Institution: UNIVERSITY OF LEEDS

Unit of Assessment: UOA 28 MODERN LANGUAGES

Title of case study: Innovative computational linguistic technologies for language service providers

1. Summary of the impact

Building on their groundbreaking research and collaborative networks, Babych and Sharoff have developed a range of language technologies which now reach major corporations, small specialist businesses, a large industrial consortium, and agencies of the EU and UN. Their translation tools have had significant industrial impact by improving efficiency, consistency and user experience, and leveraging existing data collections for new purposes. In terms of policy, the research has re-shaped attitudes toward the ownership of data by demonstrating the commercial value of pooling resources. Individual translators have also benefitted from these technologies and related CPD courses, helping them to improve document flow, terminology and translation activities.

2. Underpinning research

Beginning in 2005, Babych (Research Fellow 2005-2010; Lecturer in the Centre for Translation Studies (CTS) 2010-present) and Sharoff (Lecturer 2005-2010, Senior Lecturer 2010-present) have had a proven track record of innovative research in computational linguistics, including development of large corpora (text collections), document classification and machine translation- (MT) evaluation. Their common vision is the development and sharing of state-of-the-art technologies with professional users, addressing real world tasks in industrial environments. Some of the research projects (in particular, WebDoc, ACCURAT, TTC, HyghTra [**3, 5, 6**]) involved leading industrial partners, thus completing a circle from research output to impact, and through to new research output. The research collaboration with TAUS (Translation Automation User Society) resulted in the creation of a 2-year fellowship and a TAUS-funded PhD studentship at the University of Leeds to develop intelligent access to translation resources. This and other projects have led world-leading researchers (including Richard Forsyth and Reinhard Rapp) to join the CTS (Centre for Translation Studies) team as Research Fellows. Adam Kilgarriff, Director of Lexical Computing Ltd, was also offered an honorary fellowship based on the strength of his collaboration with the group.

The project ASSIST (2005-2007, CoI: Sharoff, RF: Babych [**1**]) resulted in development of a methodology for solving translation problems difficult for human translators, in particular discovering translation equivalents for phrases which cannot be translated directly and which are not found in dictionaries (e.g. ‘comprehensive answer’ or ‘recreational fear’). This approach adopts ideas from distributional semantics, paraphrasing using words that share similar contexts, and then recombining their dictionary translations and paraphrases of translations. These recombinations are then crosschecked against large monolingual collections of conventional phrases in the target language. Instead of offering a single translation solution, the technology delivers a ranked list of alternatives, thus inspiring user creativity.

Sharoff's research on the collection of large language corpora [**2**] led to the development of user-friendly tools for (a) harvesting large text collections for a specific domain, such as aeronautical engineering or wind energy, or a specific purpose, such as translation or language teaching, and (b) enriching these collections with linguistic information, such as syntactic properties. The project IntelliText [**iv**] further developed the corpus access infrastructure “CSAR” that now hosts one of the world’s largest and richest collections of monolingual and bilingual corpora in 15 languages. Sharoff's research on automated methods of document analysis, similarity detection and classification within and across languages [**3, 4**] was funded via a number of streams, including ASSIST [**i**]**,** WebDoc [**iii**], and TTC [**v**]. This led to better understanding of the composition of the Web in term of genres, as well as to improvements in using Web texts for language teaching, translation and text analysis.

Babych's research on MT evaluation was supported by a Leverhulme Early Career Fellowship [**ii**], followed by an invitation to join the consortium that won another FP7 grant [**vi**]. This work opened the door to automated error analysis for MT systems by combining corpus search methods with established systems used to measure the quality of output, such as BLEU [**5, 6**].

3. References to the research

Publications:

1) Serge Sharoff, Bogdan Babych, and Anthony Hartley (2006). "Using comparable corpora to solve problems difficult for human translators." In *Proc. of International Conference on Computational Linguistics and Association of Computational Linguistics*, *COLING-ACL* 2006, pp. 739–746, Sydney. ACL is the most prestigious international conference in the area of computational linguistics. The ASSIST project was rated by EPSRC reviewers as "tending to internationally leading" in terms of scientific and social impact. (16 citations) †
2) Serge Sharoff, (2006) "Open-source corpora: using the net to fish for linguistic data." In *International Journal of Corpus Linguistics* 11(4), pp. 435-462. (54 citations) †

3) Serge Sharoff (2007). "Classifying Web corpora into domain and genre using automatic feature identification." In *Proc. of the 3rd Web as Corpus Workshop*, pp. 83-94. (23 citations)

4) Serge Sharoff (2010). "In the garden and in the jungle: Comparing genres in the BNC and Internet." In Alexander Mehler, Serge Sharoff, and Marina Santini, eds, *Genres on the Web: Computational Models and Empirical Studies*, Springer, Berlin/New York. (22 citations) ‡

5) Babych, B. & Hartley (2009), A. "Automated error analysis for multiword expressions: using BLEU-type scores for automatic discovery of potential translation errors." *Linguistica Antverpiensia*: *Journal of translation and interpreting studies. Special Issue on Evaluation of Translation Technology*, 8, pp. 81-104. ‡

6) Babych, B. & Hartley, A. (2004). "Extending the BLEU MT evaluation method with frequency weightings". *Proc. of ACL 2004*, Barcelona. (62 citations) †

Evidence of quality:

In the rapidly developing field of computational linguistics, publications at leading international conferences are rated at the same level as journal publications.
The citation counts are provided by Google Scholar.
† Submitted to RAE 2008 and available on request.
‡ Submission to REF2014.

Grants:

i) ASSIST: Automatic semantic assistance for translators (EPSRC, PI: Hartley, CoI: Sharoff, 2005-2007, 30 months, £240,000 for Leeds)

ii) Translation Strategies in Comparable Corpora: (Leverhulme Early Career Fellowship, Babych, 2007-2009, 24 months, £40,000)

iii) WebDoc: Document Classification on the Web (Google Research Award, PI: Sharoff, 2009-2010, 12 months, £40,000, an unrestricted gift for further research into Web genres)

iv) Intellitext: Intelligent Access to Multilingual Document Collections (AHRC, PI: Hartley, CoI: Sharoff, 2010-2011, 12 months, £160,000)

v) TTC: Translation, Terminology and Corpora (EU FP7 grant TTC, PI: Sharoff, 2010-2012, 36 months, €369,000 for Leeds)

vi) ACCURAT: Analysis and evaluation of comparable corpora for under-resourced areas of machine translation (EU FP7, PI: Babych, 2010-2012, 30 months, €340,000 for Leeds).

vii) HyghTra: High Quality Hybrid Translation System (FP7 Marie Curie IAPP, Co-ordinator and PI Babych, 2010-2013, 48 months, £500,000 for Leeds)

4. Details of the impact

Babych's and Sharoff's success in attracting industrial funding and the impact on the products, services and work-flow of industrial and organisational partners is based on their unique combination of expertise and their shared interest in making rapidly developing natural language processing technologies accessible to a range of translation communities, thus initiating wider research and development. Their collaborative research has opened up new business opportunities and models for non-academic partners, and improved efficiency and consistency across a number of translation platforms and contexts:

Corporations: ABBYY Corp and Google Inc:
Collaboration with ABBYY Corp was based on Sharoff's work on computational lexicography [**1,3,4**] (on which he presented to an audience of 50 translation professionals in July 2011 [**A**]), and the CTS team's work on MT technologies (**5,6**). A statement from ABBYY's Director of Linguistics Research treats Sharoff's "discovery of variation in linguistic data ... coming from topics, genres or social stratification" [**3, 4**] as contributing to their "better understanding of the Web data" and helpful in their work in computational lexicography and MT [**B**]. Babych and Hartley's work on automated analysis of MT quality [**5,6**] led to invitation from ABBYY for an expert assessment of their MT technologies, which gave the corporation an "awareness and confidence in the prospects of these technologies and was used during the audit of... ABBYYMT and Semantic Analysis projects audit by governmental and commercial organisations" [**B**].

The ASSIST research into automatic Web document classification [**3**], particularly the development of ideas around "Web page classification" prompted an expression of interest from Google Inc. in 2008 [**C**]. Sharoff subsequently received a Google Research Award [**iii**], and was invited to present the CTS team's research to a group of 15 senior staff members at Google headquarters in Zurich in December 2008. Google has since integrated document classification functionality into its searches, improving effectiveness and user experience.

Small business: Lingenio GmbH:
As part of the FP7 Marie Curie IAPP (Industry-Academia Partnership and Pathways) project HyghTra [**vii**], CTS worked together with a German translation company Lingenio GmbH to produce a technology for rapid development of hybrid MT systems. Using statistical techniques to automatically build linguistic resources, Babych's and Sharoff's contribution allowed Lingenio to harvest and annotate large collections of electronic texts, extract translation equivalents, create lists of related terms across languages, and evaluate MT systems. Using this technology, the company was able to re-design and streamline their system development cycle and create MT products for a number of new languages, including Dutch, Spanish, French, Russian and Ukrainian (a range previously unattainable due to financial cost). In addition, a new development framework for MT systems has enabled Lingenio to add languages and translation directions within shorter time frames. Finally, the developer-oriented environment created for the project has proved to be effective as a tool for MT customisation, while the support of authoring documents in a non-native language has enabled individuals to increase their productivity [**D**]. A 5-day workshop involving 11 translators was jointly organised by CTS and Lingenio in April 2013, in order to share best practice and gather feedback from users.

International consortium: TAUS:
The technologies developed for the ASSIST project have been shown to support translators in discovering appropriate translation equivalents, to help developers of MT systems encode rules for linguistic analysis and translation, and to improve the accuracy of modern search engines by automatically classifying documents on the internet. After reading Sharoff's work [**1**], the Director of TAUS, an industrial consortium whose members include major translation service providers and users, contacted CTS directly. This resulted in the development of a

free on-line Linguistic Search Engine for translators (https://www.tausdata.org/), containing the world's largest collection of technical translations in over 40 languages. This entailed the introduction by TAUS of a new business model in which providers pool their collections in order to maximise mutual benefit. As the Director of TAUS confirms, "we had our own ideas and vision about data sharing, but what CTS team contributed, changed how we made that into services" [**E**]. The TAUS Linguistic Search Engine has since been integrated into the popular memoQ translation memory system and currently receives some 470,000 searches each month [**F**]. As a result, translators now import a much larger collection of documents, ensuring more consistent and terminologically appropriate translations. Uptake of TAUS membership by large industrial and non-commercial partners such as Adobe, Cisco, Google, Microsoft, Intel, DELL, Oracle, Philips, European Patent Office, etc., is further testament to the viability of this new business model and translation service.

Individual translators and governmental agencies:

After working with CTS, the Director of Lexical Computing (a provider of computational solutions in lexicography for several major dictionary publishers), remarked that the collaboration "improv[ed] our understanding of the nature of data coming from the Web," thereby enabling the successful conversion of webpages into useful resources for translators [**G**]. Sharoff's research (specifically **2, 3** and **4**) has underpinned his co-edited *A Frequency Dictionary of Russian* (Routledge, 2013), an invaluable tool for both learners and teachers, providing a list of the 5,000 most frequently used words 300 multiword constructions.

Moreover, the CTS collection of resources from the Web in 15 languages is extensively used by researchers and translators on a global scale. The following table shows the number of corpus queries made by users from outside the University of Leeds in 2012: [**H**]

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
15948	44567	20783	39052	12514	10519	8745	8713	9143	14913	22789	19687

As part of International Annual Meeting on Language Arrangements, Documentation and Publications (IAMLADP), CTS regularly organises one-week workshops to offer training in using state-of-the-art translation technologies. In these workshops, translators from large international organisations (i.e. UN, European Commission, European Parliament) learn about emerging technologies and the possibilities of integrating the latest developments in terminology extraction and MT into their professional workflow, thus enhancing productivity. The Head of the Language and Technology Support Section of the Translation Centre for the Bodies of the European Union commented that the research presented in the workshops was "very relevant for our activities and could lead to improvements in terms of efficacy and efficiency" [**I**]. Positive feedback from participants [**I**] and continuing demand for new sessions has led to the CTS workshops becoming a permanent fixture of the IAMLADP programme.

5. Sources to corroborate the impact

A) A video of Sharoff's presentation (in Russian) is available online: http://vimeo.com/27433273; and as a download from the company website: http://www.abbyy.ru/science/seminars/archive/ [accessed 30 October 2013].

B) Email testimony, Director of Linguistic Research at ABBYY Corp, *available on request.* Email of 21 October 2013].

C) Email to Sharoff from Google Inc, *available on request.* Email of 30.07.2008

D) Email testimony from Owner of Lingenio GmbH, *available on request.* Email of 17.04.2013.

E) Testimony from Director of TAUS, *available on request.* 09.07.2013

F) Statistics provided by Chief Technology Officer at Spartan Software Inc, *available on request*. Email of 23.04.2013.

G) Email testimony from Director of Lexical Computing, *available on request.* Email of 21.10.2013.

H) Query log of the corpus server (http://corpus.leeds.ac.uk), *available on request.*

I) Feedback from questionnaires from 2013 IAMLADP sessions held in Leeds (http://www.iamladp.org/about.htm), *available on request.*