1. **A framework to develop a holistic model of text readability**: We have seen that ARA research is primarily focused on textual features, especially those that focus on form. However, there are many other aspects such as conceptual difficulty, typographic features, user characteristics, task features etc, as we saw earlier. An obvious challenge would be to develop a unified model of ARA that encompasses all these aspects. However, it is not the work of one person or group, nor can it all be done in one go. So, an important first step in this direction (which can address limitations 1–2) would be to design an easily extendable framework to build a holistic model of readability by incrementally adding multiple dimensions, covering multi modal data. This would also necessitate the development of appropriate corpora and other resources suitable for this purpose.
2. **Models adaptable to new domain**: Any ARA model could still only be relevant to the target domain/audience and may not directly transfer to a new application scenario. Hence, approaches that can transfer an existing model into a new domain/audience should be developed. One potential avenue to explore in this direction is to model ARA as a ranking problem instead of classification or regression, as it was shown to generalize better than other models in the past (Xia et al., 2016). This can address the limitation 3 mentioned earlier.
3. **Creation of open and diverse datasets and tools:** Development of openly accessible corpora which suit various application scenarios, for several languages is a major challenge in ARA research, as we saw earlier. New methods to quickly create (and validate) corpora need to be developed. Whether recent developments in data augmentation can be useful for developing ARA corpora is also something that can be explored in future. For widespread adaptation of research on ARA, and to progress towards a holistic model, ready to use tools should be developed. Tools such as Coh-Metrix (Graesser et al., 2011) and CTAP[3] (Chen and Meurers, 2016) that provide a range of linguistic features typically associated with readability assessment are a step in this direction. Along with these, tools that can show the predictions of ARA models should also be developed, to address the limitations 3–4.
4. **Developing Best Practices:** To support the creation of reusable resources (corpora/code) and to be able to reproduce/replicate results and understand SOTA, a set of best practices must be developed for ARA. Some inspiration for this can be drawn from the procedures and findings of the recently conducted REPROLANG challenge (Branco et al., 2020) which conducted a shared task to replicate some published NLP research. The best practices for ARA should also include guidelines for validating the corpora and features developed, as well as recommended procedures for developing interpretable approaches. This can help one address the limitations 5–7 to some extent. This will also potentially encourage non-NLP researchers to seriously consider employing more recent ARA models in their research. Some aspects of this challenge area (e.g., validation, interpretation) demand expertise beyond NLP methods and may require inter-disciplinary collaborations.

It has to be noted that some of these challenges are not necessarily specific to ARA, and are applicable across NLP in general. This collection of ideas on challenges for future is by no means exhaustive, and we hope this survey initiates more discussion in this direction.

6 Conclusion

In this paper, we presented an overview of two decades of research on automatic readability assessment in NLP and related areas of research. During this process we identified the limitations of contemporary research and identified some challenge areas for future. Despite a large body of research, we don't yet have a clear picture of what works for ARA, and there are no off the shelf tools and resources for different kinds of researchers and practitioners interested in ARA. Further, many challenges mentioned in previous surveys still remain. Considering that readability assessment has a wide range of applications in and outside NLP as it was seen from examples in Section 1, we think it is important to address these issues and enable the a broader adaption of ARA approaches within and outside NLP.

[3] `www.ctapweb.com`

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, (Just Accepted):1–87.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2012. Towards fine-grained readability measures for self-directed language learning. In *Proceedings of the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012*, number 080, pages 11–19. Linköping University Electronic Press.

Karin Berendes, Sowmya Vajjala, Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein. 2018. Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, 110(4):518.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan van Noord, Dieter Van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with reprolang2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545.

Miriam Cha, Youngjune Gwon, and HT Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 113–119, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Xiaobin Chen and Detmar Meurers. 2018. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Kevyn Collins-Thompson and Jamie Callan. 2004. Information retrieval for language tutoring: An overview of the reap project. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 544–545.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Scott A Crossley, Hae Sung Yang, and Danielle S McNamara. 2014. What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113.

James W Cunningham and Heidi Anne Mesmer. 2014. Quantitative measurement of text difficulty: What's the use? *The Elementary School Journal*, 115(2):255–269.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? finding the optimal feature set for general readability prediction and its application to english and dutch. *Computational Linguistics*, 42(3):457–490.

Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83. Association for Computational Linguistics.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

William H DuBay. 2007. *Unlocking language: The classic readability studies*. Impact Information.

Carsten Eickhoff, Pavel Serdyukov, and Arjen P De Vries. 2011. A combined topical/non-topical approach to identifying web sites for children. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 505–514.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284.

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

Thomas François and Cédrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea, July. Association for Computational Linguistics.

Thomas François, Núria Gala, Patrick Watrin, and Cédrick Fairon. 2014. Flelex: a graded lexical resource for french foreign learners. In *International conference on Language Resources and Evaluation (LREC 2014)*.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. Svalex: a cefr-graded lexical resource for swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 213–219.

Thomas François. 2014. An analysis of a french as a foreign language corpus for readability assessment. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 13–32.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 20(2):79–97.

Nuria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a french lexicon with difficulty measures: Nlp helping to bridge the gap between traditional dictionaries and specialized lexicons. In *Electronic lexicography in the 21st century: thinking outside the paper (eLEX-2013)*.

Susan R Goldman and Carol D Lee. 2014. Text complexity: State of the art and the conundrums it raises. *the elementary school journal*, 115(2):290–300.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. Simple or complex? assessing the readability of basque texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 334–344.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79.

Elfrieda H Hiebert and P David Pearson. 2014. Understanding text complexity: Introduction to the special issue. *the elementary school journal*, 115(2):153–160.

David M Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217.

Zahurul Islam, Alexander Mehler, and Rashedur Rahman. 2012. Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 545–553.

Shoaib Jameel, Wai Lam, and Xiaojun Qian. 2012. Ranking text documents based on conceptual difficulty using term embedding and sequential discourse cohesion. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 145–152. IEEE.

Zhiwei Jiang, Qing Gu, Yafeng Yin, and Daoxu Chen. 2018. Enriching word embeddings with domain knowledge for readability assessment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 366–378.

Zhiwei Jiang, Qing Gu, Yafeng Yin, Jianxiang Wang, and Daoxu Chen. 2019. Graw+: A two-view graph propagation method with word coupling for readability assessment. *Journal of the Association for Information Science and Technology*, 70(5):433–447.

Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd international conference on computational linguistics*, pages 546–554. Association for Computational Linguistics.

Paul Kidwell, Guy Lebanon, and Kevyn Collins-Thompson. 2011. Statistical estimation of word acquisition with application to readability prediction. *Journal of the American Statistical Association*, 106(493):21–30.

Jin Young Kim, Kevyn Collins-Thompson, Paul N Bennett, and Susan T Dumais. 2012. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 213–222.

Rebecca Knowles, Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2016. Analyzing learner understanding of novel l2 vocabulary. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 126–135.

Christoph Kühberger, Christoph Bramann, Zarah Weiß, and Detmar Meurers. 2019. Task complexity in history textbooks: A multidisciplinary case study on triangulation in history education research. *History Education Research Journal*, 16(1):139–157.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Bertha A Lively and Sidney L Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(389-398):73.

Pavel Lyatoshinsky, Manolis Pratsinis, Dominik Abt, Hans-Peter Schmid, Valentin Zumstein, and Patrick Betschart. 2019. Readability assessment of commonly used german urological questionnaires. *Current urology*, 13(2):87–93.

Yi Ma, Eric Fosler-Lussier, and Robert Lofthus. 2012. Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 548–552.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 193–203.

Matej Martinc, Senja Pollak, and Marko Robnik Šikonja. 2019. Supervised and unsupervised neural approaches to text readability. *arXiv preprint arXiv:1907.11779.*

G Harry McLaughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. 2020. Readnet: A hierarchical transformer framework for web article readability analysis. In *European Conference on Information Retrieval*, pages 33–49. Springer.

Hamid Mohammadi and Seyed Hossein Khasteh. 2019. Text as environment: A deep reinforcement learning text readability assessment model. *arXiv preprint arXiv:1912.05957.*

Courtney Napoles and Mark Dredze. 2010. Learning simple wikipedia: A cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 42–50. Association for Computational Linguistics.

Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC.*

Maria Soledad Pera and Yiu-Kai Ng. 2012. Brek12: a book recommender for k-12 users. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1037–1038.

Subha Perni, Michael K Rooney, David P Horowitz, Daniel W Golden, Anne R McCall, Andrew J Einstein, and Reshma Jagsi. 2019. Assessment of use, specificity, and readability of written clinical informed consent forms for patients with cancer undergoing radiotherapy. *JAMA oncology*, 5(8):e190260–e190260.

Sarah E Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A readable read: Automatic assessment of language learning materials based on linguistic complexity. *arXiv preprint arXiv:1603.08868.*

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Luz Rello, Horacio Saggion, Ricardo Baeza-Yates, and Eduardo Graells. 2012. Graphical schemes may improve readability but not understandability for people with dyslexia. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 25–32. Association for Computational Linguistics.

Antony Sare, Aesha Patel, Pankti Kothari, Abhishek Kumar, Nitin Patel, and Pratik A Shukla. 2020. Readability assessment of internet-based patient education materials related to treatment options for benign prostatic hyperplasia. *Academic Radiology.*

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic assessment of japanese text readability based on a textbook corpus. In *LREC.*

Kathleen M Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, 115(2):184–209.

Kathleen M Sheehan. 2017. Validating automated measures of text complexity. *Educational Measurement: Issues and Practice*, 36(4):35–43.

Wade Shen, Jennifer Williams, Tamas Marius, and Elizabeth Salesky. 2013. A language-independent approach to automatic text difficulty assessment for second-language learners. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 30–38.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

Sanja Štajner and Sergiu Nisioi. 2018. A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).*

Sanja Štajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. International Joint Conferences on Artificial Intelligence.

Edward L Thorndike. 1921. The teacher's word book.

Amalia Todirascu, Thomas François, Nuria Gala, Cédric Fairon, Anne-Laure Ligozat, and Delphine Bernhard. 2013. Coherence and cohesion for the assessment of text readability. *Natural Language Processing and Cognitive Science*, 11:11–19.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Ivana Lucic. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2013. On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68.

Sowmya Vajjala and Detmar Meurers. 2014a. 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 (PITR)*, pages 21–29.

Sowmya Vajjala and Detmar Meurers. 2014b. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2):194–222.

Sowmya Vajjala, Detmar Meurers, Alexander Eitel, and Katharina Scheiter. 2016. Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 38–48.

Sheila W Valencia, Karen K Wixson, and P David Pearson. 2014. Putting text complexity in context: Refocusing on comprehension of complex text. *The Elementary School Journal*, 115(2):270–289.

Mabel Vogel and Carleton Washburne. 1928. An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, 28(5):373–381.

Tim vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep indicators. *Informatica*, 32(4).

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

QUAFF: Pilot Experiment

Michel Simard[1], Roland Kuhn[1], and Judith Rémillard[2]

[1]Multilingual Text Processing, NRC
[2]School of Translation and Interpretation, University of Ottawa

Executive Summary

This document describes the results of the QUAFF human evaluation of the quality of English-to-French machine translation (MT) outputs from the NRC's Portage system. The primary goal of the QUAFF pilot study was to find a metric for MT quality that would supplement the metric we have used almost exclusively until now, BLEU (Papineni et al., 2002). BLEU is an automatic metric based on the number of N-gram matches between MT outputs and reference translations. Though they are useful for guiding work on MT systems inside NRC, BLEU scores are too mathematically complex to explain in simple terms to our customers. We were looking for a metric based on human assessments of translation quality that would allow us to communicate in a non-technical way how much better a given commercial release of Portage is than the previous version. Ideally, the metric would cost as little as possible.

We believe that we have found a metric that meets these criteria. Recently, the tendency in human evaluations of MT quality has been to make them more and more complex – for instance, by counting the number of words required to post-edit outputs in order to attain reasonable quality, or by employing eye-trackers and timers to monitor exactly what happens during human post-editing of an MT output. We have gone in the opposite direction: simplicity. We ask the evaluators to choose between two possible translations of a source sentence. One of the translations is from *baseline* (the previous commercial release) and the other from *advanced* (the upcoming commercial release). In each example, the two outputs are presented in random order, so the evaluator does not know which output came from which system. Over a set of examples, the proportion of times the *advanced* output is preferred over the *baseline* output provides a measure of how much better the new version of Portage is than the old one.

Choosing between two translations for the same source sentence can be done quickly – for instance, much more quickly than post-editing. This means that we can obtain a large number of data points without spending a lot of money. In the experiments reported here, we asked five professional translators to perform a total of 3000 pairwise choices, paying them a total of $1500. Two test files were used, from two different domains. For the first domain, the score for the *advanced* outputs was about +1.2 BLEU higher than for the *baseline* outputs, and for the other domain, *advanced* had a BLEU advantage of about +1.0 BLEU over *baseline*[1].

Our analysis of the results:

- On individual examples, there was a fairly high level of disagreement between evaluators, and even within the same evaluator when an individual was asked to make the same pairwise choice at different times.
- Nevertheless, agreement on which **system** yielded better outputs was remarkably high. All evaluators preferred the *advanced* outputs to the *baseline* ones overall, by a ratio that varied from 1.4 to 2.1 across the five evaluators (these ratios exclude cases where the evaluators decided that they had no preference between the two outputs). On average, the preference ratio was 2.0 for the first domain and 1.8 for the second domain. Thus, BLEU and the independently performed QUAFF human evaluation were in complete agreement as to which system was better. If the *advanced* system represented an actual commercial release – it doesn't (the new commercial version of Portage we plan to release in the fall has an English-to-French BLEU score considerably higher than the *advanced* system in the QUAFF experiments) – we could tell our customers with complete accuracy, "Excluding ties, your post-editors are likely to prefer the outputs from the new release of Portage in the ratio two to one over the outputs from the previous release." A statement of this type is far more comprehensible than one involving BLEU scores.
- Statistical analysis given in this document showed that we could have reached this conclusion - *advanced* is significantly better than *baseline* – with a smaller number of pairwise comparisons. That is, for the kind of quality difference that held between the two systems compared in our experiments, we could spend even less than $1500 on a QUAFF evaluation and still reach a statistically valid conclusion. Furthermore, our analysis has shown how

[1]BLEU scores range between 0 and 1, but are typically reported in percentage points; higher scores denote better translations, and an increase of 0.5 - 1 BLEU is typically viewed as denoting a significant improvement.

to optimize statistical resolution for a fixed number of annotated examples (by spreading those examples across several evaluators, instead of having one evaluator work on many examples).

- We only had one regret about the design of the QUAFF pilot study: in retrospect, we should have insisted on a binary choice between the two translations. The questionnaire also allowed the choices "Both translations are good" and "Both translations are bad"; it is our impression that the latter was greatly overused. Of course, there will always be examples where the *baseline* output and the *advanced* output are of equivalent quality, but it is possible to estimate the proportion of such genuine ties from multiple evaluations while insisting that evaluators always pick one output or the other (how to do this is described in the Discussion section below). Future QUAFF evaluations will involve strictly binary choices.

This pilot study has shown that the QUAFF methodology of pairwise comparison between outputs from *baseline* and *advanced* versions of Portage, presented to the evaluators in random order, is a highly cost-effective way of measuring quality improvements from one Portage release to another. We recommend that a QUAFF evaluation be performed prior to each major commercial release of the software, and communicate its results to our customers. In particular, we recommend that a QUAFF evaluation be carried out prior to the release of Portage II 3.0 in autumn 2015.

1 Introduction

The goal of the QUAFF project is to create a set of lightweight benchmarks that we can apply over the years to test the progress of the Portage machine translation technology on language pairs that are used for supporting professional translators. In most cases, these professional translators will be post-editing the Portage outputs. The benchmarks will also enable us to choose whether or not to incorporate particular new techniques in Portage Shared (PS). Currently, the only metric we use on a regular basis is an automatic one: BLEU. It correlates only mildly with human judgments. Suppose that in the research branch of Portage, a certain new technique gives us a +0.5 BLEU improvement in many conditions, but that incorporating it in PS would significantly complicate the user interface. Should we build the technique into PS or not? (A gain of +0.5 BLEU is nice to have, but not overwhelming). If we knew that the new technique did not result in outputs that seemed significantly better to humans, it would clearly be preferable to omit the technique from PS.

The previous paragraph is deliberately vague about the meaning of "human judgments": in fact, humans evaluate the quality of translations in many different ways. For language pairs for which Portage's output is post-edited, it seems appropriate to supplement BLEU with metrics related to things human translators care about, such as perceptions of quality or post-editing effort (for language pairs where Portage is used for gisting, such as Arabic to English or Chinese to English, metrics related to comprehension might be more appropriate).

Portage output currently being post-edited is mainly in the direction English to French (in Canada, about 80% of translations between the two official languages are in this direction). The QUAFF project focuses on this direction. If the English to French benchmarks are successful, we may eventually set up analogous French to English benchmarks.

This report describes a pilot experiment that was carried out in early 2015, to test a first, simple evaluation protocol, which relies on straightforward pairwise comparisons.

2 Metrics

We seek metrics to supplement BLEU that reflect post-editors' perceptions of translation quality. Productivity is what translation agencies care about most, but an informal survey of the literature suggests that this is very tricky to measure:

1. The difficulty of defining what "time spent translating" means. If a translator or post-editor pauses for 90 seconds while staring out of the window, is he pondering how to translate a tricky idiom, or thinking about a TV show or a recent quarrel with a lover?

2. Enormous amounts of variability in productivity are observed, that depend on who the translator is, what his or her usual working environment is, what the text is, and even how he is being paid (the same translator might focus on quality if being paid by the hour, and on speed if paid by the word).

It is possible that measuring the change in productivity when translators post-edit Portage output will become easier because of new online tools being developed by projects like MATECAT.[2] If that happens, we will certainly try to incorporate direct measurements of productivity into our benchmarks.

[2]http://www.matecat.com/matecat/the-project

At the moment, however, trying to measure productivity would be difficult and expensive, and would take the group far out of its zone of expertise.

This leaves two obvious candidates for our human benchmarks in the short term:

Pairwise Comparison That is, the human evaluator is presented with the source (English) sentence, and two French Portage outputs in random order (randomized anew for each example): one from the *baseline* system, and one from an *advanced* system. The evaluator simply indicates a preference: "1", "2", or "=" if neither is preferred. Evaluator choices "1" and "2" are mapped by software onto the true underlying labels, ie the *baseline* or *advanced* condition. Pairwise comparison is in fact just a special case of *N-way ranking* (typically $N = 5$, with ties), which has been used extensively for the WMT shared tasks over the years Bojar et al. (2014). Ranking is appropriate for WMT, which compares several systems (e.g. as many as 18 for the English-German task in 2014), but is probably much more cognitively demanding for evaluators than pairwise comparison.

HTER ("Human-targeted Translation Error Rate"). The human evaluator is presented with the source sentence and a single translation: either from the *baseline* version of Portage, or from the *advanced* version. The evaluator post-edits the Portage output until he or she is happy with it. HTER is based on the number of words the evaluator deletes, inserts, and changes. In this case, each evaluator should only ever be given one version of the output, because otherwise they will be "primed" when they start working on a translation for a given source sentence the second time. Often as part of the HTER protocol, evaluators are explicitly instructed to minimize the changes they make to the MT output. For NRC–internal evaluation, it makes more sense to pay them a fixed amount for improving a given number of MT outputs, and let human nature take its course. (If the evaluation is done for a specific client, then the requirements or normal procedures of that client should prevail).

These two metrics both have advantages and disadvantages. Prior to this project, an informal pairwise comparison was conducted within the group, involving 110 examples, each consisting of a sentence triplet: an English sentence from Hansard and two different French translations of that sentence, one from a baseline version of Portage and one from a version of Portage that used a new technique (coarse LMs). Of course, only examples where

Preference	Evaluator 1	Evaluator 2	Evaluator 3
advanced	34%	43%	31%
baseline	15%	22%	18%
no pref.	52%	35%	43%
Improvement ratio	2.3	2.0	1.7

Table 1: Pre-pilot Pairwise comparison

baseline and *advanced* differed in at least one word were included among the 110 examples. The test was scientifically flawed – the three evaluators knew which output was *baseline* and which was *advanced* – but instructive. Although evaluators differed on many specific examples, all of them perceived an advantage for *advanced*. Table 1 shows the results. The last line in this table gives the *Improvement* ratio, i.e. the count ratio between cases where the *advanced* condition made things better and those where it made things worse, compared to the *baseline*:

$$\text{Improvement ratio} = \mathit{advanced}/\mathit{baseline}$$

Points to note:

- The main difference between the evaluators is the percentage of "no preference": Evaluator 2 put far fewer cases into that category. There were plenty of disagreements about specific examples. However, there was a consensus that the *advanced* technique makes translations better about twice as often as it makes them worse. This is very encouraging.
- Long sentences were much harder to compare than short ones. Partial solutions to this problem are discussed below.
- All evaluators had trouble explaining their preferences, either in individual cases or globally. For instance, they were unable to say exactly how the *advanced* outputs were better than the *baseline* ones. The most honest summary would be something like "On average, these translations just seem a bit better" which is not very helpful. This observation has influenced the plan, because it suggests that it may be difficult to collect detailed observations about the nature of the changes caused by new techniques (while it may often be easy to determine if a given change is globally beneficial or harmful).

The advantage of HTER is that it has been demonstrated in a number of studies to be inversely correlated with post-editor productivity, while not requiring time measurements. Its disadvantage is that it's much more expensive than pairwise comparison: it obviously takes much longer to correct a translation output by Portage than to decide whether or not you prefer it to an alternative translation by Portage.

This motivates our choice to evaluate according to pairwise comparison, which will allow us to gather a fairly large amount of data quickly and cheaply (especially if we can mitigate the "long sentence" problem). At a later stage, we should also evaluate according to HTER (or user productivity if MATECAT or other environments make measuring productivity easier). It would be interesting, at this later stage, to see how well conclusions from pairwise comparison correlate with HTER or productivity.

In any case, maybe productivity isn't the only thing we care about. Suppose a new technique – that by definition shows some BLEU improvement, since we will never be considering for inclusion in PS techniques that don't yield any BLEU gains – shows significant gains in perceived quality as revealed by blind human pairwise comparisons, and much later we find out that it doesn't yield any improvement in productivity. Would we regret including the technique in PS? Almost certainly not. Making post-editors happy will increase their acceptance of our technology, and from our point of view, that is almost as important a goal as improving their productivity. So an argument can be made for pairwise comparison as reflecting a criterion that is not productivity but that we do care about.

There is an interesting caveat to this reasoning. Imagine a technique that improves Portage's outputs somewhat when they're really awful, but not when they're good – i.e., that changes outputs from "terrible" to "bad", but not from "terrible", "bad", or "mediocre" to "good" or "excellent". Such a technique would show improvements according to pairwise comparison (and possibly BLEU), but be uncorrelated with both the satisfaction of post-editors and their productivity, since even the "bad" yet improved sentences produced by the technique will ultimately be ignored – the post-editor will basically translate from scratch.

We will need to keep an eye on this possibility (e.g., by tracking whether sentence pairs where the *advanced* version is ranked as better than the *baseline* one have lower sentence-level BLEU than average *advanced* sentences). One possibility is to cast the evaluation as "pre-post-editing": formulate the annotation question as "if you had to post-edit one of these two translations, which one would you pick?" then allow a third answer "none of these two – I'd rather translate from scratch". Another possibility is to ask evaluators two questions: 1) "Which of these two is the best translation?" And 2) "Is at least one of these good enough for post-editing purposes?" In this study, we opted for a variant of the first solution: we offered two alternative answers for situations where neither of the two translations was preferred: "Both translations are equivalently bad", and "Both translations are equivalently good".

3 Pilot Study

This QUAFF study involves comparison by human evaluators of outputs from a *baseline* version of Portage with outputs from an *advanced* version, on two different text domains.

3.1 Data

The two data experimental combinations involve two different data scenarios, from two domains of the `gc.ca` corpus:

Environment This is a "small data" scenario with about 250K English-French training sentence pairs;

Health This is a "medium data" scenario with about 500K English-French training sentence pairs.

For each data scenario, tuning and test data were drawn from the same domain as the training data. For each experimental combination, we trained *baseline* and *advanced* models, then generated *triplets* from the held-out test data: each triplet consists of the input English source sentence and the two French outputs in random order, one from the *baseline* version of Portage and one from the *advanced* version.

We know from prior experience that excessively long sentences are notoriously difficult to evaluate. For this reason, in this exercise, we excluded source sentences longer than 50 words. Very short sentences are also typically difficult to evaluate out of context, and so we also excluded sentences shorter than 5 words.

For this pilot study, our goal was to produce 1200 annotated triplets, i.e. 600 from each of the *Environment* and *Health* data scenarios. However, we wanted evaluators to work only on triplets where *baseline* and *advanced* are different, and because we did not know in advance to what extent the two experimental conditions produced different results, we actually held out much more test data than the number of triplets we planned to evaluate, in the order of several thousands for each data scenario.

3.2 Systems

Both versions of the Portage MT system were implemented in the R & D branch of Portage, and were meant to approximate commercial releases: *baseline* closely resembles the Portage version our clients are currently using, while *advanced* was our best guess at what the next commercial release will look like. The *advanced* system has all the capabilities in *baseline* (alignments from IBM2 and HMM3, batch LMIRA, advanced casing capabilities, etc.) plus the following new ones:

- DHDM - hierarchical reordering with some related sparse features
- Other sparse features – Hop-May (but not the expensive, complex ones, and no indicator features)
- Coarse LMs and coarse biLMs.

The *advanced* system did not include mix LMs or mix TMs. Some preliminary BLEU testing was done to determine which of the techniques above, or which of their variants, would be included in *advanced*. Another criterion was speed/memory usage. We ended up deciding to use two coarse LMS, both unpruned and 8-gram: the coarse LM with 200 word clusters and the coarse LM with 800 word clusters. In order to economize on storage space, we decided to use a single, pruned, 6-gram coarse biLM; it had configuration 400 bi(400,400).[3]

For each of *Environment* and *Health* training, we used the tuning weights, out of 5 different sets of tuning weights tried, that yielded the median BLEU score. The average BLEU scores for the two systems over the 5 runs on test data are given in Table 2.

3.3 Annotation

The annotation work took the form of individual annotation "tasks". In each of these task, each evaluator was shown three pieces of information:

[3]Subsequently – after the outputs were generated – we changed the definition of the next commercial release: the *advanced* version that will be given to our clients in autumn 2015 is not the same as that used for this evaluation. That does not affect the usefulness of the current study, whose goal was to see if QUAFF-style evaluation could yield actionable conclusions when comparing one version of Portage to another. Closer to the actual release date, we are planning to carry out another QUAFF evaluation comparing the true *advanced* version that will be in the next commercial release to the *baseline* in the previous commercial release.

System	Domain			
	Environment		Health	
baseline	35.92		37.77	
advanced	37.08	(+1.16)	38.80	(+1.03)

Table 2: System performance (BLEU)

S : A source-language sentence (in English).

T1 and T2 : Two target-language translations (in French).

The evaluator was then asked if he preferred translation T1 or T2, or if the two translations were equivalently good, or equivalently bad. If, for some reason, it was not possible to annotate the translations along these lines, evaluators could signal this by checking "Other", and invited to leave a comment, explaining the problem. (In fact, it was always possible to comment individual tasks.) All annotations were collected via a web-based interface. An example task is shown in Figure 1.

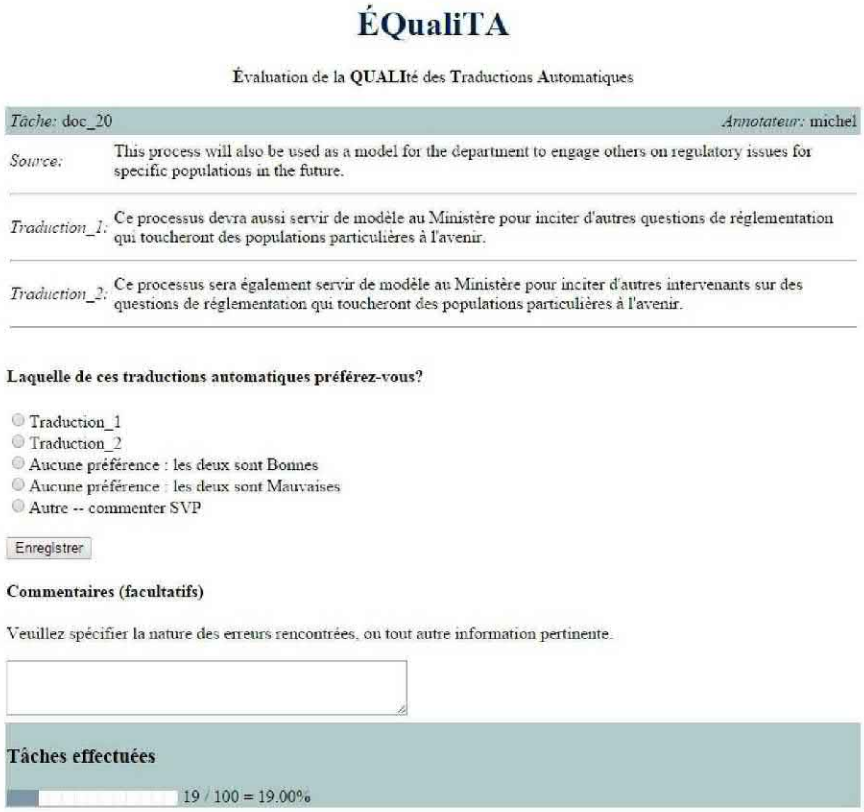
Tasks were blind and randomized: to avoid bias, output from the *baseline* and *advanced* systems was randomly shuffled within tasks, i.e. sometimes T1 was the output from the *baseline* system and T2 the output from the *advanced* system, and sometimes the other way around, and the origin of each translation was not shown to the evaluators in any way.

Five professional EN-FR translators acted as evaluators in the evaluation. Annotations were performed over a period of approximately three months, in two distinct phases:

The first phase concentrated on a set of 200 tasks, with each evaluator performing each task twice. In practice, tasks were organized in batches of 100 tasks. Each evaluator was assigned 4 batches; batches 1 and 3 were identical, as were batches 2 and 4. As a result, each of the phase 1 tasks was performed 10 times, i.e. twice by each of the five evaluators.

The second phase, which was carried out about a month later, focused on a second set of 1000 tasks: this time again, tasks were organized into batches of 100 tasks. But this time, each task was assigned to a single evaluator. This produced 1000 singly-annotated tasks. [4]

[4]This two-phase design was actually the result of a data manipulation error: translators were not supposed to perform each task twice during Phase 1. Instead, they were supposed to perform 200 "common" tasks, followed by 200 "unique" tasks. However, this error turned out to be convenient, because it allowed us to measure intra-annotator agreement (see Section 4.3).



ÉQualiTA

Évaluation de la QUALIté des Traductions Automatiques

Tâche: doc_20	*Annotateur:* michel
Source:	This process will also be used as a model for the department to engage others on regulatory issues for specific populations in the future.
Traduction_1:	Ce processus devra aussi servir de modèle au Ministère pour inciter d'autres questions de réglementation qui toucheront des populations particulières à l'avenir.
Traduction_2:	Ce processus sera également servir de modèle au Ministère pour inciter d'autres intervenants sur des questions de réglementation qui toucheront des populations particulières à l'avenir.

Laquelle de ces traductions automatiques préférez-vous?

- Traduction_1
- Traduction_2
- Aucune préférence : les deux sont Bonnes
- Aucune préférence : les deux sont Mauvaises
- Autre -- commenter SVP

Enregistrer

Commentaires (facultatifs)

Veuillez spécifier la nature des erreurs rencontrées, ou tout autre information pertinente.

Tâches effectuées

19 / 100 = 19.00%

Figure 1: An Evaluation Task

Globally, the five evaluators performed 3000 annotation tasks. Evaluators were paid \$0.50 per task. The total cost for the annotations was therefore \$1500.

4 Results

4.1 Global Results

The evaluation procedure described in the previous section produced a set of 3000 annotations, for 1200 distinct translation pairs: each pair was assigned one or several of the following labels:

advanced : The translation from the *advanced* system is preferred

baseline : The translation from the *baseline* system is preferred

both_bad : Both translations are equivalently bad

both_good : Both translations are equivalently good

other : Impossible to evaluate

For sentences with multiple annotations (Annotation Phase 1), we assigned the most frequently assigned label; in cases where *baseline* and *advanced* were tied, we assign the label *both_good*.

Table 3 shows absolute and relative counts for each label, as well as the Improvement Ratio (Section 2). Globally, these results seem to indicate a general preference for the *advanced* translations over the *baseline* translations. This is true for both the Environment and Health domains.

The *Improvement* ratio shows how much perceived quality increase is caused by the *advanced* technique, in cases where this condition makes a difference. By itself, it does not give an accurate idea of the *impact* of the *advanced* condition, because it ignores cases where not a single word differs between the *baseline* and *advanced* outputs, or where *baseline* and *advanced* differed but evaluators didn't see a quality difference (triplets labeled *both_bad* or *both_good*). To account for these situations, we calculate the percentages of test examples where *advanced* and *baseline* translations were preferred over all test examples, including those where both translations were identical. We then calculate the "*Impact percentage*" as the difference between these percentages:

$$\text{Impact percentage} = \%advanced - \%baseline$$

		Domain: Environment		Health		All Domains	
Preference:	advanced	161	(26.8%)	172	(28.7%)	333	(27.8%)
	baseline	82	(13.7%)	96	(16.0%)	178	(14.8%)
	both_good	68	(11.3%)	94	(15.7%)	162	(13.5%)
	both_bad	285	(47.5%)	235	(39.2%)	520	(43.3%)
	other	4	(0.7%)	3	(0.5%)	7	(0.6%)
Total		600		600		1200	
Improvement ratio		2.0		1.8		1.9	

Table 3: Annotation Results

	Environment	Health	All Domains
Identical translations	22.6 %	26.2 %	24.4 %
advanced is preferred	20.8 %	21.2 %	21.0 %
baseline is preferred	10.6 %	11.8 %	11.2 %
no preference	45.5 %	40.5 %	42.9 %
Impact = % *advanced* - % *baseline*	10.2 %	9.4 %	9.8 %

Table 4: Label distribution, adjusted to reflect sampling bias.

In practice, the two systems under evaluation produce identical translations for 24.4% of input sentences, i.e. 22.6% of Environment and 26.2% of Health sentences.[5] Table 4 shows label percentages adjusted to reflect this bias in the sampling of the labeled data, and the *Impact percentage*, which takes into account these adjusted percentages.[6] Here again, we observe a clear preference for the *advanced* system: globally, it produces better translations than the *baseline* for 9.8% of our test set.

[5]These percentages are computed on input sentences between 5 and 50 words. When all sentence lengths are considered, the *baseline* and *advanced* systems produce identical outputs for 31.6% of Environment sentences and 38.7% of Health sentences

[6]In this table, "*both_good*" and "*both_bad*" labels were merged into a single, "*no preference*" category.

4.2 Statistical Significance

While it seems safe to conclude from the results of the previous section that the *advanced* system indeed produces better translations that the *baseline* system, we must analyze to what extent these numbers are reliable. Specifically, we want to know whether the observed values for the Improvement Ratio are significantly greater than 1 (or, equivalently, if Impact is significantly greater than 0). But more generally, we would like to know whether the numbers of annotations and of evaluators are adequate for this kind of study, or if we could do with less annotations or fewer evaluators.

We use a bootstrap resampling approach to examine this question: we draw random samples from our annotated data (with replacement), to produce sets of different sizes, and with different number of evaluators. From each sample, we compute the Improvement ratio I. We repeat this process many times (in practice: 1000 times), to estimate the probability that $I <= 0$, i.e. the chances that an evaluation campaign lead us to conclude that the advanced system is not better than the baseline. In this sampling procedure, we assume that there may be multiple evaluators, but that each sentence is annotated by only one evaluator (no multiple annotations).

Figure 2 shows the results of this procedure. In this figure, the black curve corresponds to the situation with one evaluator: if a single evaluator is asked to annotate 50 pairs of translations, then the probability that he prefers more *baseline* translations than *advanced* translations is 0.18 (top left corner). If instead we ask him to annotate 100 pairs, this probability drops to 0.09. At 500 pairs, the probability drops below 0.01.

With two evaluators (red curve), the probability of observing less *advanced* than *baseline* labels out of 50 pairs of translations (i.e. an average of 25 pairs annotated by each translator) is 0.12, and it drops below 0.01 as soon as 250 pairs are annotated. With three evaluators (blue curve), the 0.01 significance level is reached somewhere around 175 pairs. The behavior with five annotators (green curve) is not strikingly different from that with three.

Figure 3 shows similar plots for Impact. The graph on the left is analogous to Figure 2: it shows the probability of observing a null or negative impact, as a function of the numbers of annotations and evaluators. The graph to the right shows what happens when we consider a more stringent requirement: that the impact be strictly greater than 4%. In both cases, reliable conclusions are reached faster with more evaluators.

With regard to the current exercise, with 1200 annotated sentences and 5 evaluators, the probability that the *advanced* system is not better than

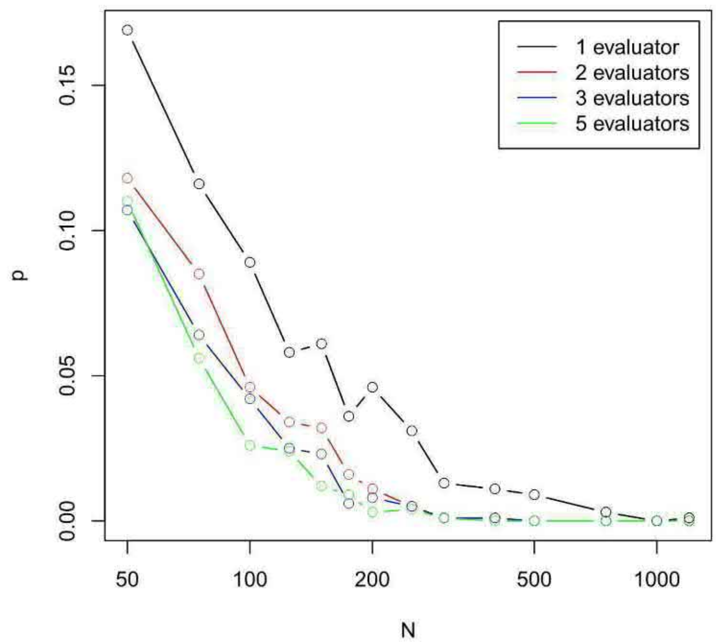


Figure 2: Probability that Improvement Ratio ≤ 1 (p), as a function of the number of annotated sentences (N), with 1, 2, 3 or 5 evaluators

the *baseline* is actually negligible.

4.3 Evaluator Agreement

As mentioned earlier, the annotation process was designed in such a way that a subset of the sampled data (200 sentences) was annotated by all evaluators. The intention was to allow the analysis of inter-annotator agreement. In addition, as a result of a manipulation error, these 200 sentences were all annotated *twice* by each evaluator, which also makes it possible to study *intra*-annotator agreement, i.e. the extent to which each evaluator assigns identical labels to identical tasks.

Table 5 shows all pairwise agreement on these 200 common tasks, measured using Cohen's κ (see Appendix A.1). Each line and column corresponds to a set of annotations for these tasks: "1.a" is the result of the

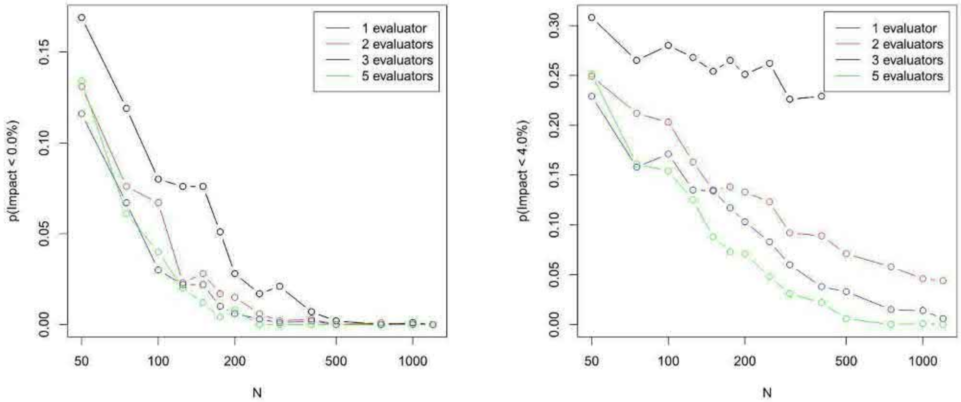


Figure 3: Probability that Impact Percentage ≤ 0 (left) and $\leq 4\%$ (right), as a function of the number of annotated sentences (N), with 1, 2, 3 or 5 evaluators

first pass of annotation by Evaluator 1, "1.b" is his second pass, "2.a" is Evaluator 2's first pass, etc. Figures in *italics* on the main diagonal are *intra*-annotator agreements; all other figures are *inter*-annotator.

Agreement between all annotations, as measured by Fleiss's κ, is 0.28. The average pairwise *inter*-annotator agreement, as measured with Cohen's κ is 0.26. The average *intra*-annotator agreement, as measured with Cohen's κ is 0.57. The last line of Table 5 gives the average κ for each evaluator. This suggests that Evaluators 2 and 3 tend to agree more with the others, while Evaluators 1 and 4 are the Black Sheep of the lot (in fact, they strongly disagree with one another more than anything else). Globally, of the 200 tasks assigned to all evaluators, only 37 were assigned the same label by all evaluators during the first annotation pass (7 *advanced*, 5 *baseline* and 25 *both_bad*); at the second annotation pass, this number goes down to 25 (7 *advanced*, 1 *baseline* and 17 *both_bad*).

There are no hard rules for interpreting κ, but the literature (Landis and Koch, 1977) suggests that $\kappa \in [0.21, 0.40]$ denotes "fair agreement", while $\kappa \in [0.41, 0.60]$ is "moderate". To better understand what these numbers mean, it is perhaps more instructive to consider specific examples. Figure 4 shows two examples of intra-annotator agreement matrices. The first one is for Evaluator 1, who displays the highest κ, i.e. the most consistent behavior. When re-labeling the data, Evaluator 1 assigns the same label to

Annotation	1.a	1.b	2.a	2.b	3.a	3.b	4.a	4.b	5.a
1.b	*0.69*								
2.a	0.19	0.14							
2.b	0.29	0.21	*0.57*						
3.a	0.28	0.24	0.38	0.43					
3.b	0.25	0.26	0.27	0.40	*0.51*				
4.a	0.10	0.08	0.39	0.31	0.26	0.21			
4.b	0.17	0.11	0.41	0.37	0.23	0.19	*0.54*		
5.a	0.37	0.28	0.22	0.30	0.35	0.30	0.18	0.21	
5.b	0.38	0.27	0.20	0.29	0.33	0.38	0.14	0.18	*0.55*
Average	**0.23**		**0.30**		**0.30**		**0.22**		**0.27**

Table 5: Annotator agreement: Pairwise annotator agreements is measured using Cohen's κ, for all annotations produced by evaluators. Figures in *italics* are *intra*-annotator agreements. Per evaluator average κ are computed over all *inter*-annotator figures for that evaluator.

individual translations 80% of the time. In practice, many examples that Evaluator 1 had initially labeled as *both_bad* or *both_good*, he relabeled as either *advanced* or *baseline* the second time around. This suggests that his judgement became more discriminant over time.

In contrast, Evaluator 3 is the least consistent ($\kappa = 0.51$), with only 65.5% identical annotations between the two rounds. Upon analysis, Evaluator 3 shows the opposite behavior: many examples for which he initially preferred the *baseline* or *advanced* translation were later re-labeled as *both_bad*. It is also interesting to note that Evaluator 3 uses the *both_bad* and *bath_good* labels much more frequently than Evaluator 1. We come back to this later.

Figure 5 shows examples of agreement matrices between pairs of evaluators. Example 1 shows relatively high agreement between evaluators 3 and 5. Many of their disagreements revolve around situations where one evaluator prefers one of the translations, while the other feels that both are bad. But there is also a surprising number of instances (16) in which one prefers the *baseline* and the other the *advanced*.

Example 2 shows an even higher level of agreement ($\kappa = 0.41$). But what this really reflects is the fact that both Evaluators 2 and 4 have a very strong tendency to use the *both_bad* label (70.5% of Evaluator 2's annotations, 72.5% of Evaluator 4's).

Example 3 shows the highest level of disagreement between two evaluators (1 and 4: $\kappa = 0.10$). In this case, evaluators disagree over most

Evaluator 1: ($\kappa = 0.69$)

	advanced	baseline	both_good	both_bad	other
advanced	84	4	0	1	0
baseline	9	50	1	2	0
both_good	3	3	5	0	0
both_bad	10	4	1	21	0
other	0	1	0	1	0

Evaluator 3: ($\kappa = 0.51$)

	advanced	baseline	both_good	both_bad	other
advanced	34	2	9	7	0
baseline	3	8	11	5	0
both_good	3	2	21	4	0
both_bad	9	7	6	68	1
other	0	0	0	0	0

Figure 4: Examples of intra-annotator agreements and disagreements

annotations: most situations where Evaluator 1 prefers either the *baseline* or *advanced* translation, Evaluator 4 labels as *both_bad*.

When designing the annotation scheme, the *both_good* and *both_bad* labels were intended to be used as last resorts, in case of ties, i.e. situations where it was not possible to decide which translation was better (or worse) than the other: *both_good* was to be used in situations where both translations required no manual corrections, and *both_bad* in all other cases.[7] When later examining individual evaluator's work, it rapidly became apparent that some evaluators had a strikingly different interpretation of the *both_bad* label, as simply meaning that none of the two proposed translations was *good*. Clearly, from a professional translator's perspective, this is the case for most MT output.

Table 6 shows the distribution of labels for each evaluator. Evaluators 2 and 4 use the label *both_bad* in the vast majority of cases (66% and 70%), illustrating the point above. In contrast, Evaluators 1 uses that label very

[7] In practice, the evaluators did not see these labels; instead, they saw short phrases summarizing their intended interpretation. These can be seen in Figure 1. The phrase for label *both_good* was "Aucune préférence: les deux sont bonnes" ("No preference: both are good"); the phrase for label *both_bad* was "Aucune préférence: les deux sont mauvaises" ("No preference: both are bad").

Example 1: 3.a VS. 5.a ($\kappa = 0.35$)

	advanced	baseline	both_good	both_bad	other
advanced	34	11	8	28	0
baseline	5	14	13	8	0
both_good	2	1	6	1	0
both_bad	11	1	3	54	0
other	0	0	0	0	0

Example 2: 2.a VS. 4.b ($\kappa = 0.41$)

	advanced	baseline	both_good	both_bad	other
advanced	9	0	1	16	0
baseline	0	9	3	11	1
both_good	0	2	6	1	0
both_bad	8	3	2	124	4
other	0	0	0	0	0

Example 3: 1.a VS. 4.a ($\kappa = 0.10$)

	advanced	baseline	both_good	both_bad	other
advanced	15	6	3	0	0
baseline	4	9	1	2	0
both_good	3	2	1	0	0
both_bad	63	44	5	32	1
other	4	1	1	2	1

Figure 5: Examples of inter-annotator agreements and disagreements

	Eval. 1	Eval. 2	Eval. 3	Eval. 4	Eval. 5
advanced	0.44	0.16	0.23	0.12	0.40
baseline	0.30	0.11	0.11	0.07	0.20
both_good	0.08	0.06	0.21	0.06	0.12
both_bad	0.17	0.66	0.44	0.71	0.27
other	0.00	0.00	0.00	0.04	0.00
Improvement	1.50	1.44	2.07	1.68	1.98
Impact	11.2%	3.8%	9.1%	3.6%	15.1%

Table 6: Label distribution per evaluator

Label	Count
both_good	85
both_bad	140
other	24
advanced + baseline	104
Total	353

Table 7: Labels to which evaluator comments are associated.

sparingly (17%), showing an interpretation of its meaning much closer to the intended. Evaluators 3 and 5 display intermediate behaviors.

It is remarkable that, despite the striking variabilities in the use of the "neutral" labels (*both_bad* and *both_good*), the relative proportions of *advanced* and *baseline* labels are quite similar between evaluators ("Improvement" in Table 6). Per-evaluator "Impact" numbers display a much wider variance, but this is normal, since they indirectly reflect the proportion of *advanced* and *baseline* labels relative to all test data.

4.4 Evaluator Comments

As mentioned earlier, it was possible for evaluators to attach free text comments to annotations, and they were explicitly invited to do so whenever they used the label *other*. In practice, this functionality was used with all labels.

In practice, many comments are non-informative, simply restating how the evaluator labeled the task. Informative comments are quite varied. We discuss below the most salient of these.

When comments were attached to tasks which the evaluators labeled as *other* (or sometimes as *both_bad/both_good*), the following reasons were most often mentioned:

Not enough context : Evaluators often found it difficult or impossible to decide which translation was better without sufficient context. This occurred in situations where specialized terminology was involved, or with very short segments.

Error/French in source : Errors in the source propagate to the output ("Garbage in, garbage out"). Evaluators sometimes identified such errors, and preferred not to evaluate translations in those cases.[8]

Segmentation error : This typically refers to a variant of the above: a situation where the input segment is badly segmented, resulting in MT errors. For example: missing words at the beginning or end of a sentence, untokenized punctuation, etc.

Identical Translations : Situations where both the *baseline* and *advanced* systems produced exactly the same output were explicitly excluded from the evaluation. There were rare cases however, where evaluators were presented pairs that appeared to be similar, although they differed by a tiny detail, such as different quote characters or apostrophes, etc.

Other comments referred to specific errors in one or both translations. Some comments were very general quality assessments, things like "meaning error", "misleading translation", "logical inconsistency", "clumsy formulation", etc. Other comments were more specific, such as:

- Missing word(s) in translation
- Added word(s) in translation (eg: inserted $ sign in numerical expression, superfluous determiners in titles)
- Verb-Subject inversion (or not) in questions
- Anglicisms and literal (non-idiomatic) translations
- Agreement errors

[8] If we were evaluating an advanced version of Portage that incorporated source error correction mechanisms, obviously, this would be a problem.

- Verb tense errors
- Wrong determiner
- Typographical error, letter case
- Punctuation error (comma)
- Anaphora error (wrong referent)
- Scoping (conjunctions, if...then construct, etc.)
- Terminology / Official names

At this stage, we have not tried to link specific types of errors to a specific system (*baseline* or *advanced*). Given the relatively low inter- and intra-annotator agreements on individual labels, we suspect that this kind of analysis would not be very reliable.

In one way or another, all comments were negative, except one, which is worth noting:

> "Incroyablement, la traduction automatique est meilleure que l'original!"

5 Discussion

The results presented in the previous section demonstrate quite convincingly that our *advanced* Portage system produces results that are significantly better than the *baseline* system. In fact, our analysis of statistical significance (Section 4.2) suggests that we could have reached the same conclusions with far fewer annotations. In practice, one important observation is the benefit of resorting to multiple evaluators.

However, our goals were quite modest here, and the only distinction we made with regard to the experimental conditions was on the text domain (*Environment* vs. *Health*). Had we wanted to draw finer distinctions (e.g. taking into account text length, associated BLEU or HTER scores, specific constructions, etc.), we might have in fact needed more data. Also, this study focused on a pair of systems that produce substantially different translations, with global performance that differs by over 1 BLEU; more annotations may be necessary when comparing systems that produce more subtly different results.

Should we eliminate the "neutral" labels from the labeling scheme in future evaluations? As noted earlier, the role of these labels is to allow evaluators to explicitly mark situations where it is not possible to differentiate between the two alternatives (more on this in Section 4.4). Without these labels, evaluators would be forced to make a choice, even when no choice is possible, which would essentially result in random assignments.

But, is it the case that evaluators who use "neutral" labels more sparingly are in fact assigning some *baseline* or *advanced* labels more-or-less randomly? If that was the case, these evaluators could be observed to be less discriminant than those who use neutral labels for all but the most clear-cut cases. Then, evaluators who use fewer neutral labels would have Improvement figures close to 1, as if a larger proportion of their *baseline* and *advanced* labels had been assigned randomly. In practice, this does not seem to be supported by evidence: the least discriminant evaluator, as given by his Improvement figure (Evaluator 2, Improvement=1.44) is the one with the second highest proportion of neutral labels (72%); and the most discriminant (Evaluator 5, Improvement=1.98) is the one with the second smallest proportion of neutral labels (39%).

Therefore, it appears we wouldn't lose much from eliminating neutral labels. In fact, in a purely binary evaluation scheme, i.e. one in which evaluators only ever make binary choices – "A is better" or "B is better" – it is still possible to estimate the percentage of examples where there is no quality difference between A and B. This can be done by assigning some examples to more than one evaluator (as was done in the first phase of evaluation for 200 examples). Different variants can be considered:

1. Every example is annotated by two people. Only examples where they agree are used to calculate the improvement statistic.

2. Every example is annotated by three people. Again, examples where they agree are the basis for calculating the improvement statistic.

3. A subset of examples is annotated by several people; the rest are each only annotated by one person. E.g., every annotator labels 100 examples in the "common subset" and 300 examples that the other annotators won't see. The numbers from the common subset are used to estimate the percentage of examples that have no quality difference.

There is a minor mathematical issue that arises in all these schemes. We want to find out what percentage of examples are "quality ties", where choices A and B are genuinely of equivalent quality. However, the percentage of disagreements underestimates the quality tie percentage.

An example: say 2 annotators label the same set of 100 examples. For each example, they must make a binary choice between translations A and B. Let's suppose they agree 90 times, and disagree 10 times. Does this mean the percentage of examples where A and B have equivalent quality = quality tie percentage = 10%?

No. Let AA denote the case where annotator 1 prefers A and annotator 2 also prefers A; let AB denote the case where #1 prefers A but #2 prefers B, etc. Now, among the quality ties (however many there are), the following outcomes should be equiprobable: AA, AB, BA, BB. If there are 10 quality ties among the 100 examples, only half of them on average should be disagreements – the AB and BA examples. I.e., only 5 will be disagreements. The other half will be of the AA or BB types – agreements.

I.e., if among the 100 examples, 10 are disagreements, the true quality tie percentage is 20%. A correction factor must be applied to account for the fact that 2 annotators will agree with each other by pure chance on half the "quality tie" examples. In general, for 2 annotators:

$$\text{quality tie percentage} = \text{disagreement percentage} \times 2$$

A second example: say 3 annotators label the same set of examples. How do we estimate the quality tie percentage? Consider the examples where quality of the two translations is tied. On these examples, the following 8 outcomes are equiprobable: AAA, AAB, ABA, BAA, BBA, BAB, ABB, BBB. Two of these, AAA and BBB, are agreements that occurred by chance. So to estimate the number of quality ties, one must take the number of disagreements and multiply by $8/6 = 4/3$. So if we had 100 examples, each labeled by three annotators, and they disagreed on 10 examples, the quality tie percentage = disagreement percentage $\times$ $4/3 = 10\% \times 1.33 = 13.3\%$.

The general formula for estimating quality tie percentage is as follows. If N annotators label each example, and $\%D$ is the percentage of examples on which they disagree, the true quality tie percentage $\%QT$ is:

$$\%QT = \%D \times \frac{2^{N-1}}{2^{N-1} - 1}$$

Given $\%QT$, it is possible to compute global statistics like the Improvement Ratio and the Impact percentage, without knowing specifically which examples are tied.

6 Conclusions

To compare the quality of translations produced by two versions of the Portage MT system, we performed a human evaluation, based on a simple pairwise comparison. The goal of this operation was to develop a reliable, effective and economical evaluation procedure that could be run periodically to support the development of the commercial version of Portage.

This pilot study focused on two versions of the Portage system: a *baseline* version, representative of the current commercial Portage offering, and an *advanced* version, which reflects the commercial version planned for Fall 2015. The evaluation, which involved five professional translators, resulted in a dataset of 1200 triples (source language segment, *baseline* translation, *advanced* translation) with labels denoting preference: *baseline* is better , *advanced* is better, or a neutral label (*both_good*, *both_bad* or *other*). From this data, we were able to conclude that the *advanced* Portage system produces better translations than its *baseline* counterpart for approximately 21% of test segments, while degrading another 11%. Therefore, the overall net positive impact of the techniques deployed in the *advanced* version is approximately 10%.

Our analysis of the statistical significance of the observed results suggests that, for the sole purpose of comparing two versions of the system on a single test set, far fewer triples need to be annotated than the 1200 that were produced. A standard confidence interval of 0.01 can be attained on Improvement Ratio with 500 annotated triples if working with a single evaluator, and as little as 200 triples if three evaluators share the annotation work. However, this observation is specific to this dataset. The quantity of annotations should be adjusted to account for the systems under comparison, and the degree of accuracy sought.

One aspect that comes out very clearly in our analysis is the importance of resorting to multiple evaluators. This can be partly explained by the relatively low agreement between annotators on specific triples. In our experiment, while all evaluators agreed that the *advanced* translations were globally better than the *baseline*, they very seldom agreed on which translation was better for specific examples. Even intra-annotator agreement was relatively low, suggesting that the task is very subjective, and that consistent labeling of triples cannot be expected in general. In particular, this implies that individual labels should not be trusted, and that this kind of evaluation procedure should probably not be used for in-depth analysis of specific translation errors.

One aspect that likely contributed to low inter-annotator agreements was the inconsistent use of neutral labels. In the absence of precise instructions, evaluators used these labels in very different ways: while some used them as last resorts for cases where no clear preference emerged, others used them systematically to denote cases where none of the translations seemed "good enough". For future evaluations, we plan to eliminate these neutral labels and to rely on other means (multiple annotations) to estimate the proportion of triples that cannot be differentiated.

References

Bojar, O., C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, et al. (2014). Findings of the 2014 workshop on statistical machine translation.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin 76*(5), 378.

Landis, J. R. and G. G. Koch (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics.

A Appendix

A.1 Measures of Inter-annotator Agreement

The typical measure of inter-annotator agreement is Cohen's κ ("kappa"). This only works for two annotators; for multiple annotators, we should use Fleiss's κ, but the idea is the same:

$$\kappa = (Pr(A)-Pr(E))/(1-Pr(E))$$

where $Pr(A)$ is the observed probability of agreement and $Pr(E)$ is the theoretical probability that annotators agree by chance.

Let's run through an example to see how this thing behaves. Say we have two annotators *Roland* and *Eric*, and these are their agreement statistics:

		Eric		
		advanced	*baseline*	"="
	advanced	20	5	10
Roland	*baseline*	5	40	15
	"="	15	10	30

There are 150 observations. Of these, Roland and Eric agree on 90 (20 *advanced*, 40 *baseline* and 30 "="): $Pr(A) = 90/150 = 0.60$. The probability of them agreeing by chance $Pr(E)$ is computed like this:

- For Roland, $Pr(\text{advanced}) = 35/150 = 0.23$, $Pr(\text{baseline}) = 60/150 = 0.4$ and $Pr(=) = 55/150 = 0.37$;
- For Eric, $Pr(\text{advanced}) = 40/150 = 0.27$, $Pr(\text{baseline}) = 55/150 = 0.37$ and $Pr(=) = 55/150 = 0.37$;
- The probability that both Eric and Roland pick *advanced* is $0.23 \times 0.27 = 0.06$;
- The probability that they both pick *baseline* is $0.4 \times 0.37 = 0.15$;
- The probability that they both pick "=" is $0.37 \times 0.37 = 0.14$.

So $Pr(E) = 0.06 + 0.15 + 0.14 = 0.35$ and $\kappa = (0.60\ 0.35)/(1\ 0.35) = 0.38$. That's a "fair" agreement on the Landis & Koch scale (Landis and Koch, 1977), and a "poor" one according to Fleiss (Fleiss, 1971).

The significance of that number is essentially just as good as our estimates of $Pr(E)$ and $Pr(A)$. In this case, the 95% confidence interval is ± 0.12, which means the real κ could be anywhere between 0.26 and 0.50.[9] Assuming twice as many observations (300 instead of 150) but the same overall distribution, the value of κ would remain the same (obviously), but the 95% confidence interval would shrink to ± 0.08 (κ in $[0.30, 0.46]$). So 100 observations is likely to be a bit tight to get a reasonable estimate of agreement. This motivates our decision to have 200 common tasks.

[9] http://graphpad.com/quickcalcs/kappa2