Simple, readable sub-sentences

Sigrid Klerke
Centre for Language Technology
University of Copenhagen
sigridklerke@gmail.com

Anders Søgaard
Centre for Language Technology
University of Copenhagen
soegaard@hum.ku.dk

Abstract

We present experiments using a new unsupervised approach to automatic text simplification, which builds on sampling and ranking via a loss function informed by readability research. The main idea is that a loss function can distinguish good simplification candidates among randomly sampled sub-sentences of the input sentence. Our approach is rated as equally grammatical and beginner reader appropriate as a supervised SMT-based baseline system by native speakers, but our setup performs more radical changes that better resembles the variation observed in human generated simplifications.

1 Introduction

As a field of research in NLP, text simplification (TS) has gained increasing attention recently, primarily for English text, but also for Brazilian Portuguese (Specia, 2010; Aluísio et al., 2008), Dutch (Daelemans et al., 2004), Spanish (Drndarevic and Saggion, 2012), Danish (Klerke and Søgaard, 2012), French (Seretan, 2012) and Swedish (Rybing and Smith, 2009; Decker, 2003). Our experiments use Danish text which is similar to English in that it has a deep orthography making it hard to map between letters and sounds. Danish has a relatively free word order and sparse morfology.

TS can help readers with below average reading skills access information and may supply relevant training material, which is crucial for developing reading skills. However, manual TS is as expensive as translation, which is a key limiting factor on the availability of easy-to-read material. One of the persistent chalenges of TS is that different interventions are called for depending on the target reader population. Automatic TS is an effective way to counter these limitations.

2 Approach

Definitions of TS typically reflect varying target reader populations and the methods studied. For our purposes we define TS to include any operation on the linguistic structure and content of a text, intended to produce new text, which

1. has semantic content similar to (a part of) the original text
2. requires less cognitive effort to decode and understand by a target reader, compared to the original text.

Operations on linguistic content may include deletion, reordering and insertion of content, paraphrasing concepts, resolving references, etc., while typography and layout are excluded as non-linguistic properties.

We cast the problem of generating a more readable sentence from an input as a problem of choosing a reasonable sub-sentence from the words present in the original. The corpus-example below illustrates how a simplified sentence can be embedded as scattered parts of a non-simplified sentence. The words in bold are the common parts which make up almost the entire human generated simplification and constitutes a suitable simplification on its own.

Original: **Der er målt** hvad der bliver betegnet som abnormt **stor**e **mængde**r af **radioaktivt materiale i hav**vand **nær** det jordskælvsramte **atomkraftværk i Japan .**

What has been termed **a**n abnormally **large amount of radioactivity has been measured in sea** water **near the nuclear power plant** that was hit by earthquakes **in Japan**

Simplified: **Der er målt** en **stor mængde radioaktivt materiale i hav**et **nær atom-kraftværk**et Fukushima **i Japan .**

A large amount of radioactivity has been measured in the **sea near the nuclear power plant** Fukushima **in Japan**

To generate candidate sub-sentences we use a random deletion procedure in combination with

Proceedings of the ACL Student Research Workshop, pages 142–149,
Sofia, Bulgaria, August 4-9 2013.

general dependency-based heuristics for conserving main sentence constituents, and then introduce a loss-function for choosing between candidates. Since we avoid relying on a specialized parallel corpus or a simplification grammar, which can be expensive to create, the method is especially relevant for under-resourced languages and organizations. Although we limit rewriting to deletions, the space of possible candidates grows exponentially with the length of the input sentence, prohibiting exhaustive candidate generation, which is why we chose to sample the deletions randomly. However, to increase the chance of sampling *good* candidates, we restrict the search space under the assumption that some general patterns apply, namely, that the main verb and subject should always be kept, negations should be kept and that if something is kept that originally had objects, those objects should also be kept. Another way in which we restrict the candidate space is by splitting long sentences. Some clauses are simple to identify and extract, like relative clauses, and doing so can dramatically reduce sentence length. Both simple deletions and extraction of clauses can be observed in professionally simplified text. (Medero, 2011; Klerke, 2012)

The next section positions this research in the context of related work. Section 4 presents the experimental setup including generation and evaluation. In Section 5, the results are presented and discussed and, finally, concluding remarks and future perspectives are presented in the last section.

3 Related work

Approaches for automatic TS traditionally focus on lexical substitution (De Belder and Moens, 2012; Specia et al., 2012; Yatskar et al., 2010), on identifying re-write rules at sentence level either manually (Chandrasekar et al., 1996; Carroll et al., 1999; Canning et al., 2000; Siddharthan, 2010; Siddharthan, 2011; Seretan, 2012) or automatically from parallel corpora (Woodsend and Lapata, 2011; Coster and Kauchak, 2011; Zhu et al., 2010) and possibly learning cues for when to apply such changes (Petersen and Ostendorf, 2007; Medero, 2011; Bott et al., 2012).

Chandrasekar et al. (1996) propose a structural approach, which uses syntactic cues to recover relative clauses and appositives. Sentence level syntactic re-writing has since seen a variety of manually constructed general sentence splitting rules, designed to operate both on dependencies and phrase structure trees, and typically including lexical cues (Siddharthan, 2011; Heilman and Smith, 2010; Canning et al., 2000). Similar rules have been created from direct inspection of simplification corpora (Decker, 2003; Seretan, 2012) and discovered automatically from large scale aligned corpora (Woodsend and Lapata, 2011; Zhu et al., 2010).

In our experiment we apply few basic sentence splitting rules as a pre-processing technique before using an over-generating random deletion approach.

Carroll et al. (1999) perform lexical substitution from frequency counts and eliminate anaphora by resolving and replacing the referring expressions with the entity referred to. Their system further include compound sentence splitting and rewriting of passive sentences to active ones (Canning et al., 2000). Research into lexical simplification remains an active topic. De Belder and Moens (2012; Specia et al. (2012) are both recent publications of new resources for evaluating lexical simplification in English consisting of lists of synonyms ranked by human judges. Another type of resource is graded word-lists as described in Brooke et al. (2012). Annotator agreement and comparisons so far shows that it is easy to overfit to reflect individual annotator and domain differences that are not of relevance to generalized systems.

In a minimally supervised setup, our TS approach can be modified to include lexical simplifications as part of the random generation process. This would require a broad coverage list of words and simpler synonyms, which could for instance be extracted from a parallel corpus like the DSim corpus.

For the majority of research in automatic TS the question of what constitutes cognitive load is not discussed. An exception is Siddharthan and Katsos (2012), who seek to isolate the psycholinguistically motivated notions of sentence comprehension from sentence acceptability by actually measuring the effect of TS on cognition on a small scale.

Readability research is a line of research that is more directly concerned with the nature of cognitive load in reading building on insights from psycholinguistics. One goal is to develop techniques and metrics for assessing the readability of unseen

text. Such metrics are used as a tool for teachers and publishers, but existing standard metrics (like Flesch-Kincaid (Flesch, 1948) and LIX (Bjornsson, 1983)) were designed and optimized for easy manual application to human written text, requiring thehuman reader to assess that the text is congruent and coherent. More recent methods promise to be applicable to unassessed text. Language modeling in particular has shown to be a robust and informative component of systems for assessing text readability (Schwarm and Ostendorf, 2005; Vajjala and Meurers, 2012) as it is better suited to evaluate grammaticality than standard metrics. We use language modeling alongside traditional metrics for selecting good simplification candidates.

4 Experiments

4.1 Baseline Systems

We used the original input text and the human simplified text from the sentence aligned DSim corpus which consist of 48k original and manually simplified sentences of Danish news wire text (Klerke and Søgaard, 2012) as reference in the evaluations. In addition we trained a statistical machine translation (SMT) simplification system, in effect translating from normal news wire text to simplified news. To train an SMT system, a large resource of aligned parallel text and a language model of the target language are needed. We combined the 25 million words Danish Korpus 2000[1] with the entire 1.75 million words unaligned DSim corpus (Klerke and Søgaard, 2012) to build the language model[2]. Including both corpora gives better coverage and assigns lower average ppl and a simlar difference in average ppl between the two sides of a held out part of the DSim corpus compared to using only the simplified part of DSim for the language model. Following Coster and Kauchak (2011), we used the phrase-based SMT Moses (Koehn et al., 2007), with GIZA++ word-alignment (Och and Ney, 2000) and phrase tables learned from the sentence aligned portion of the DSim corpus.

[1] `http://korpus.dsl.dk/korpus2000/engelsk_hovedside`

[2] The LM was a 5-gram Knesser-Ney smoothed lowercase model, built using IRSTLM (Federico et al., 2008)

4.2 Experimental setup

Three system variants were set up to generate simplified output from the original news wire of the development and test partitions of the DSim corpus. The texts were dependency-parsed using Bohnet's parser (Bohnet, 2010) trained on the Danish Treebank[3] (Kromann, 2003) with default settings[4].

1. **Split** only performed simple sentence splitting.

2. **Sample** over-generated candidates by sampling the heuristically restricted space of random lexical deletions and ranking candidates with a loss function.

3. **Combined** is a combination of the two, applying the sampling procedure of Sample to the split sentences from Split.

Sentence Splitting We implemented sentence splitting to extract relative clauses, as marked by the dependency relation `rel`, coordinated clauses, `coord`, and conjuncts, `conj`, when at least a verb and a noun is left in each part of the split. Only splits resulting in sentences of more than three words were considered. Where applicable, referred entities were included in the extracted sentence by using the dependency analysis to extract the subtree of the former head of the new sentence[5]. In case of more than one possibility, the split resulting in the most balanced division of the sentence was chosen and the rules were re-applied if a new sentence was still longer than ten tokens.

Structural Heuristics To preserve nodes from later deletion we applied heuristics using simple structural cues from the dependency structures. We favored nodes headed by a subject relation, `subj`, and object relations, `*obj`, and negating modifiers (the Danish word *ikke*) under the assumption that these were most likely to be important for preserving semantics and generating well-formed candidates under the sampling procedure described below. The heuristics were applied both to trees, acting by preserving entire subtrees and applied to words, only preserving single tokens.

[3] `http://ilk.uvt.nl/conll/post_task_data.html`

[4] Performance of the parser on the treebank test set Labeled attatchment score (LAS) = 85.65 and Unlabeled attatchment score (UAS) = 90.29

[5] For a formal description see (Klerke, 2012)

This serves as a way of avoiding relying heavily on possibly faulty dependency analyses and also avoid the risk of insisting on keeping long, complex or superfluous modifiers.

Sampling Candidates for scoring were overgenerated by randomly selecting parts of a (possibly split) input sentence. Either the selected nodes with their full sub-tree or the single tokens from the flat list of tokens were eliminated, unless they were previously selected for preservation by a heuristic. Some additional interaction between heuristics and sampling happened when the deletions were performed on trees: deletion of subtrees allow non-continuous deletions when the parses are non-projective, and nodes that were otherwise selected for keeping may nevertheless be removed if they are part of a subtree of a node selected for deletion. After pruning, all nodes that used to have outgoing `obj`-relations had the first child node of these relations restored.

4.3 Scoring

We rank candidates according to a loss function incorporating both readability score (the lower, the more readable) and language model perplexity (the lower, the less perplexing) as described below. The loss function assigns values to the candidates such that the best simplification candidate receives the lowest score.

The loss function is a weighted combination of three scores: perplexity (PPL), LIX and word-class distribution (WCD). The PPL scores were obtained from a 5-gram language model of Danish[6] We used the standard readability metric for Danish, LIX (Bjornsson, 1983)[7]. Finally, the WCD measured the variation in universal pos-tag-distribution [8] compared to the observed tag-variation in the entire simplified corpus. For PPL and LIX we calculated the difference between the score of the input sentence and the candidate.

Development data was used for tuning the weights of the loss function. Because the candidate-generation is free to produce extremely short candidates, we have to deal with candidates receiving extremely low scores. Those scores never arise in the professionally simplified text, so we eliminate extreme candidates by introducing filters on all scores. The lower limit was tuned experimentally and fixed approximately two times below the average difference observed between the two parts of the aligned DSim corpus, thus limiting the reduction in PPL and LIX to 60% of the input's PPL and LIX. The upper limit was fixed at the input-level plus 20% to allow more varied candidates through the filters. The WCD-filter accepted all candidates with a tag-variance that fell below the 75-percentile observed variance in the simplified training part of the DSim corpus. The resulting loss was calculated as the sum of three weighted scores.

Below is the loss function we minimized over the filtered candidates $\mathbf{t} \in \mathcal{T}_\mathbf{s}$ for each input sentence, $\mathbf{s}$. The notation *var*() denotes the range allowed through a hard filter. Using development data we set the values of the term weights to $\alpha = 1, \beta = 6$ and $\gamma = 2$.

$$\begin{aligned} \mathbf{t}^* &= \operatorname*{argmin}_{\mathbf{t}\in\mathcal{T}_\mathbf{s}} loss(\mathbf{s},\mathbf{t}) \\ loss(\mathbf{s},\mathbf{t}) &= \alpha\frac{\Delta LIX(\mathbf{s},\mathbf{t})}{\mathrm{var}(LIX(\mathbf{s}))} + \beta\frac{\Delta PPL(\mathbf{s},\mathbf{t})}{\mathrm{var}(PPL(\mathbf{s}))} \\ &\quad + \gamma\frac{\Delta WCD(.75,\mathbf{t})}{WCD(.75)} \end{aligned}$$

If no candidates passed through the filters, the input sentence was kept.

4.4 Evaluation

Evaluation was performed by a group of proficient Danish speaking volunteers who received written instructions and responded anonymously via an online form. 240 sentences were evaluated: six versions of each of 40 test set sentences. 48 sentences were evaluated by four judges, and the remaining by one judge each. The judges were asked to rate each sentence in terms of grammaticality and in terms of perceived beginner reader appropriateness, both on a 5-point scale, with one signifying *very good* and five signifying *very bad*. The evaluators had to rate six versions of each sentence: original news wire, a human simplified version, the baseline system, a split sentence version (Split), a sampled only version (Sample), and a version combining the Split and Sample techniques (Combined). The presentation was randomized. Below are example outputs

[6]The LM was Knesser-Ney smoothed, using the same corpora as the baseline system, without punctuation and built using SRILM (Stolcke, 2002).

[7]LIX is similar to the English Flesch-Kincaid grade level in favoring short sentences with short words. The formula is LIX $=$ *average sentence length* $+$ % *long words*, with long words being of more than 6 characters. (Anderson, 1983) calculated a conversion from LIX to grade levels.

[8]suggested by(Petrov et al., 2011)

for the baseline and the other three automatic systems:

BL: **Der er hvad der bliver betegnet som abnormt store mængder radioaktivt materiale i havvand nær** frygter **atomkraftværk .**

Split: **Der er målt hvad**. Hvad **bliver betegnet som abnormt store mængder af radioaktivt materiale i havvand nær det jordskælvsramte atomkraftværk i Japan .**

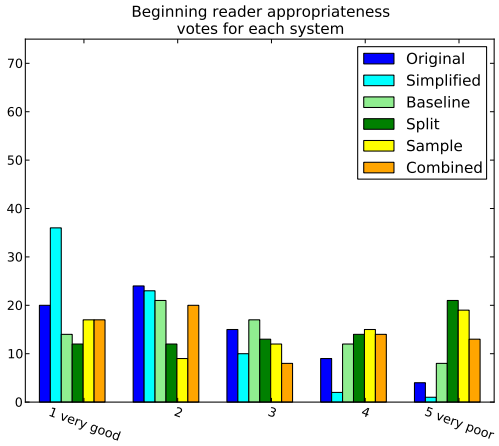
Sample: **Der er målt hvad der bliver betegnet som store mængder af radioaktivt materiale i havvand japan .**

Comb.: **Der er målt hvad**. Hvad **bliver betegnet som store mængder af radioaktivt materiale det atomkraftværk i japan .**

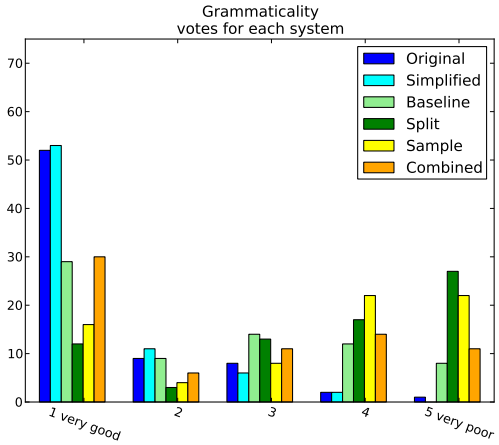
5 Results

The ranking of the systems in terms of beginner reader appropriateness and grammaticality, are shown in Figure 1. From the test set of the DSim corpus, 15 news wire texts were arbitrarily selected for evaluation. For these texts we calculated median LIX and PPL. The results are shown in Table 1. The sentences for human evaluation were drawn arbitrarily from this collection. As expected, the filtering of candidates and the loss function force the systems Sample and Combined to choose simplifications with LIX and PPL scores close to the ones observed in the human simplified version. Split sentences only reduce LIX as a result of shorter sentences, however PPL is the highest, indicating a loss of grammaticality. Most often this was caused by tagger and parser errors. The baseline reduces PPL slightly, while LIX is unchanged. This reflects the importance of the language model in the SMT system.

In the analyses below, the rating were collapsed to three levels. For texts ranked by more than one judge, we calculated agreement as Krippendorff's α. The results are shown in Table 2. In addition to sentence-wise agreement, the system-wise evaluation agreement was calculated as all judges were evaluating the same 6 systems 8 times each. We calculated α of the most frequent score (mode) assigned by each judge to each system. As shown in Table 2 this system score agreement was only about half of the single sentence agreement, which reflect a notable instability in output quality of all computer generated systems. The same tendency is visible in both histograms in Figure 1a and 1b. While grammaticality is mostly agreed upon when the scores are collapsed into three bins ($\alpha = 0.650$), proficient speakers do not agree to the same extent on what constitutes beginner reader appropriate text ($\alpha = 0.338$). The average, mean and most frequent assigned ranks are recorded in Table 3. Significant differences at $p < 0.05$ are reported in Table 4.



(a) Sentence – Beginner



(b) Sentence – Grammar.

Figure 1: Distribution of all rankings on systems before collapsing rankings.

	Orig.	Simpl.	Base	Split	Sample	Comb.
PPL	222	174	214	234	**164**	177
LIX	45 (10)	39 (8)	45 (10)	41(9)	36 (8)	**32 (7)**

Table 1: LIX and PPL scores for reference texts and system generated output. Medians are reported, because distributions are very skewed, which makes the mean a bad estimator of central tendency. LIX grade levels in parenthesis.

Reflecting the fair agreement on grammaticality, all comparisons come out significant except the human generated versions that are judged as equally grammatical and the Combined and Baseline systems that are indistinguishable in grammaticality. Beginner reader appropriateness is significantly better in the human simplified version

	Systems	Sentences
Beginner reader	0.168	0.338
Grammaticality	0.354	0.650

Table 2: Krippendorff's α agreement for full-text and sentence evaluation. Agreement on system ranks was calculated from the most frequent score per judge per system.

compared to all other versions, and the original version is significantly better than the Sample and Split systems. The remaining observed differences are not significant due to the great variation in quality as expressed in Figure 1a.

We found that our Combined system produced sentences that were as grammatical as the baseline and also frequently judged to be appropriate for beginner readers. The main source of error affecting both Combined and Split is faulty sentence splitting as a result of errors in tagging and parsing. One way to avoid this in future development is to propagate several split variants to the final sampling and scoring. In addition, the systems Combined and Sample are prone to omitting important information that is perceived as missing when compared directly to the original, although those two systems are the ones that score the closest to the human generated simplifications. As can be expected in a system operating exclusively at sentence level, coherence across sentence boundaries remains a weak point.

Another important point is that while the baseline system performs well in the evaluation, this is likely due to its conservativeness: choosing simplifications resembling the original input very closely. This is evident both in our automatic measures (see Table 1) and from manual inspection. Our systems Sample and Combine, on the other hand, have been tuned to perform much more radical changes and in this respect more closely model the changes we see in the human simplification. Combined is thus evaluated to be at level with the baseline in grammaticality and beginner reader appropriateness, despite the fact that the baseline system is supervised.

Conclusion and perspectives

We have shown promising results for simplification of Danish sentences. We have also shown that using restricted over-generation and scoring can be a feasible way for simplifying text without relying directly on large scale parallel corpora,

	Sent. – Beginner			*Sent. – Grammar*		
	$\bar{x}$	$\tilde{x}$	mode	$\bar{x}$	$\tilde{x}$	mode
Human Simp.	1.44	1	1	1.29	1	1
Orig.	2.14	1	1	1.32	1	1
Base	2.58	3	1	1.88	2	1
Split	3.31	3	5	2.44	3	3
Sample	3.22	3	5	2.39	3	3
Comb.	2.72	1	1	1.93	2	1

Table 3: Human evaluation. Mean ($\bar{x}$), median ($\tilde{x}$) and most frequent (mode) of assigned ranks by beginner reader appropriateness and grammaticality as assessed by proficient Danish speakers.

	Comb.	Sample	Split	Base	Orig.
Human Simp.	b, g	b, g	b, g	b, g	b
Orig.	g	b, g	b, g	g	
Base		g	g		
Split	g				
Sample	g				

Table 4: Significant differences between systems in experiment b: Beginner reader appropriateness and g: Grammaticality. Bonferroni-corrected Mann-Whitney's U for 15 comparisons, two-tailed test. A letter indicate significant difference at corrected $p < 0.05$ level.

which for many languages do not exist. To integrate language modeling and readability metrics in scoring is a first step towards applying results from readability research to the simplification framework. Our error analysis showed that many errors come from pre-processing and thus more robust NLP-tools for Danish are needed. Future perspectives include combining supervised and unsupervised methods to exploit the radical unsupervised deletion approach and the knowledge obtainable from observable structural changes and potential lexical simplifications. We plan to focus on refining the reliability of sentence splitting in the presence of parser errors as well as on developing a loss function that incorporates more of the insights from readability research, and to apply machine learning techniques to the weighting of features. Specifically we would like to investigate the usefulness of discourse features and transition probabilities (Pitler and Nenkova, 2008) for performing and evaluating full-text simplifications.

Acknowledgements

Thanks to Mirella Lapata and Kristian Woodsend for their feedback and comments early in the process of this work and to the Emnlp@Cph group and reviewers for their helpful comments.

References

S.M. Aluísio, Lucia Specia, T.A.S. Pardo, E.G. Maziero, H.M. Caseli, and R.P.M. Fortes. 2008. A corpus analysis of simple account texts and the proposal of simplification strategies: first steps towards text simplification systems. In *Proceedings of the 26th annual ACM international conference on Design of communication*, pages 15–22. ACM.

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

C. H. Bjornsson. 1983. Readability of Newspapers in 11 Languages. *Reading Research Quarterly*, 18(4):480–497.

B Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97. Association for Computational Linguistics.

S. Bott, H. Saggion, and D. Figueroa. 2012. A hybrid system for spanish text simplification. In *Third Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Montreal, Canada*.

Julian Brooke, Vivian Tsang, David Jacob, Fraser Shein, and Graeme Hirst. 2012. Building Readability Lexicons with Unannotated Corpora. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 33–39, Montr{é}al, Canada, June. Association for Computational Linguistics.

Y. Canning, J. Tait, J. Archibald, and R. Crawley. 2000. *Cohesive generation of syntactically simplified newspaper text*. Springer.

John Carroll, G. Minnen, D. Pearce, Yvonne Canning, S. Devlin, and J. Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of EACL*, volume 99, pages 269–270. Citeseer.

R. Chandrasekar, Christine Doran, and B Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 1041–1044. Association for Computational Linguistics.

William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, volume 2, pages 665–669. Association for Computational Linguistics.

W. Daelemans, A. Höthker, and E.T.K. Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

A. Davison and R.N. Kantor. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, pages 187–209.

J. De Belder and M.F. Moens. 2012. A dataset for the evaluation of lexical simplification. *Computational Linguistics and Intelligent Text Processing*, pages 426–437.

Anna Decker. 2003. Towards automatic grammatical simplification of Swedish text. Master's thesis, Stockholm University.

Biljana Drndarevic and Horacio Saggion. 2012. Towards Automatic Lexical Simplification in Spanish: An Empirical Study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 8–16, Montr{é}al, Canada, June. Association for Computational Linguistics.

M Federico, N Bertoldi, and M Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Ninth Annual Conference of the International Speech Communication Association*.

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

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

Sigrid Klerke and Anders Søgaard. 2012. DSim , a Danish Parallel Corpus for Text Simplification. In *Proceedings of Language Resources and Evaluation (LREC 2012)*, pages 4015–4018.

Sigrid Klerke. 2012. Automatic text simplification in danish. sampling a restricted space of rewrites to optimize readability using lexical substitutions and dependency analyses. Master's thesis, University of Copenhagen.

P Koehn, H Hoang, A Birch, C Callison-Burch, M Federico, N Bertoldi, B Cowan, W Shen, C Moran, R Zens, and Others. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

M T Kromann. 2003. The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT)*, page 217.

Julie Medero. 2011. Identifying Targets for Syntactic Simplification. In *Proceedings of Speech and Language Technology in Education*.

F.J. Och and H. Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 1086–1090. Association for Computational Linguistics.

S.E. E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *the Proceedings of the Speech and Language Technology for Education Workshop*, pages 69–72. Citeseer.

S. Petrov, D. Das, and R. McDonald. 2011. A universal part-of-speech tagset. *Arxiv preprint ArXiv:1104.2086*.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Jonas Rybing and Christian Smith. 2009. CogFLUX Grunden till ett automatiskt textförenklingssystem för svenska. Master's thesis, Linköpings Universitet.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 523–530.

V. Seretan. 2012. Acquisition of syntactic simplification rules for french. In *Proceedings of Language Resources and Evaluation (LREC 2012)*.

Advaith Siddharthan 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*, pages 17–24, Montr{é}al, Canada, June. Association for Computational Linguistics.

Advaith Siddharthan. 2010. Complex lexico-syntactic reformulation of sentences using typed dependency representations. *Proceedings of the 6th International Natural Language Generation Conference*.

Advaith Siddharthan. 2011. Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11.

L. Specia, S.K. Jauhar, and R. Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355.

L. Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, pages 30–39.

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

S. Vajjala and D. Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, pages 163–173.

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

Mark Yatskar, Bo Pang, C. Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368. Association for Computational Linguistics.

Zhemin Zhu, 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*, pages 1353–1361. Association for Computational Linguistics.