LANGUAGE IN THE AGE OF DATAISM

Špela VINTAR
University of Ljubljana, Faculty of Arts, Ljubljana

Vintar, Š. (2019). Language in the Age of Dataism. Slovenščina 2.0, 7 (1): 126–143.

DOI: http://dx.doi.org/10.4312/slo2.0.2019.1.126-143.

The digital age brings dramatic changes to language and communication; its effects can be seen in the ways we use language, the channels we use to communicate and the manners in which ideas are spread. From the other end of the spectrum, our linguistic behaviour, communications and knowledge are transformed into data which can be used or bought to feed intelligent technologies. The article presents a bird's eye view of these dynamics of change, first by focusing on the impact of digitisation on language itself, further by analysing current trends in the language industry where traditional services are being replaced by technology- and data-driven solutions, and finally by exploring the impact of these technologies on people and society at large. We make a case for digital linguistics as an interdisciplinary field of study which adopts a human-centred approach to the sociolinguistic, technological, economic, infrastructural and ethical issues emerging with regard to language in the digital age.

Key words: digitisation, language change, language industry, digital linguistics

1 INTRODUCTION

For some time now, the effects of digitisation on humanity no longer inspire just awe in the face of technological advances but increasingly raise concerns. In less than thirty years of its existence the internet has evolved from a medium charged with tremendous potential for freedom of communication, thought and global cooperation into its shadowy reverse – an environment which has become indispensable but obscured by infringements of privacy, security, dignity, intellectual property rights and competition laws. As Pasquale observes in *The Black Box Society*, the "democratization" promised by Web 2.0 had "a

different – even an opposite effect. The very power that brought clarity and cooperation to the chaotic online world also spawns murketing, unfair competition, and kaleidoscopic distortions of reality" (2015, p. 98).

With rapid advances in Artificial Intelligence, similar concerns are arising in view of the many scenarios where machine learning algorithms are already replacing human decision-making. The fundamental questions are not whether certain jobs will disappear, in which work environments will humans be replaced with robots and when this is likely to happen for most fields of human endeavour. A more complex set of questions refers to issues such as AI bias ("Is the machine fair?") and the moral status of AI ("Is it good or evil?"). In the *Cambridge Handbook of Artificial Intelligence* Bostrom and Yudkowsky lay the foundations for an ethics of AI, acknowledging that "[t]he term 'Artificial Intelligence' refers to a vast design space, presumably much larger than the space of human minds (since all humans share a common brain architecture)" (2014, p. 332) and that certain criteria which apply to humans performing social functions must also be considered in an algorithm intended to replace human judgement: responsibility, transparency, auditability, incorruptibility, predictability (2014, p. 318).

The beginnings of the age of Big Data celebrated a technological milestone: a point in time when the computational and storage capacity on the one hand and the availability of digital data on the other would no longer present a bottleneck for development. But consider the difference between data collected as sample of human activity in order to build better models, and *collective* data gathered through recording *all* human activity in order to be used, sold and resold by techno-giants and governments alike – this transition marks the beginnings of dataism, which, by Harari's definition (2016, p. 351), declares that "the universe consists of data flows, and the value of any phenomenon or entity is determined by its contribution to data processing."

It is against this background that we reflect on language in the digital age, whereby our focus shifts from language as a communicative device, language as an economic or business activity to language used as data. It will become clear that from all of these three aspects language has undergone profound changes under the influence of technology, and some of these changes may clearly be regarded as positive. In fact, while popular media will have us be-

lieve that the future of *everything* is rather bleak, language in the digital age is, in many respects, thriving.

2 DIGITISATION AND LANGUAGE CHANGE

Languages change over time, and the factors involved in this process range from social, political, technological and economic influences to interventions by normative bodies. It is therefore only to be expected that digitisation and the emergence of numerous new channels of communication would have an impact on language use, and this is often reflected in news articles with titles such as "Is the Innanet RUINING teh English Language??? ¯\(°_o)/¯"[1] or "L3t's t@lk internet"[2]. Linguists have been alert to this topic since the early days of texting (Crystal, 2008), and the expansion and diversification of digital media has given rise to numerous studies exploring their effects on language as a whole or on the use of written language by youth (Baron, 2008; Lenhart, 2008; Thurlow, 2007; Crystal, 2011). In his comprehensive and detailed review of the field of computer-mediated communication (CMC), Androutsopoulos (2011) provides a sociolinguistic set of conditions which shape "digital networked writing", defining it as "vernacular", "interpersonal and relationship-focused", "unplanned and spontaneous" and "dialogical and interaction-oriented". In a critical synthesis of research studies spanning over three decades, Androutsopoulos demonstrates that much of the language change ascribed to digital media is restricted to lexis, with notorious lists of CMC-typical acronyms and other lexical innovations from the field of technology. The effects of the internet on spoken language seem to be negligible, but the productivity of neologisms derived from social media seems boundless across (written) genres and in languages other than English.

As for netspeak ruining school writing and negatively influencing literacy, evidence is less conclusive, and it is clear that such studies are methodologically difficult to conduct. Lenhart (2008) reports on a large scale study of the attitudes and habits of US teens comparing their out-of-school written communication and school writing, and the prevailing opinion of teens was

1 https://gizmodo.com/is-the-innanet-ruining-teh-english-language-_o-1680686542, 22. 1. 2015

2 https://www.deccanherald.com/sunday-herald/sh-top-stories/l3ts-tlk-internet-668377.html, 6. 5. 2018

that texting and communicating via digital media was *not* writing, and that electronic communication had little or no impact on their written production at school. Similarly, Androutsopoulos (2011) mentions an empirical study by Dürscheid and Wagner (2010) carried out in German-speaking Swiss schools, where results suggest that out-of-school digital writing does not visibly influence institutional language production.

This is not to say that the entire landscape of language use has not dramatically changed, mainly through the emergence of new digital genres, and an "unprecedented scale of publicness" that tweets, blogs, posts, news comments and user reviews can achieve. The internet is a mixture of editorial, professionally-crafted content intertwined with vernacular, spontaneous, informal texts; a "manifestation of the intermingling of the private and the public that characterises late modernity" (Androutsopoulos, 2011, p. 10). We might add that the private/public is only one of the dimensions along which internet discourse is intermingled, other candidate variables being standard/non-standard, true/fake, predominantly textual/predominantly visual, monolingual/multilingual, human-written/machine-written, and many more.

In recent years, a number of language resources, tools and methods have been developed which allow researchers to ask not just whether internet language is different, but *how* different it is. Such studies attempt to quantify the degree to which a certain language variety deviates from standard language, whereby basic corpus pre-processing steps such as lemmatization and PoS-tagging need to be fundamentally adapted or even developed anew to accommodate the transformations and innovations found across genres of the web. In an interesting study of tweets in three closely related languages of former Yugoslavia, Serbian, Croatian and Slovene, Miličević et al. (2017) performed a thorough investigation of spelling transformations and report on a number of similarities and differences. In all three languages, frequent transformations include the omission of diacritics, repetition of certain vowels for emphasis and omission or transformation of word-final vowels or suffixes. In general, the transformation frequency is highest in Slovene (17%) and lowest in Serbian (10%), with Croatian in the middle (13%), and if the omissions of diacritics are not counted Slovene drops to 15% and Serbian to just over 3% of transformed tokens. This difference is significant – it means that in an average

Slovene tweet between 4 and 5 words will be spelled in a non-standard way, while in Serbian only one or none. It would appear, at least for these three languages, that the tendency of a language towards the use of non-standard forms correlates with the level of digital maturity of its country,[3] which is an unexpected finding.

On the other hand, the authors of the study observe that transformations in Serbian, while lower in frequency, occur at more varied positions and indicate a more playful and creative use of language than Slovene or Croatian. On the whole, twitterese and other types of internet discourse mirror layers and layers of social, cultural, political, economic and historical circumstances, and therefore any study of computer-mediated communication limiting itself to just linguistic features necessarily remains incomplete. More importantly, in the same way that virtual communities are communities with their own sociological features, cyberlanguage is a language form in its own right whose properties cannot be described in terms of deviation or transformation from its standard or spoken relatives.

Digitisation affects language beyond the scope of netspeak and genres predominantly residing on the internet. Today, texts are created with the aid of AI technologies and although these are trained on large samples of human language, neural networks may have given rise to a new set of dialects. We are referring mostly to machine translation and the various levels of post-editing applied before such texts are made public. As shown by recent surveys of the language industry which we present in more detail in the next section, the use of MT is growing in all strands of professional translation, but few studies have systematically analysed the properties of post-edited texts. A recent paper by Toral (2019) fills this gap by addressing the question whether human translation and post-edited machine translation differ significantly in terms of several quantifyable features: lexical variety, lexical density, length ratio and part-of-speech sequences.

The underlying intuition is that translations produced by humans from machine-translated drafts must be somehow different from translations produced by humans from scratch, and Toral performs a number of experiments

3 For Slovenia and Croatia, see the Digital Economy and Society Index 2019, for Serbia the I-DESI 2018.

across six language pairs verifying the existence of *post-editese*. As his results show, post-edited texts have lower lexical variety and density than human translations, and their sentence length and PoS sequences are closer to the source than the target language. This is in line with the so-called "translation universals", the properties of translations which appear across language pairs and include phenomena such as normalization, shining-through and source language interference (Baker, 1993; Mauranen & Kujamäki, 2004). Toral's experiment thus proves two important things: firstly, that MT has a lower percentage of content words than HT and is therefore lexically simpler, and secondly that humans striving to improve on MT and create a human-like translation fail to do so, at least as far as lexical variety and PoS sequences are concerned. The author concludes with a cautionary note that "the extensive use of PE rather than HT may have serious implications for the target language in the long term, for example that it becomes impoverished" (2019, p. 279). It remains interesting though that – as Toral himself and several other authors point out (Green, 2013; Bowker & Buitrago Ciro, 2015) – humans do not necessarily perceive HT as better or more acceptable than PE. A recent study by Screen (2019) compares the quality of human and post-edited translations from the end-user perspective. The experiment uses both eye-tracking and end-user assessments of readability and comprehensibility, and the results show no statistically significant difference or inferiority of post-edited texts.

3 AI AND THE LANGUAGE INDUSTRY

In the previous section we briefly discussed some instances where digitisation has an impact on language itself, both within and beyond the scope of internet communication. We now turn our attention to the economic sector of language-related services generally referred to as the language industry, which traditionally revolved around translation and interpreting but is increasingly diversified and, as we shall see, datafied. The importance of aggregating translation data became apparent with the emergence of Translation Memory tools, commonly known as Computer-Aided Translation or simply CAT tools from the early 1990s. With growing needs for fast translation and localization in the globalized world the idea that past translation projects should be stored in bilingual segments and recycled in order to boost productivity seemed per-

fectly logical. However, the reactions of translators to CAT tools were reserved at best, with much opposition to the notion that translation work could be conceived as being repetitive and recyclable.

As with most novelties, the technology gradually became mainstream and is considered indispensable today – according to the latest 2018 Language Industry Survey (LIS 2019, p. 17) less than 1% of language services companies report that they are not using CAT tools. An interesting historical trivium is that as early as 1997 Trados Translator's Workbench, the predecessor of today's market-leading SDL Trados Studio product suite, boasted the use of neural networks for their fuzzy matching algorithms, thus anticipating the AI era in translation technologies.

The development of statistical MT engines and their growing accessibility brought about another shift, namely that of MT becoming a pre-processing step in professional translation, thereby generating the demand for post-editing. Despite the fact that numerous studies have demonstrated significant productivity gains even with early SMTs (O'Brian, 2007; Guerberof, 2008) the sentiments of practicing translators towards PE remain mixed to this day, as a recent survey by the American Translators Association shows (Zetzsche, 2019). The sentiment however is not shared by language service providers. According to the results of the Language Industry Survey for 2016, 2017 and 2018, the use of MT is growing steadily both by companies and individuals. The latest survey, which is considered representative for Europe but not the rest of the world, states that the number of companies and individuals who are not using MT at all has dropped to 31% and 38%, respectively (LIS, 2016, 2017, 2018).

With the arrival of neural Machine Translation (NMT), the language industry was transported into the age of AI. Even if several respondents of the aforementioned ATA survey on "(Why) Do you use MT?" answered "To get a good laugh", numerous studies have been performed to prove that NMT systems generally outperform SMT models by two or more BLEU points (Bentivogli et al., 2016; Way 2018), whereby several authors warn that BLEU may be under-reporting the difference in quality. According to error analyses, NMT produces fewer morphological errors (–19%), lexical errors (–17%), and substantially fewer word order errors (–50%) than its closest statistical competitor, and on average requires about a quarter fewer edits compared to the best

phrase-based SMT (Way, 2018).

It is thus not surprising that the report issued after the annual TAUS Global Content Conference (TAUS, 2019), an event which attracted 130 world's largest players in translation and localization, begins with a chapter titled The Quantum Leap and proclaims that "the NMT revolution of the last few years has pretty much wiped out all previous technologies. In addition to this, MT post-editing (PE) has become mainstream, currently the most widely used set-up is MT in conjunction with some degree of human PE." The size of the Machine Translation market was estimated at 433 million USD in 2016[4] and was expected to grow at an annual rate of 19%. Google Translate's *daily* throughput exceeds the volume that all translators in the world translate in a year.

According to some estimates, MT is expected to reach the point of *human parity* by 2029, but on the other hand the language industry has voiced several concerns regarding the use of NMT in business solutions. The first has to do with the robustness of NMT when dealing with different types of content and different domains. This clearly presents a challenge for language service providers, as varying levels of MT quality may have an impact on productivity, return-on-investment and the payment schemes used for PE. A second challenge is the sentence-based mode of processing for most NMT systems which may result in incoherent and inconsistent translations. Research is being conducted on paragraph- or document-level NMT which would allow systems to translate content, not isolated sentences (TAUS 2019, p. 5–6).

Comparing reports about the language industry from Europe, such as the LIS (2016, 2017 and 2018), and those from more globally oriented organisations such as TAUS (Massardo et al., 2016; Keynotes Summer, 2019) or GALA[5], it appears that the global or US-based view of the language industry anticipates more dramatic changes driven by technology and envisages translation as a utility available to everyone, everywhere and on every device. All reviewed studies however agree in forecasting a rapidly growing demand for translations and other language services, in fact, these demands even today quite significantly surpass the capacities of human language service providers.

4 Global Market Insights: Machine Translation Market Size, https://www.gminsights.com/industry-analysis/machine-translation-market-size

5 http://www.gala-global.org

One obvious consequence of this fact is that the majority of translations reach their audiences as raw MT, and that even in professional translation varying levels of quality are required. Both of these facts are hard to digest for a typical professional translator who was trained to strive towards a single and universal highest quality standard, and the position of most translator training institutes regarding quality remains unchanged.

4 THE DATAFICATION OF TRANSLATION

There is another important trend we can discern from the reports, and it concerns data. Translation memories and bilingual corpora have been considered important assets for some time now, and issues related to ownership, data protection and intellectual property rights have been a hot topic of debate for over a decade (Smith, 2009). The Language Industry Survey for 2017 (LIS, 2018) introduced for the first time a question about the transfer of user rights or ownership to the client, and responses indicated that approximately half of the respondents would never transfer those rights, while the other half would do so sometimes. The results for 2018 reveal a strong trend towards this transfer, and a breakdown of responses by company size shows that for larger companies the transfer of user rights or ownership is now almost mandatory. Large companies work for large clients, and these adhere to the dataism motto that data is the new fuel.

Another TAUS publication titled The Translation Industry in 2022 (Massardo & van der Meer, 2017) identifies Data as one of the six drivers of change and contains a valuable explanation of the difference between language data and translation data. While the former consists of translation memories, corpora, lexicographical and terminological collections, the latter is essentially metadata (2017, p. 18):

> *Translation data is typically metadata: data about translation that can be harvested downstream the closure of a translation project/job/task, such as content type, language pair(s), domain, subject, number of characters/words/lines, quote/price, scheduled time, time spent, technologies used, translation stats (e.g. source - translation memory match, automatically propagated, machine translated - edited, approved) date and time of last saving, etc.*

> *The analysis of translation data can provide a very valuable insight into the translation processes to find the best resource for a job, to decide what to translate and which technology to use for which content.*

Eavesdropping on the debates amongst the tech giants such as Amazon, Apple, Google, Microsoft, Adobe, and the largest LSPs such as Lionbridge, SDL and TransPerfect, the power of data and the central role of AI remain recurring topics. Language data markets have been established, but a lot of data collection goes on backstage using home-grown solutions. Machine translation is but the most obvious application fuelled by data; there is much demand for other intelligent services such as speech processing, user profiling, sentiment analysis, question answering, social network analytics, and there is a pronounced trend towards machine learning for a better management of multilingual workflows.

In view of these developments it would appear that language as a business, not unlike other technology-driven businesses, is under threat of monopolisation by the big players who simultaneously own the bulk of the data, develop the smartest technologies and increasingly own research infrastructures significantly more powerful than those provided by the academia or public research funding.

5 A CASE FOR DIGITAL LINGUISTICS

We have examined some of the challenges that language is facing in the digital age; it is now time to reflect on the possible measures to be taken by researchers, academia, practitioners and policy makers in order not to be reduced to mere instruments of change but assume an active role, and possibly direct the course of development into one which is fairer and more inclusive for all members of society.

The advances that Artificial Intelligence is enabling in natural language processing are truly impressive, and scientific progress is accelerated by the enormous amount of private funding flowing into research and by e.g. Google's policy[6] to openly share some of its AI tools with the community, thus enhancing competition. Clearly though, it will be increasingly hard for researchers to keep up with the speed of discoveries produced by the techno-giants.

6 https://ai.google/tools/

It is important to remember that the role of science in these – or any other – times is not to blindly compete in the race towards singularity, but to provide critical insights, analyse impact, advocate responsibility, and safeguard the ethical principles fundamental to our society. With regard to the ethics of AI, strong initiatives are underway within leading research institutes, such as the Future of Humanity Institute[7], the IEEE[8] or the Foundation for Responsible Robotics[9], and the European Commission has recently passed a communication titled Building Trust in Human Centric Artificial Intelligence, which defines AI "not as an end in itself, but a tool that has to serve people with the ultimate aim of increasing human well-being" (EC COM, 2019, p. 168).

Returning to language and AI, ethical concerns regarding the use of human translations to train MT systems have been raised by Kenny (2011), especially because "the role of translators in creating vital data has been mostly downplayed or ignored" by MT developers. She also touches upon another interesting ethical question, namely the (im)possibility of computers communicating like humans. According to Melby and Warner (1995) and Kenny (ibid.), "in order to communicate with others, we must have agency, which involves the capacity to make real choices for which we take responsibility, and we must also regard our interlocutors as having agency. […] Without agency, we are reduced to the status of machines and there is no dynamic general language." It is needless to point out that from today's perspective, with chatbots and automated dialogue systems lurking around every corner of the internet, the ethics of communication seems a considerably more complex issue.

A more recent contribution to the discussion about language resources and the ethics of their reuse was made by Moorkens et al. (2016) who systematically describe the practices prevalent in the language industry regarding data ownership, the "disempowered" translator in precarious working positions and the legal situation "in which laws of copyright are effectively bypassed in content collection, curation, and exploitation, [and which] permits resource holders to retain data at a cost to disempowered human writers and translators". The authors' recommendations for translators include collective bar-

7 https://www.fhi.ox.ac.uk/

8 https://ethicsinaction.ieee.org/

9 https://responsiblerobotics.org/

gaining, informing themselves about their legal rights and using TM metadata more effectively in order to explicitly assign usage rights to their assets.

Establishing fair practices for data sharing and a transparent regulative system for its collection and processing is just one of the challenges we need to face up to, and the present situation gives little grounds for optimism. As Pasquale writes, "top legal scholars have already analogized the power relationships in virtual worlds and cloud computing to medieval feudalism" (2015, p. 218). Considering all the other profound changes that language and communication are undergoing in the digital society, some of which we have discussed above, it becomes clear that to understand and adequately describe these phenomena an interdisciplinary approach is required, and that linguistics alone, even with all its applied subfields, lacks the methodological inventory to approach this task. Analysing large communication networks, proposing new workflows of content creation, developing intelligent knowledge solutions or modelling emotions, to name but a few non-futuristic scenarios, all require a combination and integration of knowledge from different domains.

If solutions for the processing of natural language were traditionally developed by computational linguists, we are now entering an era in which AI technologies are becoming mainstream in many areas of everyday life, and we may well imagine the not-so-distant future when these now separate intelligences begin interacting to solve complex problems, much like intelligent humans do. As we have demonstrated before, any intelligent technology imposed on the human society has a social dimension in that it modifies the social practices that were in place before, and it may also have ethical, legal, psychological and other dimensions.

We thus propose the term digital linguistics to designate a human-centred approach to digitally-driven language and communication as well as the study thereof, utilizing methodologies and theoretical backgrounds from a range of "feeder" disciplines: linguistics, including computational, corpus, cognitive, socio- and psycholinguistics; computer and information science, including machine learning, data mining, knowledge modelling and AI; social sciences, including law, journalism, communication and media studies; and the relevant humanities, in particular ethics, psychology and philosophy. The list is not exhaustive and serves primarily to emphasize the interdisciplinary nature of digital linguistics.

We further believe it is paramount that universities and other higher education institutions respond not only to the skills gap reported by employers, but more importantly to the expectations and concerns of the civil society which already feels insecure in the "feudalism" of digital communication channels. One attempt to bridge this education gap is the joint master's degree in Digital Linguistics in preparation by a consortium of three universities, Ljubljana, Zagreb and Brno, expected to launch in 2021/2022. The model curriculum was developed within the recently concluded DigiLing[10] project and is based on the findings of a trans-European survey of language-related needs amongst employers.[11]

6 CONCLUSIONS

It seems that digitisation affects language in ways different from what the average person or even linguist might expect. The examples selected for discussion above show that the language of internet communications develops under its own rules, not dissimilar to other language varieties known from pre-internet times. Contrary to urban myth, teenagers do know how to draw the boundary between formal and informal writing, while adults or even language professionals have a hard time distinguishing between human and post-edited translations and do not have a clear preference for either. Machine-translated and post-edited texts are found increasingly acceptable by end-users despite the fact that they exhibit pronounced features of the source language.

Word embeddings and neural networks allow us to discern semantic change (Hamilton et al., 2016) or translate between languages for which no parallel data exist (Johnson et al., 2017), but at the same time language professionals feel disempowered as their intellectual property rights are being ignored in the global data collection frenzy. In this article we attempted to present a selection of recent trends involving language and communication in the digital age, and their implications may range from fantastic to catastrophic, depending on one's point of view. A concluding thought might be that as academics and researchers we should strive towards objectivity and realism in the face of the complex challenges, but also towards a responsible stance and a keen interest in the dynamics of change, the only constant of our times.

10 https://www.digiling.eu

11 https://www.digiling.eu/deliverables

ACKNOWLEDGMENT

This article was presented as keynote lecture at the INFuture 2019: Knowledge in the Digital Age conference and is republished here with the permission of the INFuture programme board.

REFERENCES

Androutsopoulos, J. (2011). Language change and digital media: a review of conceptions and evidence. In N. Coupland & T. Kristiansen (Eds.), *Standard languages and language standards in a changing Europe* (pp. 145–161). Oslo: Novus Press.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). Amsterdam: John Benjamins.

Baron, Naomi S. (2008). *Always on: Language in an Online and Mobile World*. Oxford: Oxford University Press.

Building Trust in Human Centric Artificial Intelligence. EC COM(2019) 168. European Commission, 8. 4. 2019.

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. In *Proceedings of EMNLP 2016*.

Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. In W. Ramsey & K. Frankish (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316–334). Cambridge: Cambridge University Press.

Bowker, L. & Buitrago Ciro, J. (2015). Investigating the usefulness of machine translation for newcomers at the public library. *Translation and Interpreting Studies*, *10*(2), 165–186.

Hamilton, W.L., Leskovec, J. & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the Association for Computational Linguistics (ACL), Berlin.

Crystal, D. (2008). *Txtng: The gr8 db8*. Oxford: Oxford University Press.

Crystal, D. (2011). *Internet Linguistics*. London: Routledge.

Dürscheid, C., Wagner, F. & Brommer, S. (2010). *Wie Jugendliche schreiben: Schreibkompetenz und neue Medien*. Berlin & New York: de Gruyter.

Green, S., Heer, J. & Manning, C. D. (2013). The efficacy of human post-editing for language translation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 439–448).

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F. & Eisenstein, J. (2016). The social dynamics of language change in online networks. In *International Conference on Social Informatics* (pp. 41–57). Cham: Springer.

Guerberof, A. (2009). Productivity and quality in MT post-editing. In *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*. August 29, 2009. Ottawa, ON, 8ff.

Harari, Y. N. (2015). *Homo Deus: a Brief History of Tomorrow*. London: Harvill Secker.

Johnson, M., Schuster, M., Quoc V. Le, Krikun, M., Wu, Y., Chen, Z. et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* 5, 339–351.

Kenny, D. (2011). The ethics of machine translation. In *Proceedings of the New Zealand Society of Translators and Interpreters Annual Conference* 2011. Auckland, New Zealand.

Lenhart, A. (2008). Writing, technology, and teens. Washington, DC: Pew Internet and American Life Project. Retrieved from http://pewresearch.org/pubs/808/writing-technology-and-teens

[LIS 2016] 2016 Language Industry Survey – Expectations and Concerns of the European Language Industry. Retrieved from https://www.euatc.org/industry-surveys/item/download/5_57a02b9c45602ea9f7daf4440a7b2979

[LIS2017] 2017 Language Industry Survey – Expectations and Concerns of the European Language Industry. Retrieved from https://ec.europa.eu/info/sites/info/files/2017_language_industry_survey_report_en.pdf

[LIS2018] 2018 Language Industry Survey – Expectations and Concerns of the European Language Industry. Retrieved from https://ec.europa.eu/info/sites/info/files/ 2018_language_industry_survey_report_en.pdf

Massardo, I.; van der Meer, J. (2017). *The translation industry in 2022*. TAUS BV, De Rijp, The Netherlands.

Massardo, I., van der Meer, J., Khalilov, M. (2016). TAUS Translation Technology Landscape Report. September 2016, TAUS BV, De Rijp, The Netherlands.

Mauranen, A., Kujamäki, P. (2004, Eds.). *Translation universals: do they exist?* Amsterdam: John Benjamins.

Melby, A. K., Warner, C. T. (1995). *The possibility of language: a discussion of the nature of language, with implications for human and machine translation*. Amsterdam & Philadelphia: John Benjamins Publishing.

Miličević, M., Ljubešić, N. & Fišer, D. (2017). Birds of a feather don't quite tweet together. In D. Fišer & M. Beißwenger (Eds.), *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World* (pp. 14–43). Ljubljana: Faculty of Arts.

Moorkens, J., Lewis, D., Reijers, W., Vanmassenhove, E. & Way, A. (2016). Translation resources and translator disempowerment. In *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

O'Brien, S. (2007). An Empirical Investigation of Temporal and Technical Post-Editing Effort. *Translation and Interpreting Studies*, *2*(1), 83–136.

Pasquale, F. (2015) *The black box society*. Cambridge, MA: Harvard University Press.

Screen, B. (2019). What effect does post-editing have on the translation product from an end-user's perspective?. *Journal of specialised translation*, *31*, 133–157.

Smith, R. (2009). Copyright issues in translation memory ownership. *ASLIB Translating and the Computer* 31.

TAUS Keynotes Summer 2019. A Review of the TAUS Global Content Conference in Salt Lake City, UT (USA). TAUS Signature Editions, Amsterdam. Retrieved from www.taus.net

Toral, A. (2019). Post-editese: an exacerbated translationese. *Proceedings of the Machine Translation Summit XVII Volume 1: Research track* (pp. 273–281). Dublin, Ireland: EAMT.

Thurlow, C. (2007). Fabricating youth: new-media discourse and the technologization of young people. S. Johnson & A. Ensslin (Eds.), *Language in the Media* (pp. 213–233). London: Continuum.

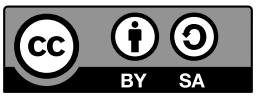
Way A. (2018). Quality Expectations of Machine Translation. In J. Moorkens, S. Castilho, F. Gaspari, S. Doherty (Eds.)*, Translation Quality Assessment: From Principles to Practice* (pp. 159–178). Cham: Springer. doi: https://doi.org/10.1007/978-3-319-91241-7_8

Zetzschke, J. (2019). (How) Do You Use MT? *The Tool Box Journal*, 19-11-306(2019). Retrieved from http://www.internationalwriters.com/toolkit/current.html#LETTER.BLOCK5

JEZIK V ČASU DATAIZMA

Digitalizacija vnaša korenite spremembe v jezik in komunikacijo, saj vpliva na načine izražanja, sredstva komunikacije in poti, prek katerih se širijo nove ideje. Obenem živimo v času, ko se naše jezikovno vedenje, naša sporočila in znanje skrbno beležijo, ti podatki pa se uporabljajo in prodajajo za urjenje pametnih tehnologij. V prispevku skušamo zaobjeti dinamiko teh sprememb s širše perspektive, in sicer se najprej osredotočimo na vpliv digitalizacije na sam jezik, nato analiziramo sodobne težnje v jezikovni industriji, kjer opažamo, da mnoge tradicionalne jezikovne storitve nadomeščajo tehnološko podprte in na podatkih temelječe rešitve, nazadnje pa obravnavamo vpliv tehnologij na človeka in družbo kot celoto. Iz širšega okvirja razprave izpeljemo utemeljitev in opredelitev področja digitalnega jezikoslovja kot interdisciplinarne vede, ki se z jezikom v digitalni dobi ukvarja s humanistično-družboslovnega izhodišča in vanj vključuje jezikoslovne, tehnološke, družbenoekonomske, infrastrukturne, kognitivne, etične in pravne vidike.

Ključne besede: digitalizacija, jezikovne spremembe, jezikovna industrija, digitalno jezikoslovje



To delo je ponujeno pod licenco Creative Commons: Priznanje avtorstva-Deljenje pod enakimi pogoji 4.0 Mednarodna. / This work is licensed under the Creative Commons Attribution-Share-Alike 4.0 International.

https://creativecommons.org/licenses/by-sa/4.0/