is almost twice as likely that two annotators will agree than disagree on how to rank two versions of a sentence. Although this result is strongly significant, we note that it is a surprisingly low agreement given that the chance agreement is high for two people ranking three items.

The relatively low systematic agreement could arise either from annotators ranking only a few traits systematically (e.g. syntax errors rank low when present and otherwise ranking is random) or it could result from annotators following fully systematic but only slightly overlapping strategies for ranking (e.g. one ranks by number of long words while another ranks by sentence length which would tend to overlap).

4.4.2 *Eye Tracking*

Our second analysis tests how well the subjective ranking of sentences correspond to eye movements. We expect that more complex text will slow down readers, and we want to know whether the perceived readability reflects the variation we observe in eye-movement measures. Again using the τ_B association, we now assign ranks within each sentence-triple based on each eye-tracking measure and compare these pseudo-rankings to the *typical rank* assigned by the annotators.[8] We find that neither sentence length or any of the eye tracking measures are significantly associated with the *typical rank*. This means that

[8] This approach introduces ties which are handled by the τ_B statistic but influences the result notably since each ranking task only includes 3 items.

	Difference in medians									
	Sys – Exp		Sys – Ori		Ori – Exp		Exp – Brk		Brk – Ori	
avg. rank	**0.25**	*	**-0.47**	***	**-0.73**	***	**-1.51**	***	**0.78**	***
ms	-125	–	**-3173**	***	**2830**	***	-33	–	**-2797**	***
Fix	-0.8	–	**-14.0**	***	**12.3**	***	-1.3	–	**-11**	***
ms/w	50.1	–	-50.1	–	**267**	**	**-217**	**	-50.1	–
fix/w	0.1	–	**0.4**	***	**-0.17**	**	-0.19	–	**0.36**	***
%-regr	**4**	**	1	–	**9**	**	**11**	***	2	–
n-words	**-1 w**	*	**-2 w**	*	5 w	–	0 w	–	-5 w	–

Table 2: Influence of sentence variant and brokenness on perceived readability and eye movements. When comparing Expert (Exp), Original (Ori) and System (Sys) 109 sentences are included while for Broken (Brk) only 27 sentences are compared. Stars denote significance levels: *: $p < .05$, **: $p < .01$, ***: $p < .001$

we do not observe any correlation between sentences' perceived readability and the sentence length, the time it takes to read it or the speed or number of fixations or proportion of regressions recorded.

One potential reason why we do not observe the expected association between rank and eye movements can be that several of our eye tracking measures are expected to vary differently with sentence length and complexity, whereas readers' readability rankings are not necessarily varying consistently with any of these dimensions as participants are forced to conflate their experience into a one-dimensional evaluation.

In order to investigate whether the eye movements do in fact distinguish between length and complexity in sentences, we compare how readers read and rank long original sentences, short expert simplifications and short, syntactically broken system output.

The system output was post hoc categorized by syntactic acceptability by the main author and a colleague, resulting in a sample of 27 sentence triples with syntactically unacceptable system and a sample of 109 fully syntactically acceptable sentence triples. This allows us to compare the following four groups, Original, Expert, Unbroken System and Broken System.

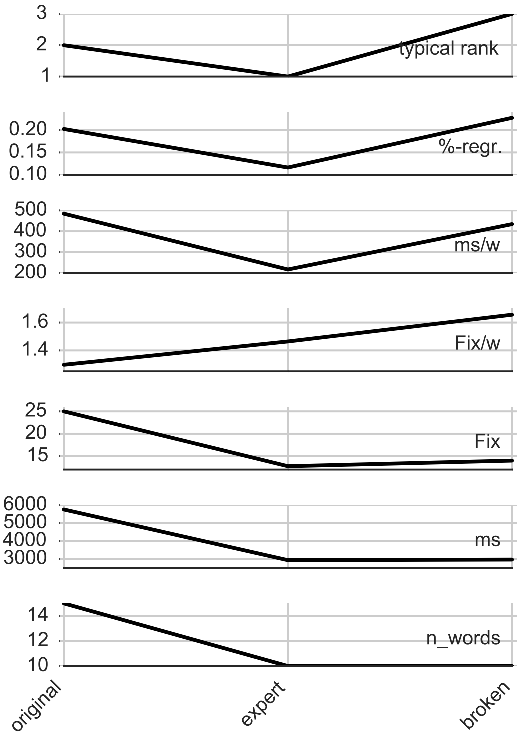


Figure 4: Interaction of sentence type and brokenness on perceived readability and eye movements. (N=27)

We compare all eye-movement measures and ranking for each pair of groups[9] and test whether the measures differ significantly between groups using the Wilcoxon signed-rank test. We report the comparisons as the difference between the medians in Table 2. This is similar to an unnormalized Cohen's d effect size, but using the median as estimate of the central tendency rather than the mean. We observe that all group-wise comparisons receive significantly different average ranks, ranging from the Unbroken System scoring a quarter of a rank-position better than the Expert compressions to the Broken System output fairing 1.51 rank positions worse than the Expert group.

Note that Broken System is also ranked significantly below the Original newswire sentences, signaling that bad syntax has a stronger impact on perceived readability than length. Even though the sample of Broken System sentences is small, overall reading time and number of fixations distinguish the long Original sentences from both the short Expert simplifications

[9] We use the larger sample whenever the group Broken System is not part of the comparison.

and Broken System outputs, that are comparably short. We also observe that the number of fixations per word is consistently lower for the long Original sentences compared to the other, shorter groups. Importantly, we observe that two measures significantly distinguish Expert simplifications from syntactically Broken System output, namely reading time per word, which is slower for Broken System syntax and proportion of regressions which is much higher in Broken System sentences. In addition and as the only eye-tracking measure, proportion of regressions also distinguishes between Unbroken System output and Expert simplifications, indicating a 4 percentage point increase in proportion of regressions when reading Unbroken System output.

In Figure 4 we show how the medians of all the measures vary in the small subset that contain Broken System output, Expert compressions and Original newswire. The figure illustrates how the different aspects of reading behavior reflect length and syntax differently, with regressions most closely following the subjective ranking (top).

4.5 DISCUSSION

In the following section we discuss weaknesses and implications of our results.

4.5.1 *Learning and Scoring the Compression Model*

It is important to note that the compression model inherently relies on the expert compressed data, which means it penalizes any deviation from the single gold compression. This behavior is sub-optimal given that various good simplifications usually can be produced by deletion and that alternative good compressions are not necessarily overlapping with the gold compression. One example would be to pick either part of a split sentence which can be equally good but will have zero overlap and count as an error. Our results suggest that the framework is still viable to learn a useful model, which would need a post-processing syntax check to overcome the syntax errors arising in the deletion process.

We note that the model produces more aggressive deletions than the experts, sometimes producing sentences that sound more like headlines than the body of a text. It is surprising that this is the case, as it is typically considered easier to improve the readability slightly, but we speculate that the behavior could reflect that the parts of the training data with headline-like characteristics may provide a strong, learnable pattern. However, from an application perspective, it would be simple to exploit this in a stacked model setup, where models trained to exhibit different characteristics present a range of alternative simplifications to a higher-level model.

From inspections of the output we observe that the first clause tends to be kept. This may be domain-dependent or it may reflect that PoS-tags and parsing features are more reliable in the beginning of the sentence. This could be tested in the future by applying the model to text from a domain with different information structure.

4.5.2 *Implications for System Development*

We found that the very simple compression model presented in this paper was performing extensive simplifications, which is important in light of the fact that humans consider it harder to produce more aggressive simplifications. We trained our model on a relatively small, specialized compression corpus. The Simple English Wikipedia simplification corpus (SEW) (Coster and Kauchak, 2011), which has been used in a range of statistical text simplification systems (Coster and Kauchak, 2011; Woodsend and Lapata, 2011b; Zhu, Bernhard, and Gurevych, 2010), is far bigger, but also noisier. We found fewer than 50 sentence pairs fitting our compression criteria when exploring the possibility of generating a similar training set for English from the SEW. However, in future work, other, smaller simplification corpora could be adapted to the task, providing insight into the robustness of using compression for simplification.

4.5.3 *Implications for Evaluation Methodology*

In many natural language generation and manipulation setups, it is important that the system is able to recognize acceptable output, and it is typical of this type of setup that neither system-intrinsic scoring functions or as standard automatic evaluation procedures are reliably meeting this requirement. In such cases it is common to obtain expensive specialized human evaluations of the output. Our results are encouraging as they suggest that behavioral metrics like regressions and reading time that can be obtained from naïve subjects simply reading system output may provide an affordable alternative.

4.5.4 *Brokenness in NLP output*

The experiments we have presented are targeting a problem specific to the field of computer manipulation of texts. In contrast to human-written text, language generation systems typically cannot fully guarantee that the text will be fluent and coherent in both syntax and semantics. Earlier research in readability has focused on how less-skilled readers, like children, dyslectic readers and second-language readers, interact with natural text, often in paragraphs or longer passages. It is important to determine to what extent the existing knowledge in these fields can be transferred to computational linguistics.

4.6 CONCLUSION

We have compared subjective evaluations and eye-movement data and shown that human simplifications and automatic sentence compressions of newswire produce variations in eye movements.

We found that the main source of difficulty in processing machine-compressed text is ungrammaticality. Our results further show that both the human simplifications and the grammatical automatic sentence compressions in our data are easier to process than the original newswire text.

Regressions and reading speed were found to be good candidates for robust, transferrable measures that, with increasing access to eye-tracking technology, are strong candidates for being directly incorporated into language technologies.

We have shown that these measures can capture significant differences in skilled readers' reading of single sentences across subjects and with ecologically valid stimuli. In future research we wish to explore the possibility of predicting relevant reading behavior for providing feedback to NLP systems like automatic text simplification and sentence compression.

5 STUDY 2: READING METRICS FOR ESTIMATING TASK EFFICIENCY WITH MT OUTPUT

ABSTRACT

We show that metrics derived from recording gaze while reading, are better proxies for machine translation quality than automated metrics. With reliable eye-tracking technologies becoming available for home computers and mobile devices, such metrics are readily available even in the absence of representative held-out human translations. In other words, reading-derived MT metrics offer a way of getting cheap, online feedback for MT system adaptation.

5.1 INTRODUCTION

What's a good translation? One way of thinking about this question is in terms of what the translations can be used for. In the words of Doyon, Taylor, and White, (1999), "a poor translation may suffice to determine the general topic of a text, but may not permit accurate identification of participants or the specific event." Text-based tasks can thus be ordered according to their tolerance of translation errors, as determined by actual task outcomes, and task outcome can in turn be used to measure the quality of translation (Doyon, Taylor, and White, 1999).

Machine translation (MT) evaluation metrics must be both adequate and practical. Human task performance, say participants' ability to extract information from translations, is perhaps the most adequate measure of translation quality. Participants' direct judgements of translation quality may be heavily biased by perceived grammaticality and subjective factors, whereas task performance directly measures the usefulness of a translation. Of course different tasks rely on different aspects of texts, but some texts are written with a single purpose in mind.

In this paper, we focus on logic puzzles. The obvious task in logic puzzles is whether readers can solve the puzzles when given a more or less erroneous translation of the puzzle. We assume task performance on logic puzzles is an adequate measure of translation quality *for logic puzzles*.

Task-performance is not always a practical measure, however. Human judgments, whether from direct judgments or from answering text-related questions, takes time and requires recruiting and paying individuals. In this paper, we propose various metrics derived from natural reading behavior as proxies of task-performance. Reading has several advantages over other human judgments: It is fast, is relatively unbiased, and, most importantly, something that most of us do effortlessly all the time. Hence, with the development of robust eye tracking methods for home computers and mobile devices, this can potentially provide us with large-scale, on-line evaluation of MT output.

This paper shows that reading-derived metrics are better proxies of task-performance than the standard automatic metric BLEU. Note also that on-line evaluation with BLEU is biased by what held-out human translations you have available, whereas reading-derived metrics can be used for tuning systems to new domains and new text types.

In our experiments, we include simplifications of logic puzzles and machine translations thereof. Our experiments show, as a side result, that a promising approach to optimizing machine translation for task performance is using text simplification for pre-processing the source texts. The intuition is that translation noise is more likely to make processing harder in more complex texts.

5.1.1 *Contributions*

- We present an experimental eye-tracking study of 20 participants reading simplifications and human/machine translations of 80 logic puzzles.[1]

- This is, to the best of our knowledge, the first study to correlate reading-derived metrics, human judgments and BLEU with task performance for evaluating MT. We show that human judgments do not correlate with task perfor-

[1] The data will be made available from https://github.com/coastalcph

mance. We also show that reading-derived metrics correlate significantly with task performance ($-.36 < r < -.35$), while BLEU does not.

- Finally, our results suggest that practical MT can benefit much from incorporating sentence compression or text simplification as a pre-processing step.

5.2 SUMMARY OF THE EXPERIMENT

In our experiments, we presented participants with 80 different logic puzzles and asked them to solve and judge the puzzles while their eye movements were recorded. Each puzzle was edited into five different versions: the original version in English (L2), a human simplification thereof (s(·)), a human translation into Danish (L1) and a machine translation of the original (M(·)), as well as a machine translation of the simplification (M(s(·))). Consequently, we used 400 different stimuli in our experiments. The participants were 20 native speakers of Danish with proficiency in English.

We record fixation count, reading speed and regression proportion (amount of fixations landing on previously read text) from the gaze data. Increased attention in the form of reading time and re-reading of previously read text are well-established indicators of increased cognitive processing load, and they correlate with typical readability indicators like word frequency, length and some complex syntactic structures (Holmqvist et al., 2011; Rayner, 1998; Rayner and Pollatsek, 2013). We study how these measures correlate with MT quality, as reflected by human judgments and participants' task performance.

We thereby assume that the chance of quickly solving a task decreases when more resources are required for understanding the task. By keeping the task constant, we can assess the relative impact of the linguistic quality of the task formulation. We hypothesise that our five text versions (L1, L2, M(·), s(·), M(s(·))), can be ranked in terms of processing ease, with greater processing ease allowing for more efficient task solving.

The experiments are designed to test the following hypothesized partial ordering of the text versions (summarized in Table 3): text simplification (s(·)) eases reading processing relative to second language reading processing (L2) while professional human translations into L1 eases processing more (**H1**). In addition, machine translated text (M(·)) is expected to ease the

Math
A DVD player with a list price of $100 is marked down 30%. If John gets an employee discount of 20% off the sale price, how much does John pay for the DVD player?
1: 86.00
2: 77.60
3: 56.00
4: 50.00

Conclude
Erin is twelve years old. For three years, she has been asking her parents for a dog. Her parents have told her that they believe a dog would not be happy in an apartment, but they have given her permission to have a bird. Erin has not yet decided what kind of bird she would like to have.
Choose the statement that logically follows
1: Erin's parents like birds better than they like dogs.
2: Erin does not like birds.
3: Erin and her parents live in an apartment.
4: Erin and her parents would like to move.

Evaluate
Blueberries cost more than strawberries.
Blueberries cost less than raspberries.
Raspberries cost more than both strawberries and blueberries.
If the first two statements are true, the third statement is:
1: TRUE
2: FALSE
3: Impossible to determine

Infer
Of all the chores Michael had around the house, it was his least favorite. Folding the laundry was fine, doing the dishes, that was all right. But he could not stand hauling the large bags over to the giant silver canisters. He hated the smell and the possibility of rats. It was disgusting.
This paragraph best supports the statement that:
1: Michael hates folding the laundry.
2: Michael hates doing the dishes.
3: Michael hates taking out the garbage.
4: Michael hates cleaning his room.

Figure 5: Logic puzzles of four categories. The stimuli contain 20 of each puzzle category.

processing load, but less so than machine translation of simplified text (M(s(·))), although both of these machine translated versions are still expected to be more demanding than the professionally translated original text (L1). Table 3 provides an overview of the hypotheses and the expected relative difficulty of processing each text version.

H1:	L1	$\prec$ s(·) $\prec$	L2
H2:	L1	$\prec$ M(s(·)) $\prec$ M(·) $\prec$	L2

Table 3: Expected relative difficulty of processing. L1 and L2 are human edited texts in the participants' native and non-native language, respectively, s(·) are manually simplified texts, M(·) are machine translated texts and M(s(·)) are machine translations of manually simplified texts.

5.2.1 *Summary of the findings*

Our experimental findings are summarized as follows: The data supports the base assumption that L1 is easier than L2. We only find *partial* support for H1; While s(·) tends to be easier to comprehend than L2, also leading to improved task performance, s(·) is ranked as easier to process than L1 as often as the opposite, hypothesised ranking. This indicates that our proficient L2 readers may be benefitting as much from simplification as from translation in reasoning tasks.

We also only find *partial* support for H2: The relative ordering of the human translations, L1, and the two machine translated versions, M(s(·)) and M(·), is supported and we find that the simplification improves MT a lot with respect to reading processing. However, participants tended to perform better with the original L2 logic puzzles compared to the machine translated versions.

In other words, MT hurts while both manual simplification and translation help even proficient L2 readers. In sum, simplification seems necessary if L2-to-L1 MT is to ease comprehension, and not make understanding harder for readers with a certain L2 command level.

Importantly, we proceed to study the correlation of our eye-tracking measures, human judgments and BLEU (Papineni et al., 2002) with task performance. There has been considerable work on how various automatic metrics correlate with human judgments, as well as on inter-annotator consistency among humans judging the quality of translations (Callison-Burch et al., 2008). Various metrics have been proposed over the years, but BLEU (Papineni et al., 2002) remains the *de facto* state-of-the-art evaluation metric. Our findings, related to evaluation, are, as already mentioned, that (a) human judgments surprisingly do not correlate with task performance, and that (b) the reading-derived metrics Time and Fixations correlate strongly with task performance, while BLEU does not. This, in our view, questions the validity of human judgments and the BLEU metric and shows that reading-derived MT metrics may provide a better feedback in system development and adaptation.

5.3 DETAILED DESCRIPTION OF THE EXPERIMENT

5.3.1 *Stimuli*

In this section, we describe the texts we have used for stimuli, as well as the experimental design and our participants.

We selected a set of 80 logic puzzles written in English, all with multiple-choice answers.[2] The most important selection criterium was that participants have to reason about the text and cannot simply recognize a few entities directly to guess the answer. The puzzles were of four different categories, all designed to train logic reasoning and math skills in an educational context. We chose 20 of each of the four puzzle categories to ensure a wide variety of reasoning requirements. Figure 5 shows an example question from each category.

The English (L2) questions and multiple choice answer options were translated into Danish (L1) by professional translators. The *question text* was manually simplified by the lead author (s(·)). Both of the English versions were machine-translated into Danish (M(·), M(s(·))).[3] This results in the five versions of the question texts, which were used for analysis. The multiple-choice answer options were not simplified or machine translated. Thus the participants saw either the original English answers or the human-translated Danish answers, matching the language of the question text. The average number of words and long words in each of the five versions are reported in Table 4.

[2] From LearningExpress, (2005).

[3] Google Translate, accessed on 29/09/2014 23.33 CET.

	# Long words		# Words	
Variant	mean	std	mean	std
L2	9.56	6.67	38.33	19.29
s(·)	8.78	5.90	35.78	17.43
L1	10.22	6.97	38.87	21.28
M(s(·))	9.70	6.75	35.19	19.07
M(·)	10.35	6.74	36.53	19.04

Table 4: Mean and standard deviation of number of words and number of words with more than seven letters per question for all five versions.

Simplification is not a well-defined task and is often biased intentionally to fit a target audience or task. To allow for comparison with parallel simplification corpora, we classified the applied simplification operations into the following set of seven abstract simplification operations and present their relative proportion in Table 5: Sentence splitting. information deletion and information reordering, discourse marker insertion (e.g., *and*, *but*), anaphora substitution (e.g., *Zoe's garden* vs. *the garden*), other lexical substitutions (e.g., *dogwoods* vs. *dogwood trees*) and paraphrasing (e.g., *all dogwoods* vs. *all kinds of dogwood trees*). On average 2.0 simplification operations was performed per question, while a total of 28.7% of the questions were left unchanged during simplification. All simplified questions still required the reader to understand and reason about the text. The simplifications were performed with the multiple answer texts in mind; leaving any information referenced in the answers intact in the question, even when deleting it would have simplified the question text.

5.3.2 *Experimental design*

The experiment followed a Latin-square design where each participant completed 40 trials, judging and solving 40 different puzzles, eight of each of the five versions.

A trial consisted of three tasks (see Figure 6): a comprehension task, a solving task and a comparison task. Each trial was preceded by a 1.5 second display of a fixation cross. The remainder of the trial was self-paced. During the entire trial - i.e.,

Simplification	%
Lexical substitution	27.4
Paraphrase	24.2
Deletion	23.1
Information reordering	11.3
Anaphora substitution	7.5
Discourse marker insertion	4.3
Sentence splitting	2.2

Table 5: Simplification operations (SOps). The total number of applied SOps was 186, the average number of SOps applied per question was 2.0 (std 1.3).

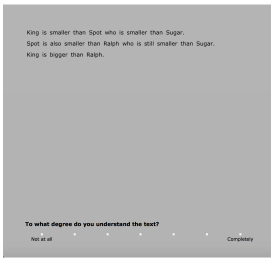
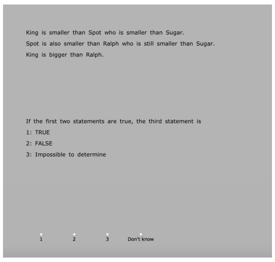
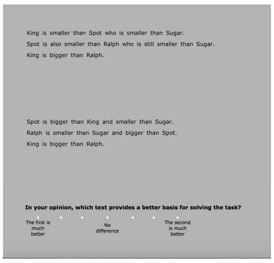
One trial		
1: Read to comprehend	2: Solve puzzle	3: Compare versions
 King is smaller than Spot who is smaller than Sugar. Spot is also smaller than Ralph who is still smaller than Sugar. King is bigger than Ralph. **To what degree do you understand the text?** Not at all · · · · · Completely	 King is smaller than Spot who is smaller than Sugar. Spot is also smaller than Ralph who is still smaller than Sugar. King is bigger than Ralph. If the first two statements are true, the third statement is 1: TRUE 2: FALSE 3: Impossible to determine 1 2 3 Don't know	 King is smaller than Spot who is smaller than Sugar. Spot is also smaller than Ralph who is still smaller than Sugar. King is bigger than Ralph. Spot is bigger than King and smaller than Sugar. Ralph is smaller than Sugar and bigger than Spot. King is bigger than Ralph. **In your opinion, which text provides a better basis for solving the task?** The first is much better · No difference · The second is much better

Figure 6: Illustration of one trial. Each trial consists of three individual tasks. The top third of the screen displays the target text and is fixed for the duration of the entire trial.

for the duration of the three tasks - the question text was presented on the top part of the screen. In the comprehension task, the participant was asked to rate the comprehensibility of the question text on a 7-point Likert scale that was presented at the bottom part of the screen. This score is called Comprehension, henceforth. This is our rough equivalent of human judgments of translation quality. For the solving task, the multiple-choice answer options was presented in the middle part of the screen below the question text and the participant indicated an answer or "don't know" option in the bottom part of the screen. The measure Efficiency, which was also introduced in Doherty and O'Brien, (2014), is the number of correct answers given for a version, C_v over the time spent reading and solving the puzzles of that version, S_v: $E = \frac{C_v}{S_v}$. This score is our benchmarking metric below.

In the last task, Comparison, a different version of the same question text was presented below the first question text, always in the same language. Participants were asked to assess which version provided a better basis for solving the task using a 7-point Likert scale with a neutral midpoint. The three leftmost options favored the text at the top of the screen, while the three rightmost choices favored the text at the lower half of the screen.

Each participant completed three demo trials with the experimenter present. Participants were kept naïve with regards to the machine translation aspect of the study. They were instructed to solve the puzzles as quickly and accurately as possible and to judge Comprehension and Comparison quickly. Each session included a 5-10 minute break with refreshments halfway through. At the end of the experiment a brief questionnaire was completed verbally. All participants completed the entire session in 70–90 minutes.[4]

5.3.2.1 *Apparatus*

The stimuli were presented in black letters in the typeface Verdana with a letter size of 20 pixels (ca. .4° visual angle) on a light gray background with 100 pixels margins. The eye tracker was a Tobii X120, recording both eyes with 120hz sampling rate. We used Tobii Studio standard settings for fixation detection. The stimuli was presented on a 19" display with a resolution of 1920 x 1080 pixels and a viewing distance of ca 65 cm. Here we focus on the initial reading task and report total reading time per word (Time), number of fixations per word (Fixations) and proportion of regressions (Regressions). The calculations of the eyetracking measures are detailed in Section 5.4.3.

5.3.2.2 *Participants*

We recruited participants until we obtained a total of 20 recordings of acceptable quality. In this process we discarded two participants due to sampling loss. Another two participants were dismissed due to unsuccessful calibration. All participants completed a pre-test questionnaire identifying themselves as native

[4] Participants received a voucher for 10 cups of tea/coffee upon completion.

Danish speakers with at least a limited working proficiency of English. None of the participants had been diagnosed with dyslexia, and all had normal or corrected to normal vision. The 20 participants (4 males) were between 20 and 34 years old (mean 25.8) and minimum education level was ongoing bachelor's studies.

5.4 RESULTS

The mean values for all metrics and the derived rankings of the five versions are presented in Table 6. Significance is computed using Student's paired *t*-test, comparing each version to the version with the largest measured value. Table 7 presents correlations with task performance (Efficiency) for each measure. We describe the correlations, and their proposed interpretation, in Section 5.4.4.

Version			μ			Rankings
	L1	M(s(·))	M(·)	s(·)	L2	
Compr.	5.58	**4.51	**4.50	5.61	5.46	s(·) ≺ L1 ≺ L2 ≺ M(s(·)) ≺ M(·)
Compa.	1.62	**−.54	**−1.07	.43	**−.43	L1 ≺ M(s(·)) ≺ M(·) | s(·) ≺ L2
Eff.	.94	.90	**0.80	1.0	.87	s(·) ≺ L1 ≺ M(s(·)) ≺ L2 ≺ M(·)
Time	.54	.62	.65	.55	.54	L1 ≺ L2 ≺ s(·) ≺ M(s(·)) ≺ M(·)
Fix.	1.51	1.66	1.83	1.55	1.60	L1 ≺ s(·) ≺ L2 ≺ M(s(·)) ≺ M(·)
Regr.	17.77	18.46	19.15	15.55	16.55	s(·) ≺ L2 ≺ L1 ≺ M(s(·)) ≺ M(·)

Table 6: Mean values for the five text versions. Comprehension and Comparison are Likert scale scores respectively ranging from 0 to 7 and from −3 to 3, Efficiency is correct answers relative to reading speed, Time is seconds per word, Fixations is number of fixations per word and Regressions is proportion of re-fixations (**: Student's paired t-test relative to largest mean value $p < 0.001$)

5.4.1 *Subjective measures*

We elicited subjective evaluations of text comprehension and pairwise comparisons of versions' usefulness for solving the puzzles. Note that participants evaluate MT output significantly lower than human-edited versions.

We treated the pairwise Comparison scores as votes, counting the preference of one version as equally many positive and negative votes on the preferred version and the dis-preferred version, respectively. With this setup, we maintain zero as a neutral evaluation. Comparison was only made within the same language, so the scores should not be interpreted across languages. Note, however, how Comparison results show a clear ranking of versions within each language.

5.4.2 *Task performance measures*

The task performance is reported as the Efficiency, i.e., correct answers per minute spent reading and solving puzzles. We observe that the absolute performance ranges from 48% to 52% correct answers. This is well above chance level (27%), and does not differ significantly between the five versions, reflecting that the between-puzzles difference in difficulty level, as expected, is much larger than the between-versions difference.

Efficiency, however, reveals a clearer ranking. Participants were less efficient solving logic puzzles when presented with machine translations of the original puzzles. The machine translations of the simplified puzzles actually seemingly *eased* task performance, compared to using the English originals, but differences are not statistically significant. The simplified English puzzles led to the best task performance.

5.4.3 *Eye-tracking measures*

The reading times in seconds per word (Time) are averages over reading times while fixating at the question text located on the upper part of the screen during the first sub-task of each trial (judging comprehension). This measure is comparable to normalized total reading time in related work. Participants spent most time on the machine translations, whether of the original texts or the simplified versions.

The measure Fixations similarly was recorded on the question part of the text during the initial comprehension task, normalized by text length, and averaged over participants and versions. Again we observe a tendency towards more fixations on machine translated text, and fewest on the human translations into Danish.

	Data used	r	$p \leq .001$
COMPREHENSION	all	.25	-
	M(s(·))	.36	-
	M(·)	-.27	-
COMPARISON	all	.13	-
	M(s(·))	.06	-
	M(·)	.26	-
TIME	all	-.35	✓
	M(s(·))	-.19	-
	M(·)	-.54	-
FIXATIONS	all	-.36	✓
	M(s(·))	-.26	-
	M(·)	-.57	-
REGRESSIONS	all	-.17	-
	M(s(·))	.01	-
	M(·)	-.33	-
BLEU	M(s(·))	-.13	-
	M(·)	-.17	-

Table 7: Correlations with EFFICIENCY (Pearson's r). BLEU only available on translated text. Correlation reported on these subsets for comparability.

Finally, we calculated REGRESSIONS during initial reading as the proportion of fixations from *the furthest word read* to a preceding point in the text. Regressions may indicate confusion and on average account for 10-15% of fixations during reading (Rayner, 1998). Again we see more regressions with machine translated text, and fewest with simplified English puzzles.

5.4.4 *Correlations between measures*

We observe the following correlations between our measures. All correlations with EFFICIENCY are shown in Table 7. First of all, we found no correlations between subjective measures and eye-tracking measures nor between subjective measures and task performance. The two subjective measures, however, show a strong correlation (Spearman's $r = .50$ $p < .001$). EFFIciency shows significant negative correlation with both of the eye-tracking measures Time (Pearson's $r = -.35$ $p < .001$ and Fixations (Pearson's $r = -.36$ $p < .001$), but not Regressions . Within the group of eye-tracking measures Time and Fixation exhibit a high correlation ($r = 0.94$ $p < .001$). Regressions is significantly negatively correlated with both of these (Pearson's $r = -.38$ $p < .001$ and Pearson's $r = -.43$ $p < .001$, respectively).

We obtain BLEU scores (Papineni et al., 2002) by using the human-translated Danish text (L1) as reference for both of the MT outputs, M(·) and M(s(·)). The overall BLEU score for M(·) version is .691, which is generally considered very good, and .670 for M(s(·)). The difference is not surprising, since M(s(·)) inputs a different (simpler) text to the MT system. On the other hand, given that our participants tended to be more efficiently comprehending and solving the logic puzzles using M(s(·)), this already indicates that BLEU is not a good metric for talking about the usefulness of translations of instructional texts such as logic puzzles.

Our most important finding is that BLEU does not correlate with Efficiency, while two of our reading-derived metrics do. In other words, the normalised reading time and fixation counts are better measures of task performance, and thereby of translation quality, than the state-of-the-art metric, BLEU in this context. This is an important finding since reading-derived metrics are potentially also more useful as they do not depend on the availability of professional translators.

5.5 discussion

Several of our hypotheses were in part falsified. L2 is solved more efficiently by our participants than M(·), not the other way around. Also, M(s(·)) is judged as harder to comprehend than s(·) and consistently ranked so by all metrics. These observations suggest that MT is not assisting our participants despite the fact that L2 ranks lower than L1 in four out of five comparisons. Our participants are university students and did not report to have skipped any questions due to the English text suggesting generally very good L2 skills.

If we assume that Efficiency – as a measure of task performance – is a good measure of translation quality (or usefulness), we see that the best indicator of translation quality that only takes the initial reading into account are Fixations and Time. This indicates that Fixations and Time may be better MT benchmarking metrics than BLEU.

5.6 related work

Eye tracking has been used for MT evaluation in both post-editing and instruction tasks (Castilho et al., 2014; Doherty and O'Brien, 2014).

Doherty, O'Brien, and Carl, (2010) also used eye-tracking measures for evaluating MT output and found fixation count and gaze time to correlate negatively with binary quality judgments for translation segments, whereas average fixation duration and pupil dilation were not found to vary reliably with the experimental conditions. A notable shortcoming of that study is that the translated segments in each category were different, making it impossible to rule out that the observed variation in both text quality and cognitive load was caused in part by an underlying variation in content complexity.

This shortcoming was alleviated in a recent re-analysis of previous experiments (Doherty, Kenny, and Way, 2012; Doherty and O'Brien, 2014) which compares the usability of raw machine translation output in different languages and the original, well-formed English input. In order to test usability, a plausible task has to be set up. In this study the authors used an instructional text on how to complete a sequence of steps using a software service, previously unknown to the participants. MT output was obtained for four different languages and three to four native speakers worked with each output. Participants' subjective assessment of the usability of the instructions, their performance in terms of efficiency and the cognitive load they encountered as measured from eye movements were compared across languages. The results of this study supports the previous finding that fixation count and total task time depends on whether the reader worked with the original or MT output, at least when the quality of the MT output is low. In addition, goal completion and efficiency (total task time relative to goal completion) as well as the number of shifts (between instructions and task performance area) were shown to co-vary with the text quality.

Castilho et al., (2014) employed a similar design to compare the usability of lightly post-edited MT output to raw MT output and found that also light post-editing was accompanied by fewer fixations and lower total fixation time (proportional to total task time) as well as fewer attentional shifts and increased efficiency.

In contrast, Stymne, Danielsson, et al., (2012) found no significant differences in total fixation counts and overall gaze time (proportional to total task time), when directly comparing output of different MT systems with expected quality differences. However, they showed that both of these two eye-tracking measures were increased for the parts of the text containing errors in comparison with error-free passages. In addition, they found gaze time to vary with specific error types in machine translated text.

From an application perspective, Specia, (2011) suggested the time-to-edit measure as an objective and accessible measure of translation quality. In their study it outperformed subjective quality assessments as annotations for a model for translation candidate ranking. Their tool was aimed at optimizing the productivity in post-editing tasks.

Eye tracking can be seen as a similarly objective metric for fluency estimation (Stymne, Danielsson, et al., 2012). The fact that eye tracking does not rely on translators makes annotation even more accessible.

Both Doherty and O'Brien, (2014) and Castilho et al., (2014) found subjective comprehensibility, satisfaction and likelihood to recommend a product to be especially sensitive to whether the instructional text for the product was raw MT output. This suggests that the lower reliability of subjective evaluations as annotations could be due to a bias against MT-specific errors. Only Stymne, Danielsson, et al., (2012) report the correlations between eye movement measures and subjective assessments and found only moderate correlations.

This work is to the best of our knowledge the first to study the correlation of reading-derived MT metrics and task performance. Since we believe task performance to be a more adequate measure of translation quality – especially when the texts are designed with a specific task in mind – we therefore believe this to be a more adequate study of the usefulness of reading-derived MT metrics than previous work.

5.7 CONCLUSION

We presented an eye-tracking study of participants reading original, simplified, and human/machine translated logic puzzles. Our analysis shows that the reading-derived metrics Time and Fixations obtained from eye-tracking recordings can be used to assess translation quality. In fact, such metrics seem to be much better proxies of task performance, i.e., the practical usefulness of translations, than the state-of-the-art quality metric, BLEU.

6

STUDY 3: IMPROVING SENTENCE COMPRESSION BY LEARNING TO PREDICT GAZE

ABSTRACT

We show how eye-tracking corpora can be used to improve sentence compression models, presenting a novel multi-task learning algorithm based on multi-layer LSTMs. We obtain performance competitive with or better than state-of-the-art approaches.

6.1 INTRODUCTION

Sentence compression is a basic operation in text simplification which has the potential to improve statistical machine translation and automatic summarization (Berg-Kirkpatrick, Gillick, and Klein, 2011; Klerke, Castilho, et al., 2015), as well as helping poor readers in need of assistive technologies (Canning et al., 2000). This work suggests using eye-tracking recordings for improving sentence compression for text simplification systems and is motivated by two observations: (i) *Sentence compression is the task of automatically making sentences easier to process by shortening them.* (ii) *Eye-tracking measures* such as first-pass reading time and time spent on regressions, i.e., during second and later passes over the text, *are known to correlate with perceived text difficulty* (Rayner, Pollatsek, et al., 2012).

These two observations recently lead Klerke, Castilho, et al., (2015) to suggest using eye-tracking measures as metrics in text simplification. We go beyond this by suggesting that eye-tracking recordings can be used to induce better models for sentence compression for text simplification. Specifically, we show how to use existing eye-tracking recordings to improve the induction of Long Short-Term Memory models (LSTMs) for sentence compression.

Our proposed model *does not require* that the gaze data and the compression data come from the same source. Indeed, in this work we use gaze data from readers of the Dundee Corpus to improve sentence compression results on several datasets.

While not explored here, an intriguing potential of this work is in deriving sentence simplification models that are personalized for individual users, based on their reading behavior.

Several approaches to sentence compression have been proposed, from noisy channel models (Knight and Marcu, 2002) over conditional random fields (Elming et al., 2013) to tree-to-tree machine translation models (Woodsend and Lapata, 2011a). More recently, Filippova, Alfonseca, et al., (2015) successfully used LSTMs for sentence compression on a large scale parallel dataset. We do not review the literature here, and only compare to Filippova, Alfonseca, et al., (2015).

Our contributions

- We present a novel multi-task learning approach to sentence compression using labelled data for sentence compression and a disjoint eye-tracking corpus.
- Our method is fully competitive with state-of-the-art across three corpora.
- Our code is made publicly available at https://bitbucket.org/soegaard/gaze-mtl16.

6.2 Gaze during reading

Readers fixate longer at rare words, words that are semantically ambiguous, and words that are morphologically complex (Rayner, Pollatsek, et al., 2012). These are also words that are likely to be replaced with simpler ones in sentence simplification, but it is not clear that they are words that would necessarily be removed in the context of sentence compression.

Demberg and Keller, (2008) show that syntactic complexity (measured as dependency locality) is also an important predictor of reading time. Phrases that are often removed in sentence compression—like fronted phrases, parentheticals, floating quantifiers, etc.—are often associated with non-local dependencies. Also, there is evidence that people are more likely to fixate on the first word in a constituent than on its second word (Hyönä and Pollatsek, 2000). Being able to identify constituent borders is important for sentence compression, and reading fixation data may help our model learn a representation of our data that makes it easy to identify constituent boundaries.

In the experiments below, we learn models to predict the first pass duration of word fixations and the total duration of regressions to a word. These two measures constitute a perfect separation of the total reading time of each word split between the first pass and subsequent passes. Both measures are described below. They are both discretized into six bins as follows with only non-zero values contributing to the calculation of the standard deviation (SD):

0: measure = 0 or
1: measure < 1 SD below reader's average or
2: measure < .5 SD below reader's average or
3: measure < .5 above reader's average or
4: measure > .5 SD above reader's average or
5: measure > 1 SD above reader's average

first pass duration measures the total time spent reading a word first time it is fixated, including any immediately following re-fixations of the same word. This measure correlates with word length, frequency and ambiguity because long words are likely to attract several fixations in a row unless they are particularly easily predicted or recognized. This effect arises because long words are less likely to fit inside the fovea of the eye. Note that for this measure the value 0 indicates that the word was not fixated by this reader.

regression duration measures the total time spent fixating a word after the gaze has already left it once. This measure belongs to the group of late measures, i.e., measures that are sensitive to the later cognitive processing stages including interpretation and integration of already decoded words. Since the reader by definition has already had a chance to recognize the word, regressions are associated with semantic confusion and contradiction, incongruence and syntactic complexity, as famously experienced in garden path sentences. For this measure the value 0 indicates that the word was read at most once by this reader.

See Table 8 for an example of first pass duration and regression duration annotations for one reader and sentence.

Words	First Pass	Regressions
Are	4	4
tourists	2	0
enticed	3	0
by	4	0
these	2	0
attractions	3	0
threatening	3	3
their	5	0
very	3	3
existence	3	5
?	3	5

Table 8: Example sentence from the Dundee Corpus

6.3 SENTENCE COMPRESSION USING MULTI-TASK DEEP BI-LSTMS

Most recent approaches to sentence compression make use of syntactic analysis, either by operating directly on trees (Cohn and Lapata, 2008, 2009; Filippova and Strube, 2008; Nomoto, 2007; Riezler et al., 2003) or by incorporating syntactic information in their model (Clarke and Lapata, 2008; McDonald, 2006). Recently, however, Filippova, Alfonseca, et al., (2015) presented an approach to sentence compression using LSTMs with word embeddings, but without syntactic features. We introduce a third way of using syntactic annotation by jointly learning a sequence model for predicting CCG supertags, in addition to our gaze and compression models.

Bi-directional recurrent neural networks (bi-RNNs) read in sequences in both regular and reversed order, enabling conditioning predictions on both left and right context. In the forward pass, we run the input data through an embedding layer and compute the predictions of the forward and backward states at layers $0, 1, \ldots$, until we compute the softmax predictions for word i based on a linear transformation of the concatenation of the of standard and reverse RNN outputs for location i. We then calculate the objective function derivative for the sequence using cross-entropy (logistic loss) and use backpropagation to calculate gradients and update the weights

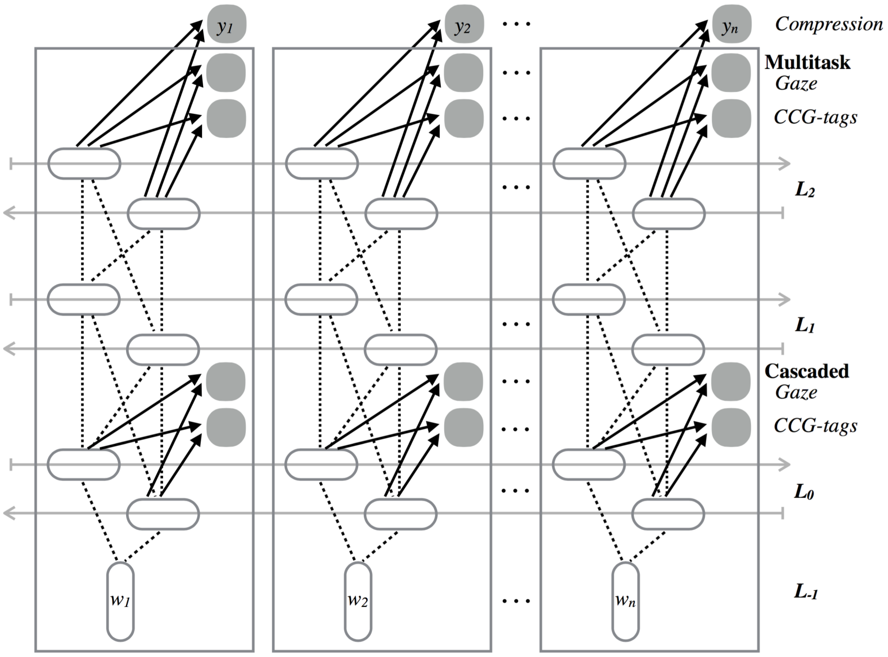


Figure 7: Multitask and cascaded bi-LSTMs for sentence compression. Layer L_{-1} contain pre-trained embeddings. Gaze prediction and CCG-tag prediction are auxiliary training tasks, and loss on all tasks are propagated back to layer L_0.

accordingly. A deep bi-RNN or k-layered bi-RNN is composed of k bi-RNNs that feed into each other such that the output of the ith RNN is the input of the $i + 1$th RNN. LSTMs (Hochreiter and Schmidhuber, 1997) replace the cells of RNNs with LSTM cells, in which multiplicative gate units learn to open and close access to the error signal.

Bi-LSTMs have already been used for fine-grained sentiment analysis (Liu, Joty, and Meng, 2015), syntactic chunking (Huang, Xu, and Yu, 2015), and semantic role labeling (Zhou and Xu, 2015). These and other recent applications of bi-LSTMs were constructed for solving a single task in isolation, however. We instead train deep bi-LSTMs to solve additional tasks to sentence compression, namely CCG-tagging and gaze prediction, using the additional tasks to regularize our sentence compression model.

Specifically, we use bi-LSTMs with three layers. Our baseline model is simply this three-layered model trained to predict compressions (encoded as label sequences), and we consider two extensions thereof as illustrated in Figure 7. Our first extension, Multi-task-LSTM, includes the gaze prediction task during training, with a separate logistic regression classifier for this

S:	Regulators Friday shut down a small Florida bank, bringing to 119 the number of US bank failures this year amid mounting loan defaults.
T:	Regulators shut down a small Florida bank
S:	Intel would be building car batteries, expanding its business beyond its core strength, the company said in a statement.
T:	Intel would be building car batteries

Table 9: Example compressions from the Google dataset. S is the source sentence, and T is the target compression.

purpose; and the other, Cascaded-LSTM, predicts gaze measures from the inner layer. Our second extension, which is superior to our first, is basically a one-layer bi-LSTM for predicting reading fixations with a two-layer bi-LSTM on top for predicting sentence compressions.

At each step in the training process of Multi-task-LSTM and Cascaded-LSTM, we choose a random task, followed by a random training instance of this task. We use the deep LSTM to predict a label sequence, suffer a loss with respect to the true labels, and update the model parameters. In Cascaded-LSTM, the update for an instance of CCG super tagging or gaze prediction only affects the parameters of the inner LSTM layer.

Both Multi-task-LSTM and Cascaded-LSTM do multi-task learning (Caruana, 1993). In multi-task learning, the induction of a model for one task is used as a regularizer on the induction of a model for another task. Caruana, (1993) did multi-task learning by doing parameter sharing across several deep networks, letting them share hidden layers; a technique also used by Collobert et al., (2011) for various NLP tasks. These models train task-specific classifiers on the output of deep networks (informed by the task-specific losses). We extend their models by moving to sequence prediction and allowing the task-specific sequence models to also be deep models.

	Sents	Sent.len	Type/token	Del.rate
		Training		
Ziff-Davis	1000	20	0.22	0.59
Broadcast	880	20	0.21	0.27
Google	8000	24	0.17	0.87
		Test		
Ziff-Davis	32	21	0.55	0.47
Broadcast	412	19	0.27	0.29
Google	1000	25	0.42	0.87

Table 10: Dataset characteristics. Sentence length is for source sentences.

6.4 EXPERIMENTS

6.4.1 *Gaze data*

We use the Dundee Corpus (Kennedy, Hill, and Pynte, 2003) as our eye-tracking corpus with tokenization and measures similar to the Dundee Treebank (Barrett, Agić, and Søgaard, 2015). The corpus contains eye-tracking recordings of ten native English-speaking subjects reading 20 newspaper articles from *The Independent*. We use data from nine subjects for training and one subject for development. We do not evaluate the gaze prediction because the task is only included as a way of regularizing the compression model.

6.4.2 *Compression data*

We use three different sentence compression datasets, Ziff-Davis (Knight and Marcu, 2002), Broadcast (Clarke and Lapata, 2006), and the publically available subset of Google (Filippova, Alfonseca, et al., 2015). The first two consist of manually compressed newswire text in English, while the third is built heuristically from pairs of headlines and first sentences from newswire, resulting in the most aggressive compressions, as exemplified in Table 9. We present the dataset characteristics in Table 10. We use the datasets as released by the authors

LSTM	Gaze	Ziff-Davis	Broadcast			Google
Baseline		0.5668	0.7386	0.7980	0.6802	0.7980
Multitask	FP	0.6416	0.7413	0.8050	0.6878	0.8028
	Regr.	0.7025	0.7368	0.7979	0.6708	0.8016
Cascaded	FP	0.6732	**0.7519**	0.8189	**0.7012**	**0.8097**
	Regr.	**0.7418**	0.7477	**0.8217**	0.6944	0.8048

Table 11: Results (F_1). For all three datasets, the inclusion of gaze measures (first pass duration (FP) and regression duration (Regr.)) leads to improvements over the baseline. All models include CCG-supertagging as an auxiliary task. Note that Broadcast was annotated by three annotators. The three columns are, from left to right, results on annotators 1–3.

and do not apply any additional pre-processing. The CCG supertagging data comes from CCGbank,[1] and we use sections 0-18 for training and section 19 for development.

6.4.3 *Baselines and system*

Both the baseline and our systems are three-layer bi-LSTM models trained for 30 iterations with pre-trained (Senna) embeddings. The input and hidden layers are 50 dimensions, and at the output layer we predict sequences of two labels, indicating whether to delete the labeled word or not. Our baseline (Baseline-LSTM) is a multi-task learning bi-LSTM predicting both CCG supertags and sentence compression (word deletion) at the outer layer. Our first extension is Multitask-LSTM predicting CCG supertags, sentence compression, and reading measures from the outer layer. Cascaded-LSTM, on the other hand, predicts CCG supertags and reading measures from the initial layer, and sentence compression at the outer layer.

6.4.4 *Results and discussion*

Our results are presented in Table 11. We observe that across all three datasets, including all three annotations of Broadcast, gaze features lead to improvements over our baseline 3-layer

[1] http://groups.inf.ed.ac.uk/ccg/

bi-LSTM. Also, Cascaded-LSTM is consistently better than Multitask-LSTM. Our models are fully competitive with state-of-the-art models. For example, the best model in Elming et al., (2013) achieves 0.7207 on Ziff-Davis, Clarke and Lapata, (2008) achieves 0.7509 on Broadcast,[2] and the LSTM model in Filippova, Alfonseca, et al., (2015) achieves 0.80 on Google with much more training data. The high numbers on the small subset of Google reflects that newswire headlines tend to have a fairly predictable relation to the first sentence. With the harder datasets, the impact of the gaze information becomes stronger, consistently favouring the cascaded architecture, and with improvements using both first pass duration and regression duration, the late measure associated with interpretation of content. Our results indicate that multi-task learning can help us take advantage of inherently noisy human processing data across tasks and thereby maybe reduce the need for task-specific data collection.

ACKNOWLEDGMENTS

Yoav Goldberg was supported by the Israeli Science Foundation Grant No. 1555/15. Anders Søgaard was supported by ERC Starting Grant No. 313695. Thanks to Joachim Bingel and Maria Barrett for preparing data and for helpful discussions, and to the anonymous reviewers for their suggestions for improving the paper.

[2] On a "randomly selected" annotator; unfortunately, they do not say which. James Clarke (p.c) does not remember which annotator they used.

7

STUDY 4: PREDICTING LEXICAL COMPLEXITY FROM USER GAZE

ABSTRACT

We show that information available in near-real-time from individual fixations during reading can be leveraged for lexical complexity detection. This has promising implications for NLP system evaluation.

7.1 INTRODUCTION

Reading behaviour, in particular recordings of readers' gaze patterns, reflect aspects of text complexity, from surface traits such as word length, spelling regularity and frequency (Rayner and Pollatsek, 2013), to functional distinctions such as content/function word status (Carpenter and Just, 1983; Rayner and Duffy, 1986), parts-of-speech (POSs) (Barrett and Søgaard, 2015a), and specific high-level syntactic features like underspecification in garden-path sentences (Malsburg and Vasishth, 2013).

With recent advances in eye tracking technology (Krafka et al., 2016; Ooms et al., 2015, inter alia) this data is becoming increasingly easily available, making recruitment of remote, skilled readers with access to a webcam a plausible future alternative for natural language processing system evaluation and annotation purposes.

This paper explores how a signal from gaze recordings alone can be exploited for detecting cases of lexical complexity from individual readings of individual stimuli. We propose to learn from gaze data to detect individual words that display high complexity. This is achieved by recording readers reading short microblog texts with non-standard language use and then detecting the complex words from only local features of the gaze recording, relying on a detail-rich representation of partial, individual fixation sequences, i.e. sub-paths of the full recorded scanpaths.

The motivation for exploring this combination of task and data is that detecting and locating excessively complex text in the output of natural language processing (NLP) systems could benefit users, but current measures of text complexity fail when applied outside narrowly-defined standard English, for which curated word lists and grammaticality tests can be relied on (Louis, 2012; Siddharthan and Katsos, 2012; Vajjala and Meurers, 2013).

It is also important to acknowledge that while annotators can be trained to annotate according to linguistic formalisms with high agreement, we cannot train skilled readers to acquire conscious insight into the automatic, unconscious cognitive processes involved in reading, and estimate the mental processing cost incurred per word. This is reflected by low agreement scores on lexical complexity annotations and difficulty in training systems to replicate such annotations (Specia, Jauhar, and Mihalcea, 2012).

EYE MOVEMENT CONTROL The motivation for using gaze data is the direct influence of text traits on eye movements. Gaze behaviours in reading are made up of saccades (rapid eye movements scanning the text), interspersed with fixations (short periods of almost stationary gaze direction during which the central 1-2° of the visual field is sampled for cognitive processing).

Existing gaze measures in psycholinguistic research are designed to robustly detect effects over multiple carefully-controlled stimuli, averaged over pools of homogeneous subjects to cancel out random noise, in order to ultimately test hypotheses about how cognition works.

In contrast, NLP system evaluation is ultimately about being able to judge the performance of some system against other systems. This goal poses the following requirements: that the smaller the unit of evaluation (e.g. full test set $\gg$ full text $\gg$ sentence $\gg$ word), the more specific an evaluation of systems' comparative strengths and weaknesses can be made. Due to the aggressive aggregation, using psycholinguistic gaze measures directly leaves us with only very coarse-grained evaluations (Doherty, O'Brien, and Carl, 2010; Klerke and Søgaard, 2013; Rello, Kanvinde, and Baeza-Yates, 2012; Stymne, Danielsson, et al., 2012). In addition, better evaluations are possible with more annotated test data, which lead to a preference for evaluation frameworks with fewer annotators whenever justifiable, another proposition at odds with traditional psycholinguistic metrics. So while some attempts succeed at using aggregate gaze statistics for NLP tasks by averaging over all readers' readings of the same text (Barrett and Søgaard, 2015a; Blache and Rauzy, 2011), or leverage a large collection of overlapping individual scanpaths (Klerke, Goldberg, and Søgaard, 2016), from an application-centered perspective, replacing a few expert annotators with many readers is not an appealing alternative.

If gaze data is to provide an attractive alternative to current annotation practices, it is necessary to identify a representation which allow us to leverage the information presumably contained in individual readers' local gaze behaviour.

CONTRIBUTION In this paper we: **(1)** demonstrate that systematic cognitive responses to lexical complexity are present and learnable from behavioural data sources; **(2)** present and evaluate a novel language-independent approach to evaluating lexical complexity which can be applied in near-real-time to a single individual's reading of the text; and **(3)** make the collected corpus of microblog reading and code for pre-processing available at `annonymised`.

7.2 BACKGROUND

It is often relevant to detect and handle lexical complexity, because users of NLP systems, and in particular learners, can benefit from systems that are able to monitor and react to a reader's progress. Learners suffer the most from overly complex text because their cognitive resources are already tasked with meeting a given learning objective, so actively handling unnecessary text complexity can, by freeing up cognitive resources, positively affect learning outcomes. This is a primary motivation for readability research in the NLP community (Carroll et al., 1999; François and Miltsakaki, 2012; Petersen and Ostendorf, 2007; Rello, Baeza-Yates, et al., 2013).

From a different perspective, many researchers in NLP now work with text that lacks critical editing, including microblogs, user reviews, crowd-sourced subtitles, comment threads, and other user-generated content. A primary challenge for systems processing these types of text, is that they display high ratios of non-canonical language use and domain-specific jargon, often collectively treated as "noise" by systems that were originally trained on canonical text, such as news corpora (Baldwin et al., 2013; Han and Baldwin, 2011). In contrast, skilled readers both produce and make sense of these text types. However, the fast pace of language change mastered by skilled readers is not feasible to emulate simply through continuous corpus creation. This is why we propose to detect in near-real-time the lexical items that are most likely to obstruct readers, through learning directly from recorded reading behaviour.

7.2.1 *Clear-cut cases of lexical complexity*

In our experiments we use word frequency as the primary indicator of lexical complexity (Cmpl). In addition to word frequency, we focus on three indicators of lexical complexity which can be annotated automatically and serve as a proxy for the wide range of factors that may obstruct reading, namely: (1) Len = token length; (2) Ppl = prefix-spelling surprisal (or "perplexity"); and (3) Mrk = whether or not a given token is prefixed with a visually-salient Twitter-style markup character (i.e. @ or #). See Section 7.3.1 for details of the sampling method.

The factors Len, Mrk and Ppl are chosen as indicators because they have consistently been found to increase reading times and affect task performance (Carpenter and Just, 1983; Rayner, 1998, 2009). While other factors, such as garden path constructions and the presence of ambiguous words used in their less common meaning also can increase reading times (Slattery et al., 2013; Spivey-Knowlton, Trueswell, and Tanenhaus, 1993, for instance), these aspects are not included as sampling factors because they are challenging to annotate and have less consistent effects on reading time.

In addition, annotations of the factors Len, Mrk and Ppl are all simple to obtain automatically, and with some control over the factors, we are able to perform a factorial analysis to explore the hypothesis that microblog reading is affected by these factors in a way that resembles existing data from the psycholinguistic literature.

7.3 METHOD

To be able to test whether near-real-time gaze patterns reflect lexical complexity of relevance as NLP system feedback, we recorded the gaze of readers reading a sample of tweets containing the above examples of lexical complexity. After verifying that tweet-reading is affected by lexical complexity, we train models to detect whether individual fixations land on complex words. Finally, we analyse how well these models generalise to complexity detection in general text. This section details how each of these steps are conducted.

7.3.1 *Tweet sampling*

Tweets consisting of 70–100 characters were sampled to ensure that they fit on one line while being long enough to display variable syntax. In particular, English-language tweets were selected,[1] and then sub-sampled into 16 distinct groups based on the general and factor-specific criteria outlined below. The final collection consisted of 244 tweets.[2]

Using a reference corpus of subtitles,[3] which resemble tweets better than newswire text, all high-frequency words (in the top 15K) were considered non-Cmpl tokens. Cmpl tokens, in contrast, were very low-frequency or OOV (outside the top 50K). The sampled tweets contain, respectively, none, one or several non-canonical Cmpl tokens.

For contexts with exactly one Cmpl token, the sample contains 15 tweets of each of the full Cartesian product of combinations of factors, allowing us to perform a full two-level factorial analysis on this sub-sample, to validate that our hypothesised clear-cut examples of lexical complexity produce effects on gaze measures comparable to effects reported in the psycholinguistic literature.

For each factor we define the levels high and low, as specified in Table 12. Mrk is a binary indicator of whether the token is prefixed with any of the salient markers. The high and low levels of the factors Len and Ppl are cut off to exclude a central section of their respective distribution which can prevent small effects from being swamped by noise in the analysis.

The factor Ppl is measured as character-based perplexity over the first five letters of the token (excluding markup, numbers and punctuation) in a 9-gram character language model (Stolcke, 2002).

[1] Sampled from a subset (10m) of a 57m in-house collection of tweets with the English language tag.

[2] Randomly selected from the larger sub-sample and manually filtered to exclude offensive language and tweets with passages of three or more tokens of non-English language or all-upper case writing.

[3] https://invokeit.wordpress.com/frequency-word-lists/

Factor level		Description	Examples
Len	low	4–7 characters (including markup and punctuation)	*boringg, nuch, #alward, @Raaab*
	high	11–21 characters (including markup and punctuation)	*Gainesville, UnPoppovich, #SoundsGood*
Mrk	low	without # or @ prefix	*Drob, DramaFever, bandwagoners*
	high	with # or @ prefix	*#SoundsGood, #socute, @ItzGary-Wong*
Ppl	low	Perplexity of the first five alphabetic characters in top tertile of letter fivegrams	*kolors, #bruins, jordanlwatson, @spicykimchi*
	high	Perplexity of first five alphabetic characters in bottom tertile of letter fivegrams	*wkwk, @tjmpb, hoodboogers, #howaboutno*

Table 12: The three factors: Len = length, Mrk = salient visual marking, and Ppl = spelling surprise. In multi-Cmpl contexts, the high level trait must affect at least one Cmpl while the low level must hold for all Cmpl tokens in the tweet.

7.3.2 *Gaze recording*

Below, we detail the stimuli, apparatus, participants and experimental procedure, before presenting the factorial analysis of the recorded data, which serves to verify that the sampling design for sampling clear-cut examples of naturally occurring lexical complexity produces effects on gaze behaviour comparable to effects observed in fully controlled psycholinguistic experimental designs.

Participants read the collection of tweets one message at a time, and manually categorised each tweet into one of 10 standard document categories.[4] Participants were encouraged to proceed as quickly as possible and finish the experiment within 50 minutes.

[4] The ten top-level categories of the Dewey Decimal System (Dewey, 1891). While most tweets don't fit these categories perfectly, the task required readers to read for comprehension.Reading and categorising one tweet constitutes one trial, totalling 244 self-paced trials grouped into eight blocks with self-paced pauses in-between and three training-trials which were completed with the experimenter present.

Eye movements were recorded with a Tobii X2-60 tracker.[5] Out of 25 recruited participants, recordings from 18 were included in the analyses.[6]

GAZE EFFECTS We perform a factorial analysis to assess both the main effects of individual factors and their interactions (Box, S. J. Hunter, and W. G. Hunter, 2005). With three two-level factors, the analysis can be conceptualised as a cube covering the space of variance over all three factors (see Figure 8). Here, the lower left front corner represents measures on Cmpl tokens with all factors at their low level i.e. every low-frequency token (Cmpl) which is not saliently marked (low Mrk), starts with an unsurprising string of letters (low Ppl) and is between four and seven characters long (low Len). The opposite top right rear corner represents measures on Cmpl tokens with all factors at their high level.

Each of the eight corners — representing the Cartesian product of the three two-level factors — is represented by 15 tweets. All possible factor interactions can then be estimated by directly comparing measures from the corners, edges or planes of the cube that represent the relevant factor combinations. We report effects on fixation count, total gaze duration and the regression proportion, i.e. the proportion of the total gaze duration that was spent as part of a backwards-directed saccade, into the Cmpl token. All fixations longer than three seconds are excluded as outliers.

For each measure we take the median of each participant's reading of the 15 tweets sampled for each combination of factor-levels. Each reader thus constitutes an independent rerun of the

[5] The stimuli were presented in black letters in the typeface Verdana with a letter size of ca. .4° visual angle (20 px) on a light grey background with 150 pixels margins (ca. 3° visual angle). Presentation order counterbalancing and fixation detection were carried out using standard settings in Tobii Studio (Studio, 2008), and the IV-T fixation filter (Olsen, 2012).

[6] Two subjects did not finish the experiment within the set time, one recording was interrupted because of technical problems, two recordings had more than 30% lost samples, the tracker could not be calibrated for one participant, and two participants did not comply with the task instructions. All participants received a movie pass as compensation.The average age of participants was 31.6 years (SD = 12.3 years) and gender was balanced (9 each of male and female). English was the only reading language of 6 participants, 8 read one other language, and 2 read 2–3 other languages. They self-assessed their English skills as Expert (11), Very Good (4) or Good (3).Participants' familiarity with reading Twitter messages varied from none (5), to monthly (2) and weekly, (4) to daily readers (5). Few reported that they tweet themselves, with only 7 tweeting monthly or less, and 11 non-tweeters.

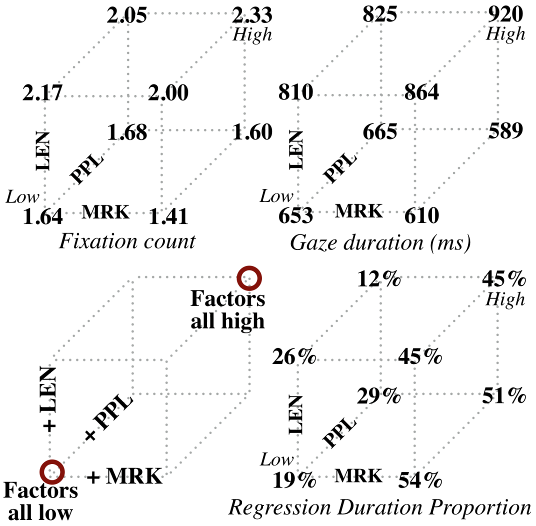


Figure 8: Median raw gaze measures on Cmpl tokens for all factor combinations.

full experiment. The variance between the independent reruns determines the standard error (SE), which in turn determines the effect size needed to reach a desired p level (±2.3 SE at $p < 0.05$) (Box, S. J. Hunter, and W. G. Hunter, 2005).

Figure 8 shows the estimated medians of each measure for each combination of low and high factor-levels and Table 13 reports the main and interaction effects as percent change relative to the all-low-factor Cmpl token (the lower left corner of the boxes in Figure 8), centered per participant and reported as percent change.

The factorial analysis demonstrates that the factors affect gaze systematically and differently during tweet reading. Starting by comparing the results on tokens in tweets without Cmpl tokens (shown in the top-row of Table 13), the number of fixations per token (1.08) is similar to the results reported for English reading in Liversedge et al., (2016). They find a lower re-fixation rate and gaze duration of 23% and 211 ms respectively, whereas our data shows 32% re-fixations and total gaze durations of 372 ms per word. However, this could be due to the categorisation task employed, which requires some semantic interpretation, which is our motivation for also comparing our results to those of Schotter et al., (2014) on a more demanding semantic proofreading task. Here we find that our results are on par with their reported total gaze durations of 341–476ms per word, dependent on the predictability and frequency of the token. The increases we observe on Cmpl

	Fixations	Gaze time	Regr. time
*non-*Cmpl *tokens*	*1.08 fix.*	*372 ms*	*32%*
all-low Cmpl *tokens*	*1.64 fix.*	*653 ms*	*19%*
Effect in percent change on Cmpl token reading.			
Len	**57.0** ± 1.8	**43.0** ± 2.4	−**12.5** ± 1.6
Mrk	−**5.2** ± 1.8	1.4 ± 2.4	**58.5** ± 1.6
Ppl	0.3 ± 1.8	0.2 ± 2.4	−0.4 ± 1.6
Len × Mrk	**11.5** ± 1.8	**16.0** ± 2.4	−2.0 ± 1.6
Len × Ppl	**8.5** ± 1.8	4.7 ± 2.4	−**11.6** ± 1.6
Mrk × Ppl	1.9 ± 1.8	0.2 ± 2.4	0.5 ± 1.6
Len × Mrk × Ppl	**15.9** ± 1.8	**13.9** ± 2.4	**12.4** ± 1.6

Table 13: Raw measures of fixation count, gaze duration and regression duration on non-Cmpl tokens and Cmpl tokens with all factors at their low value in italics. Below; factors' main and interaction effects and SE as percent change relative to the all-low-factor Cmpl tokens. Boldface indicates that the effect is significant at $p < 0.05$

tokens on both fixation count and gaze duration are larger than typically reported, which may be explained by the fact that, unlike the low-frequency target words and non-words typically studied in the literature, most of the Cmpl tokens are specific to the microblog genre, and while usually not dictionary words, are often constructed in a way that allow and encourage decoding, e.g. by CamelCasing.

When Cmpl token length increases, the primary apparent strategy is to increase both number and duration of fixations, by 57% and 43% respectively, while decreasing regressions. In contrast, when Cmpl tokens are prefixed with salient visual markers (and are not long or with unusual initial spelling), a different strategy appears to come into play: delaying attention to a re-reading phase, with a 58.5% increase in the proportion of gaze time spent after a backward directed saccade. This indicates that the Mrk factor can possibly lead readers to favour skipping these salient tokens entirely on first encounter. Thirdly, in this dataset the factor Ppl has no significant main effect, but contributes significantly through factor interaction, dampening the main effects of both of the two other factors.

Though the above factors are not our main interest, they confirm that our factors, through our instantiation of clear-cut examples of lexical complexity, form a valid proxy for reading obstruction.

7.3.3 *Complexity prediction*

In order to pinpoint lexical complexity from gaze behaviour, we need to move beyond assessing general effects of controlled text traits on gaze behaviour, as done in psycholinguistics as evidence of cognitive processes. We therefore propose to treat gaze behaviour as a binary sequence tagging problem: for each item in a sequence, we label it as belonging either to the non-complex or the complex class.

However, in sequence prediction tasks in NLP, the sequence usually follows the linear order of the text. However, as gaze behaviour does not follow the linear order of the text, but rather forms a sequence of fixations ordered in time (which are only piece-wise parallel to the printed word sequences), we propose to learn to label the fixation sequence in time, i.e. the order that individual eye movements were executed in, labelling each fixation as reflecting the reading of a word from the complex or the non-complex class.

Under this formulation, we represent each fixation as a vector of information about the previous and following fixations, effectively representing the fixated word by aspects of the current fixation and a window over the neighbouring fixations, rather than the neighbouring words. Note that this means any particular token can be represented multiple times in the dataset: once for each fixation for each reader, or not at all, in the case all participants happened to skip over that particular token.

Table 14 lists the features we extract for each fixation and its associated window. Notice that we use no lexical features and no features outside the current window in order to estimate just how predictive the local gaze behaviour sequence is of the text. This both prevents the model from memorising labels of concrete tokens between readings, and allows for near-real-time prediction from individual readings.

We use an asymmetric window covering the current fixation plus one previous and two following fixations, covering the timeframe in which the ongoing processing of a word can plausibly be influencing gaze behaviour (Rayner, Pollatsek, et al., 2012).

N	Feature	Description
	from (-1; +2) fixation window	
4	Saccade-in length and directions	Change in word position from previous fixation.
4	Relative saccade-in length and direction	Distance from currently fixated word.
4	Saccade-in direction	Word position of previous fixation is behind, further ahead or same as current.
4	Fixation duration	Fixation duration (quartiles).
4	Relative fixation duration	Change relative to individual median first fixation duration (quartiles).
4	Fixation duration change	Fixation duration from previous fixation is higher, lower or same.
	from current fixation only	
1	Fixation count	Number of fixations on the current word so far.
1	Word position	The currently fixated word position (per display).
2	BOS, EOS	Beginning and end of displayed text.
	combinations	
6	Full window	Concatenation of fixation-features over full window.
66	Feature pairs	Concatenation of pairs of fixation-features at same window position.
28	Window pairs	Concatenation of full window of two fixation-features.

Table 14: Features derived from fixations in the gaze window from -1 to +2 fixations from the current. Total number of features is $N = 128$

	Twitter (T)	Dundee (D)
Displays	800	244
Readers	10	18
Scanpaths	3,629	7,899
Fixations	48,023	407,105

Table 15: Datasets used for prediction.

Test set	Positive label
T	All Cmpl tokens in Twitter dataset
D(l)	Token length $>$ 7 characters, Dundee dataset
D(lf)	Token frequency $<$ top 50K, Dundee dataset
D(llf)	Intersection of D(l) and D(lf)

Table 16: Test set overview.

predicting complexity from individual gaze behaviour. To perform the sequence prediction we train a conditional random field (CRF: Lafferty, McCallum, and Pereira, (2001)).[7] We use 90% of the scanpaths from the full Twitter dataset (see Table 15) for training, and the remaining 10% for testing. All reported CRF models were trained for a maximum of 500 iterations with L2-regularised stochastic gradient descent ($c = .1$), discarding features that were observed < 5 times. Bootstrap sampling was used for significance testing.

We compare model performance to a chance baseline, reflecting the positive label frequency in each dataset, and two other baselines: (1) dur-BL, which predicts all long fixation durations (upper quartile) as falling on Cmpl tokens; and (2) fix-BL, which predicts that all fixations returning to a word for the third time or more is targeting a complex word. These are hard baselines, as it is physiologically impossible to see all letters clearly in a single fixation on long words. While it may not be necessary to look at all letters in common and predictable words, any previously unseen word would be impossible to decode without spending several fixations on it, pushing up both duration and fixation count.

[7] We used the CRFSuite (Okazaki, 2007) implementation via github.com/TeamHG-Memex/sklearn-crfsuite

TESTING ON READING IN OTHER DOMAINS Since the tweets were sampled to provoke maximally diverging gaze responses, we set up a test to gauge whether the learned patterns appear and are associated with lexical complexity in a standard reading scenario, using the Dundee corpus (Kennedy, Hill, and Pynte, 2003) where few words deviate as notably as the Twitter examples. The Dundee data is labelled in three different ways with increasing similarity to the CMPL labels in the Twitter data as described in Table 16. 10% of the scanpaths are held out for testing.

7.4 RESULTS

Our ultimate research question is whether we can use non-aggregated gaze representations to predict token-level difficulty. Table 9 presents results based on training on Twitter data and on the larger Dundee corpus. With the high prevalence of CMPL tokens in the Twitter dataset and a detailed representation of local behaviour surrounding individual fixations, we can see that it is possible to learn to detect systematic gaze reactions. For the models trained on the much larger and less extreme Dundee corpus, only the fairly frequent long words in D(L) provide a gaze behaviour signal that the model is able to learn.

To measure the generalisability of the Twitter-based model, we test it on out-of-domain data from the Dundee corpus, as shown in Table 10. We find that long and low-frequency words in a canonical reading scenario are identified at above chance levels, and above models trained in-domain (shown in Table 9).

Recall that the three versions of the Dundee test sets, D(L), D(LF), and D(LLF), are the same data, only with different labelling strategies, which means we can compare scores on all three to assess whether the learned gaze patterns reflect either length or low frequency, or both. This comparison reveals that the Twitter model has learned some of both, and while the highest precision is reached on long words, the high recall scores on low-frequency tokens indicate that it has learned some gaze reaction patterns typical for these types of lexical complexity, but that the learned patterns are also indicative of something else which is evidently not captured in the naive automatic labelling strategies for the Dundee datasets. We discuss this further in the error analysis in the following section.

	Baselines, F1			Model performance	
Data	%-BL	dur-BL	fix-BL	F1	Prec/Rec
T	19.9	29.0*	27.9*	**44.1***	55.8 / 36.5
D(l)	30.1	30.7	18.6	**42.9***	50.6 / 37.2
D(lf)	4.6	10.7*	**12.3***	5.0	12.0 / 3.2
D(llf)	3.4	8.3*	**11.6***	6.3*	12.5 / 4.2

Figure 9: Baselines and in-domain model performance. Improvement over chance baseline is marked by * ($p < .01$), and boldface indicates the best score on each test set.

	Twitter model	
Test set	F1	Prec/Rec
D(l)	34.4*	48.5 / 26.7
D(lf)	11.2*	6.8 / 31.9
D(lff)	9.4*	5.5 / 34.4

Figure 10: Twitter model tested out-of-domain. Significant improvements over a chance baseline are marked by * ($p < .01$).

7.5 Error analysis and discussion

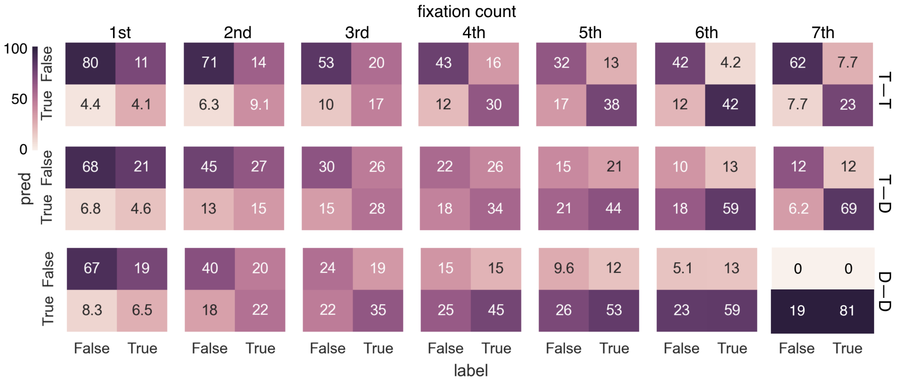
To give a more nuanced account of what the models are learning, we compare their concrete predictions on examples as shown in Table 17, and identify conditions under which their performance diverges, as shown in Figure 11.

In the examples, unambiguous true and false positive predictions are shown once per reader per word, i.e. whenever all of one reader's fixations on a given word are classified as falling on an Cmpl token. False negatives are also annotated, i.e. when all fixations by one reader on a given target-word are classified as non-Cmpl. As the models are free to chose different labels for subsequent fixations to the same word, words with more fixations are less likely to be unambiguously labelled. Examples with a relatively large portion of unambiguous predictions are shown to be able to highlight where the models and participants (dis-)agree most consistently. Note that the number of readers varies per sentence.

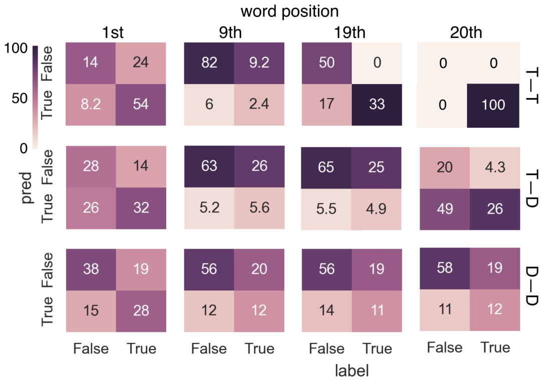
T (4)	Im a girl that dont believe in **nuch**$_{oooo}$ but i be dam if i dont believe in$_{x}$ us
T (5)	**Cntt**$_{x}$ believe **thatt**$_{oooo}$ **iThinkk**$_{ooo}$ **#Oomfs**$_{oooo}$ is **cutee**$_{ooo}$ **.....Omgg**$_{oooo}$ **watts goinn**$_{ooo}$ on n **Lifee**$_{ooo}$
D (3)	animal$^{x}_{x}$ **population**$^{xx}_{oo}$ has been under **increasing**$^{xoo}_{xoo}$ threat, from wayward$^{x}_{o}$ **fishermen,**$^{ooo}_{ooo}$ or from feral dogs, goats and rats^{x} **escaping**$^{xx}_{xoo}$ from human^{x} **settlements**$^{xoo}_{xx}$ or boats. **Recently**^{o} the$^{x}_{x}$ Ecuador^{x} **authorities**$^{xo}_{xoo}$ started$_{x}$ trying$_{x}$ to **eradicate**$^{xxx}_{oo}$ the **invading**$^{ooo}_{xoo}$ **species,**$^{xoo}_{xxo}$ and have now$_{x}$ **succeeded**$^{xoo}_{oo}$ freeing^{x} some entire islands from **non-native**$^{xx}_{o}$ **creatures.**$^{x}_{o}$ A giant^{x} **tortoise**$^{xo}_{oo}$ **breeding**$^{o}_{xo}$ **programme**$^{xo}_{oo}$ is in place, rearing infants$_{x}$ in a
D (2)	about **contemporary**$^{oo}_{oo}$ **relationships**$^{oo}_{oo}$ that's right for now, but it's not going to be right for ever. **"Discounting**^{o} the oddity finding$_{xx}$ myself **bracketed**$^{xo}_{oo}$ with her as a^{x} **contemporary**^{o} of J Fowles (b1926) P **Fitzgerald**$^{oo}_{oo}$ (b1916) – which rather^{x} **stretches**$^{o}_{o}$ the notion of **contemporaneity**$^{xx}_{o}$ beyond the bounds of my sense – I find this úmble **dismissal**$^{oo}_{oo}$ of her craft^{x} by Joanna as being mere **"acuity"**$^{oo}_{oo}$ quite

Table 17: Examples of text displays with unambiguous classifications of true and false positives (*x*) and false negatives (*o*) per reader. From Twitter (T) and Dundee (D) test sets with number of readers in parenthesis. Target Cmpl tokens are bold. Predictions are sub- or super-scripted for the model trained on T or D, respectively.

The most apparent pattern is that the Twitter model is reluctant to identify both long and atypical words as Cmpl tokens. However, it is possible that syntactic anomalies are triggering this effect, indicated by the identification of the word *finding* as the only consistently complex word in the context *Discounting the oddity finding myself bracketed* where the preposition *of* appears to have been omitted. This is consistent with the hypothesis that the Cmpl tokens in microblogs trigger a gaze strategy that supports creative guesswork, and that such a strategy might turn out to be more similar to what is used when disentangling obscure syntax in standard newswire text (where fewer odd words may be encountered). However, a comparison of model performance on the Dundee test set broken up by POS tags or dependency labels did not reveal notable differences between the models.



(a)



(b)

Figure 11: Performance confusion matrices per re-fixation (a) and per word position (b). Comparison of Twitter in-domain (T–T), Twitter out-of-domain (T–D) and Dundee in domain (D–D) using the D(L) test set. Each diagonal from top left depicts correct predictions in percent. Darker diagonals indicate better performance, whereas a dark horizontal bar indicates overfitting to the majority class. Shading is normalised per label (due to class-imbalance)

In Figure 11 we show two different breakdowns of model performance where the models differed notably, namely on the fixation count (i.e. whether it was the first, second, or later fixation to a given token) and on the word position (i.e. the fixated word's location at the display).

The pattern that emerges for the fixation count is that the Twitter model learns to distinguish very well between late fixations to Cmpl and non-Cmpl tokens (as seen on the in-domain test, top row) and is able to transfer this to the out-of-domain test set (middle row). In comparison, the Dundee model (bottom row) quickly overfits to the majority-class on late fixations. This pattern directly reflects how late fixations are too sparse in the Dundee data for that model to learn valuable complex patterns of behaviour which are however sufficiently common to be learned from the more extreme Twitter data.

For word positions, we see a different pattern, where sparsity of long Twitter texts during training has led the Twitter model to overfit this bit of information, a behaviour that may possibly be aggravated by a potentially biased distribution of Mrk tokens.

Importantly, we did not see notable differences between models across readers, which indicate that the learned behaviours were general, with differences across readers in the levels to which the identifiable behaviours are applied.

7.6 conclusion

We found that local gaze behaviour in individual readings reveals information about text at the lexical level, presenting a first approach to adapting eye movement data for near-real-time NLP system feedback. Using sequence prediction, we have shown that learned patterns in gaze behaviour generalise across reading scenarios.

This work has focused on identifying a salient subset of what is considered lexical complexity. This was a necessary limitation because broader definitions of lexical complexity are not agreed upon. Because the possible variations in skilled readers' gaze behaviour is limited, this approach has the potential to lead to a more consistent conceptualisation of lexical complexity, based on users' gaze response.

Other applications in NLP may benefit from a similar approach, such as text normalisation and simplification, where the contribution of individual edits is otherwise hard to assess, as well as in fluency evaluation of generative systems, where competing valid solutions are otherwise hard to rank.

In addition, through the comparability to psycholinguistic literature provided by the factorial analysis, this work serves to validate work on lexical complexity in psycholinguistics based on aggregate measures and presents evidence in support of the hypothesis that skilled readers apply distinct, highly responsive eye movement strategies during reading.

Part III

CONCLUSION

8 SUMMARY OF FINDINGS AND PERSPECTIVES

This thesis has tackled the question of whether reading behaviour can be leveraged for improving NLP systems' ability to handle text complexity.

The research presented here answers the central research question; whether readers' gaze behaviour reflect information about text complexity at sentence- and word-level. Furthermore, the present work also explores how gaze data can be employed for improving applications in NLP.

Study 1 shows that gaze measures reflect the kind of differences in sentences that are relevant for evaluating and improving NLP applications. The study compares readings of original and simplified sentences as well as automatically simplified sentences with and without grammatical errors.

In particular, while ungrammatical simplifications are short, leading to fewer fixations and shorter total reading times, increases in both proportion of regressions, reading time per word and the number of fixations per word are observed. This pattern uniquely describes the group of system-corrupted sentences in the data.

The experiments in Study 2 support the finding that eye movements reflect information about text complexity, specifically the quality of translations. When comparing a range of evaluations to a task-based measure of usability, reading time and fixation count are the only two measures that correlate significantly with usability.

Study 3 demonstrates that gaze information can improve the performance of a sentence compression model. The improvement is achieved by using gaze prediction as an auxiliary task in a deep-learning framework.

The deep learning model with auxiliary task prediction is a promising architecture which allows using disjoint datasets to learn deep RNNs for tasks where available training datasets are otherwise too small for deep learning.

The observed performance improvement from employing a low-level gaze data representation—which is a rich but noisy data source—encourages further exploration of representing detail rich behavioural data sources. This approach may prove superior to aggregating the noise away before using noisy sources of behavioural data in learning for NLP. Moreover the success of this representation points towards a potential for making better use of the effort spent recording gaze data, also where statistical power is low in a traditional sense.

Study 4 shows that representing gaze data at the level of individual eye movements provides learnable information about local lexical complexity. In addition, the learned gaze behaviour patterns are found to generalise across readers and reading scenarios.

The gaze representation demonstrates a new approach to learning from noisy, raw records of user behaviour. An interesting perspective is the possibility of allowing applications to respond to gaze behaviour in near-real-time. The gaze representation achieves this flexibility by describing each fixation in order of execution by a feature vector of information extracted only from a narrow sliding window over the recorded scanpath.

This work shows that it is possible and worthwile to bridge the divergent perspectives on text complexity between NLP and psycholinguistics.

In the experiments it proved challenging to insert a text-based notion of text complexity pervasive in NLP into a psycholinguistic experimental framework designed to model *the average reader*. In contrast, it was found to be both fruitful and promising to implant the notion of complexity as pertaining to *one reader's localized behaviour* into experimental frameworks from NLP.

BIBLIOGRAPHY

Amancio, Marcelo Adriano and Lucia Specia (2014). "An Analysis of Crowdsourced Text Simplifications." In: *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*.

Asmussen, Jørg (2001). "Korpus 2000." In: *Korpuslingvistik (NyS30)*.

Baldwin, Timothy, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang (2013). "How Noisy Social Media Text, How Diffrnt Social Media Sources?" In: *Proceedings of IJCNLP 2013*.

Barrett, Maria, Željko Agić, and Anders Søgaard (2015). "The Dundee Treebank." In: *Proceedings of The 14th International Workshop on Treebanks and Linguistic Theories*.

Barrett, Maria and Anders Søgaard (2015a). "Reading behavior predicts syntactic categories." In: *Proceedings of CoNLL 2015*.

— (2015b). "Using reading behavior to predict grammatical functions." In: *Proceedings of Workshop on Cognitive Aspects of Computational Language Learning, EMNLP 2015*.

Barzilay, Regina and Noemie Elhadad (2003). "Sentence alignment for monolingual comparable corpora." In: *Proceedings of EMNLP 2003*. ACL.

Bender, Emily M (2013). "Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax." In: *Synthesis Lectures on Human Language Technologies*.

Benjamin, Rebekah George (2012). "Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty." In: *Educational Psychology Review* 24.1.

Berg-Kirkpatrick, Taylor, Dan Gillick, and Dan Klein (2011). "Jointly learning to extract and compress." In: *Proceedings of ACL 2011*.

Bjornsson, Carl-Hugo (1983). "Readability of Newspapers in 11 Languages." In: *Reading Research Quarterly* 18.4.

Blache, Philippe and Stéphane Rauzy (2011). "Predicting Linguistic Difficulty by Means of a Morpho-Syntactic Probabilistic Model." In: *Proceedings of Pacific Asia Conference on Language, Information and Computation*. Vol. 25.

Bohnet, Bernd (2010). "Very high accuracy and fast dependency parsing is not a contradiction." In: *Proceedings of the 23rd International Conference on Computational Linguistics*. ACL.

Box, George EP, Stuart J Hunter, and William Gordon Hunter (2005). *Statistics for experimenters: design, innovation, and discovery*. Vol. 2. Wiley-Interscience New York.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder (2008). "Further meta-evaluation of machine translation." In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Proceedings of ACL 2008.

Canning, Yvonne, John Tait, Jackie Archibald, and Ros Crawley (2000). *Cohesive generation of syntactically simplified newspaper text*. Springer, pp. 47–63.

Carl, Michael, Pushpak Bhattacharya, and Kamal Kumar Choudhary, eds. (2012). *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*. Mumbai, India: The COLING 2012 Organizing Committee.

Carpenter, Patricia A and Marcel Adam Just (1983). "What your eyes do while your mind is reading." In: *Eye movements in reading: Perceptual and language processes*.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait (1999). "Simplifying text for language-impaired readers." In: *Proceedings of EACL 1999*. Vol. 99. Citeseer.

Caruana, Rich (1993). "Multitask learning: a knowledge-based source of inductive bias." In: *Proceedings of ICML 1993*.

Castilho, Sheila, Sharon O'Brien, Fabio Alves, and Morgan O'Brien (2014). "Does post-editing increase usability? A study with Brazilian Portuguese as Target Language." In: *Proceedings of EAMT 2014*.

Clarke, James and Mirella Lapata (2006). "Constraint-based sentence compression an integer programming approach." In: *Proceedings of COLING 2006*.

— (2008). "Global inference for sentence compression: An integer linear programming approach." In: *Journal of Artificial Intelligence Research*.

Cohn, Trevor and Mirella Lapata (2008). "Sentence compression beyond word deletion." In: *Proceedings of COLING 2008*.

— (2009). "Sentence compression as tree transduction." In: *Journal of Artificial Intelligence Research*.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). "Natural language processing (almost) from scratch." In: *The Journal of Machine Learning Research* 12.

Coster, William and David Kauchak (2011). "Simple English Wikipedia: a new text simplification task." In: *Proceedings of ACL 2011*. ACL.

Demberg, Vera and Frank Keller (2008). "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity." In: *Cognition* 109.

Demberg, Vera, Frank Keller, and Alexander Koller (2013). "Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar." In: *Computational Linguistics* 39.4.

Dewey, Melvil (1891). *Decimal Classification and Relative Index for Libraries, Clippings, Notes, Etc*. Library bureau.

Doherty, Stephen, Dorothy Kenny, and Andrew Way (2012). "A user-based usability assessment of raw machine translated technical instructions." In: *Proceedings of AMTA 2012*.

Doherty, Stephen and Sharon O'Brien (2014). "Assessing the Usability of Raw Machine Translated Output: A User-Centered Study Using Eye Tracking." In: *International Journal of Human-Computer Interaction* 30.1.

Doherty, Stephen, Sharon O'Brien, and Michael Carl (2010). "Eye tracking as an MT evaluation technique." In: *Machine translation* 24.1.

Doyon, Jennifer, Kathryn B Taylor, and John S White (1999). "Task-Based Evaluation for Machine Translation." In: *Proceedings of Machine Translation Summit VII*. Vol. 99.

Elming, Jakob, Anders Johannsen, Sigrid Klerke, Emanuele Lapponi, Héctor Martínez Alonso, and Anders Søgaard (2013). "Down-stream effects of tree-to-dependency conversions." In: *Proceedings of NAACL-HLT 2013*.

Falkenjack, Johan and Arne Jönsson (2014). "Classifying easy-to-read texts without parsing." In: *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*.

Feng, Lijun, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad (2010). "A comparison of features for automatic readability assessment." In: *Proceedings of the 23rd International Conference on Computational Linguistics*. ACL.

Filippova, Katja, Enrique Alfonseca, Carlos A Colmenares, Lukasz Kaiser, and Oriol Vinyals (2015). "Sentence Compression by Deletion with LSTMs." In: *Proceedings of EMNLP 2015*.

Filippova, Katja and Michael Strube (2008). "Dependency tree based sentence compression." In: *Proceedings of the Fifth International Natural Language Generation Conference*.

Flesch, Rudolph (1948). "A new readability yardstick." In: *Journal of applied psychology* 32.3.

François, Thomas and Eleni Miltsakaki (2012). "Do NLP and machine learning improve traditional readability formulas?" In: *Proceedings of the Workshop on Predicting and Improving Text Readability for target reader populations*. Montréal, Canada: ACL.

Frank, Stefan L and Rens Bod (2011). "Insensitivity of the human sentence-processing system to hierarchical structure." In: *Psychological science* 22.6.

Goldberg, Yoav (2015). "A primer on neural network models for natural language processing." In: *arXiv preprint arXiv:1510.00726*.

Han, Bo and Timothy Baldwin (2011). "Lexical normalisation of short text messages: Makn sens a #twitter." In: *Proceedings of ACL 2011*. ACL.

Heilman, Michael and Noah A Smith (2010). "Extracting simplified statements for factual question generation." In: *Proceedings of the Third Workshop on Question Generation*.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory." In: *Neural computation* 9.8.

Holmqvist, Kenneth, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Hovy, Dirk, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H Hovy (2013). "Learning whom to trust with mace." In: *Proceedings of NAACL-HLT 2013*.

Huang, Zhiheng, Wei Xu, and Kai Yu (2015). "Bidirectional LSTM-CRF models for sequence tagging." In: *arXiv preprint arXiv:1508.01991*.

Hyönä, Jukka and Alexander Pollatsek (2000). "Processing of Finnish compound words in reading." In: *Reading as a perceptual process*.

Keller, Frank (2010). "Cognitively plausible models of human language processing." In: *Proceedings of ACL 2010*. ACL.

Kennedy, Alan, Robin Hill, and Joël Pynte (2003). "The dundee corpus." In: *Proceedings of ECEM 2003*.

Klerke, Sigrid, Sheila Castilho, Maria Barrett, and Anders Søgaard (2015). "Reading metrics for estimating task efficiency with MT output." In: *Proceedings of Workshop on Cognitive Aspects of Computational Language Learning, EMNLP 2015*.

Klerke, Sigrid, Yoav Goldberg, and Anders Søgaard (2016). "Improving sentence compression by learning to predict gaze." In: *Proceedings of NAACL-HLT 2016*. San Diego, California: ACL.

Klerke, Sigrid and Anders Søgaard (2012). "DSim , a Danish Parallel Corpus for Text Simplification." In: *Proceedings of LREC 2012*.

— (2013). "Simple, readable sub-sentences." In: *Proceedings of ACL 2013*.

Knight, Kevin and Daniel Marcu (2000). "Statistics-based summarization-step one: Sentence compression." In: *Proceedings of AAAI/IAAI 2000*.

— (2002). "Summarization beyond sentence extraction: a probabilistic approach to sentence compression." In: *Artificial Intelligence* 139.

Krafka, Kyle, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba (2016). "Eye Tracking for Everyone." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kromann, Matthias T (2003). "The Danish Dependency Treebank and the DTAG treebank tool." In: *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*.

Lafferty, John, Andrew McCallum, and Fernando Pereira (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." In: 1.

LearningExpress (2005). *501 Challenging Logic and Reasoning Problems*. 501 Series. LearningExpress.

Levenshtein, Vladimir I (1966). "Binary codes capable of correcting deletions, insertions and reversals." In: *Soviet physics doklady*. Vol. 10.

Lin, Chin-Yew (2004). "Rouge: A package for automatic evaluation of summaries." In: *Proceedings of Workshop on Text summarization branches out, ACL 2004*. Barcelona, Spain.

Liu, Pengfei, Shafiq Joty, and Helen Meng (2015). "Fine-grained opinion mining with recurrent neural networks and word embeddings." In: *Proceedings of EMNLP 2015*.

Liversedge, Simon P, Denis Drieghe, Xin Li, Guoli Yan, Xuejun Bai, and Jukka Hyönä (2016). "Universality in eye movements and reading: A trilingual investigation." In: *Cognition* 147.

Louis, Annie (2012). "Automatic metrics for genre-specific text quality." In: *Proceedings of NAACL-HLT 2012*. ACL.

Malsburg, Titus von der and Shravan Vasishth (2013). "Scanpaths reveal syntactic underspecification and reanalysis strategies." In: *Language and Cognitive Processes* 28.10.

McDonald, Ryan (2006). "Discriminative Sentence Compression with Soft Syntactic Evidence." In: *Proceedings of EACL 2006*.

McDonald, Ryan and Joakim Nivre (2007). "Characterizing the Errors of Data-Driven Dependency Parsing Models." In: *Proceedings of EMNLP-CoNLL 2007*.

Mitchell, Jeff, Mirella Lapata, Vera Demberg, and Frank Keller (2010). "Syntactic and semantic factors in processing difficulty: An integrated measure." In: *Proceedings of ACL 2010*. ACL.

Nomoto, Tadashi (2007). "Discriminative sentence compression with conditional random fields." In: *Information Processing and Management: an International Journal* 43.6.

Okazaki, Naoaki (2007). *CRFsuite: a fast implementation of conditional random fields (CRFs)*. URL: `http://www.chokkan.org/software/crfsuite/`.

Olsen, Anneli (2012). "The Tobii I-VT Fixation Filter." In: *Tobii Technology*.

Ooms, Kristien, Lien Dupont, Lieselot Lapon, and Stanislav Popelka (2015). "Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental set-ups." In: *Journal of Eye Movement Research* 8.1.

Papineni, Kishore, Salim Roukus, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: a method for automatic evaluation of machine translation." In: *Proceedings of ACL 2002*. Philadelphia, Pennsylvania.

Petersen, Sarah E and Mari Ostendorf (2007). "Text simplification for language learners: a corpus analysis." In: *Proceedings of Workshop on Speech and Language Technology in Education*. Citeseer.

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2011). "A universal part-of-speech tagset." In: *Arxiv preprint ArXiv:1104.2086*.

Rauzy, Stéphane and Philippe Blache (2012). "Robustness and processing difficulty models. A pilot study for eye-tracking data on the French Treebank." In: *Proceedings of 24th International Conference on Computational Linguistics*.

Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research." In: *Psychological bulletin* 124.3.

— (2009). "Eye movements and attention in reading, scene perception, and visual search." In: *The quarterly journal of experimental psychology* 62.8.

Rayner, Keith, Kathryn H Chace, Timothy J Slattery, and Jane Ashby (2006). "Eye movements as reflections of comprehension processes in reading." In: *Scientific Studies of Reading* 10.3.

Rayner, Keith and Susan A Duffy (1986). "Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity." In: *Memory & Cognition* 14.3.

Rayner, Keith and Alexander Pollatsek (2013). "Basic processes in reading." In: *The Oxford Handbook of Cognitive Psychology*.

Rayner, Keith, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr (2012). *Psychology of reading*. Psychology Press.

Rello, Luz, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion (2013). "Frequent words improve readability and short words improve understandability for people with dyslexia." In: *Human-Computer Interaction–INTERACT 2013*.

Rello, Luz, Gaurang Kanvinde, and Ricardo Baeza-Yates (2012). "A Mobile Application for Displaying More Accessible eBooks for People with Dyslexia." In: *Procedia Computer Science* 14.

Rello, Luz, Martin Pielot, Mari-Carmen Marcos, and Roberto Carlini (2013). "Size matters (spacing not): 18 points for a dyslexic-friendly Wikipedia." In: *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. ACM.

Riezler, Stefan, Tracy H King, Richard Crouch, and Annie Zaenen (2003). "Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar." In: *Proceedings of NAACL-HLT 2003*.

Sackett, David L et al. (1979). "Bias in analytic research." In: *Journal of Chronic Diseases 1979* 32.

Schotter, Elizabeth R, Klinton Bicknell, Ian Howard, Roger Levy, and Keith Rayner (2014). "Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading." In: *Cognition* 131.1.

Schwarm, Sarah E and Mari Ostendorf (2005). "Reading Level Assessment Using Support Vector Machines and Statistical Language Models." In: *Proceedings of ACL 2005*.

Siddharthan, Advaith (2006). "Syntactic simplification and text cohesion." In: *Research on Language & Computation* 4.1.

— (2014). "A survey of research on text simplification." In: *ITL-International Journal of Applied Linguistics* 165.2.

Siddharthan, Advaith and Napoleon Katsos (2012). "Offline Sentence Processing Measures for testing Readability with Users." In: *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Montréal, Canada: ACL.

Slattery, Timothy J, Patrick Sturt, Kiel Christianson, Masaya Yoshida, and Fernanda Ferreira (2013). "Lingering misinterpretations of garden path sentences arise from competing syntactic representations." In: *Journal of Memory and Language* 69.2.

Specia, Lucia (2011). "Exploiting objective annotations for measuring translation post-editing effort." In: *Proceedings of EAMT 2011*.

Specia, Lucia, Sujay Kumar Jauhar, and Rada Mihalcea (2012). "Semeval-2012 task 1: English lexical simplification." In: *Proceedings of SemEval 2012*. ACL.

Spivey-Knowlton, Michael J, John C Trueswell, and Michael K Tanenhaus (1993). "Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses." In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 47.2.

Stolcke, Andreas (2002). "SRILM – an extensible language modeling toolkit." In: *Proceedings of the Seventh International Conference on Spoken Language Processing*.

Studio, Tobii (2008). "1.2 User Manual." In: *Tobii Technology AB*.

Stymne, Sara, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester (2012). "Eye Tracking as a Tool for Machine Translation Error Analysis." In: *Proceedings of LREC 2012*.

Stymne, Sara, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre (2013). "Statistical machine translation with readability constraints." In: *Proceedings of NODALIDA 2013*.

Tsarfaty, Reut, Djamé Seddah, Sandra Kübler, and Joakim Nivre (2013). "Parsing morphologically rich languages: Introduction to the special issue." In: *Computational Linguistics* 39.1.

Vajjala, Sowmya and Detmar Meurers (2013). "On The Applicability of Readability Models to Web Texts." In: *Proceedings of ACL 2013*.

— (2014). "Exploring Measures of Readability for Spoken Language: Analyzing linguistic features of subtitles to identify age-specific TV programs." In: *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*.

Von der Malsburg, Titus and Shravan Vasishth (2011). "What is the scanpath signature of syntactic reanalysis?" In: *Journal of Memory and Language* 65.2.

Woodsend, Kristian and Mirella Lapata (2011a). "Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming." In: *Proceedings of EMNLP 2011*.

— (2011b). "WikiSimple: Automatic Simplification of Wikipedia Articles." In: *Proceedings of AAAI 2011*.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles (2015). "Problems in current text simplification research: New data can help." In: *Transactions of the ACL* 3.

Zhou, Jie and Wei Xu (2015). "End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks." In: *Proceedings of ACL 2015*.

Zhu, Zhemin, Delphine Bernhard, and I. Gurevych (2010). "A monolingual tree-based translation model for sentence simplification." In: *Proceedings of The 23rd International Conference on Computational Linguistics*. ACL.