
# EURALEX XIX

## Congress of the European Association for Lexicography

**Lexicography for inclusion**

**7-11 September 2021**

Ramada Plaza Thraki

Alexandroupolis, Greece

www.euralex2020.gr

**Proceedings Book**
**Volume 1**

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris



2020 Edition

# Augmented Writing and Lexicography: A Symbiotic Relationship?

**Køhler Simonsen H.**

*Copenhagen Business School, Denmark*

**Abstract**

We live in an age of disruption and technological innovations, and lexicography as a scientific discipline and practice is witnessing a fundamental paradigm shift, cf. also (Fuertes-Olivera 2016), who talks about a "Cambrian Explosion", (Simonsen 2016), who discusses the need for a new "Lexicographic Business Model" and (Tarp 2019), who refers to the paradigm shift in lexicography as "Tradition and Disruption in Lexicography". Like many other disciplines, lexicography is operating within the framework of the "Fourth Industrial Revolution", cf. (Schwab 2015), and it seems to be facing many fundamental challenges.

One of these challenges is Augmented Writing (AW), cf. (Banks 2019; G2.com 2019; Marconi 2017 and Simonsen 2020a, 2020b), who discuss AW and how it affects journalism, communication and lexicography respectively.

The objective of this article is to discuss AW from a lexicographical perspective and to what extent the two disciplines may form a value-adding symbiotic relationship. Based on empirical data from a test of 32 AW technologies, the article discusses this question and presents a number of theoretical considerations on how AW and lexicography might develop a symbiotic relationship drawing on Colson (2019), Fadel et al. (2017), Liew (2013), Tarp (2019), and Simonsen (2020a, 2020b).

**Keywords**: Augmented Writing; Writing Assistants; Lexicographically Augmented Writing

## 1 Introduction

It is always dangerous to make predictions, especially when it comes to the impact of technology. Even the quite famous corporate turnaround expert, Jim Keyes, the then CEO of Blockbuster, got it very wrong when he predicted, "Neither RedBox nor Netflix are even on the radar screen in terms of competition" (Rapier 2020). He was very wrong. As we all know, Blockbuster went bankrupt only two years later.

Some would no doubt argue that this has nothing to do with lexicography. Others would argue that similar disruptive developments are already taking place in lexicography. One thing is certain. We can all learn from history.

One example of direct relevance for this article is *Write Assistant*, which has almost outcompeted virtually all established and renowned dictionary publishers in Denmark. Admittedly, this is just one example and one small country, but the adoption curve of disruptive technology is almost exponential and very much international. Consequently, there is an imminent need for discussing AW and the role it may have in lexicography.

Fortunately, lexicography is a strong science and discipline, and it has helped people understand, communicate and learn for thousands of years and it has much to offer. This article discusses how AW and lexicography can form a symbiotic relationship.

## 2 Research Question, Method, Data and Delimitations

The underlying research question of this paper is to answer the overall question: How can AW and lexicography form a symbiotic relationship?

The article draws on empirical insights from a structured test of 32 different AW technologies, (see also Simonsen 2020a; 2020b for a detailed discussion of the 32 AW technologies). The structured test and analysis of the AW technologies focused on parameters such as task types, degree of autonomy, workspace integration and lexicographic augmentation potential.

The analysis and discussion in this paper are delimited to AW technologies supporting text production and text analysis (sentiment analysis).

## 3 Literature Review

Computer-Assisted Language Learning (CALL) has no doubt played an important role in the development of AW. A particularly relevant contribution of CALL applications and dictionaries of relevance is Abel( 2009:5), who states that it

is crucial to categorize CALL applications based on their "central element and/or the starting point". A similar categorization is used here. The central element and/or starting point of most AWs is AI and most AWs aim at providing automatic lexical error correction and text production. CALL applications typically have a dictionary as its central element or language learning as its primary purpose. AW technologies thus seem to differ from CALL applications.

For the past 50 years publishing houses, computer linguists and lexicographers have developed a large number of language technological solutions, which have largely led to increased efficiency for translators and communicators. One landmark development started already in the 1970s when researchers discussed the possibility of translators using segments of already translated texts (Kay 1997). This led to the development of translation memory systems and *Sdltrados* was one of the first TM systems in use. Today, translators and professional text producers primarily use web-based systems like *Sdltrados* or *Wordfast*. This development to some extent also plays a role in modern AW.

Another landmark development was the many language technology solutions developed by computer linguists, IT experts and lexicographers. L2 writing research has been central to lexicography and language technology for the past 50 years, and recent research seems to focus on computer-supported collaborative writing, which in many ways is something much more advanced than the single-user AW technologies analysed here (Strobl 2014). Other contributions on L2 writing research and computer-supported collaborative writing are discussed by Arnold et al. (2009), De la Colina and García Mayo (2007), Elola and Oskoz (2010), Kessler et al. (2012), Kost (2011), and Storch (2005), to mention just a few.

Other landmark developments, which may have served as inspiration for many AW technologies, are based on computer-based writing instructions for text producers and learners (Allen et al. 2016) and the tool Writing Aid Dutch, which offers students process-oriented writing support (De Wachter et al. 2014). Furthermore, Frankenberg-Garcia et al. (2019) discuss a writing assistant, which is designed to help EAP writers with collocations, and Wanner et al. (2013) published a seminal discussion of writing assistants and automatic lexical error detection.

However, the above writing aids or writing assistants are not based on AI, and AW is widely different from many existing CALL applications and other types of language technological solutions because AW to a very high degree is based on AI and very often do not even use lexicographical data as the "central element and/or the starting point" (Abel 2009).

Recent landmark developments include Granger and Paquot (2015), who outline theoretical blueprints of a needs-driven online academic writing aid, Strobl et al. (2019), who offer a very useful review of different technologies for digital support for academic writing and, of course, the very relevant contributions by Tarp et al. (2017) and Tarp (2019), who discuss new challenges in lexicography based on the L2 writing assistant *Write Assistant* referred to at the beginning of this article.

Consequently, we need to develop theoretical considerations on lexicography and AW – because AW seems to need lexicography. However, before I do that, it is time to reflect on the insights from the empirical data.

## 4 Analysis and discussion

The analysis of the 32 AW technologies combined with the literature review of relevant theoretical contributions led to three overall findings. The first important finding based on the test of the 32 AW technologies made it possible to create an overall typology. It was found that the surveyed AW services can be divided into five overall groups.

**Group 1:** Spelling and grammar checkers such as Grammarly or WhiteSmoke. This category of tools is most often fully workspace-integrated and helps the user with automatic spelling and grammar recommendations. As shown below in Figure 1, Grammarly also includes an automatic tone of voice detector in addition to its grammar checker.

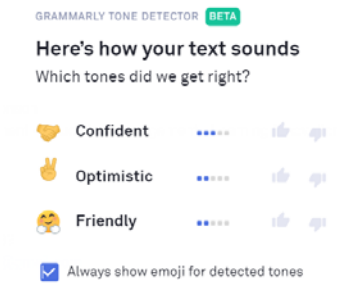


Figure 1: Grammarly.

**Group 2:** Text production robots such as TalktoTransformer or Articoolo. This category of tools is most often only browser-based and helps the user by autonomously producing texts based on just a few keywords. As shown below in Figure 2, TalktoTransformer automatically creates a text with just a few words using the GPT-2 Natural Language Understanding model. As will appear TalktoTransformer starts satisfactorily, but then the AI goes seriously astray.

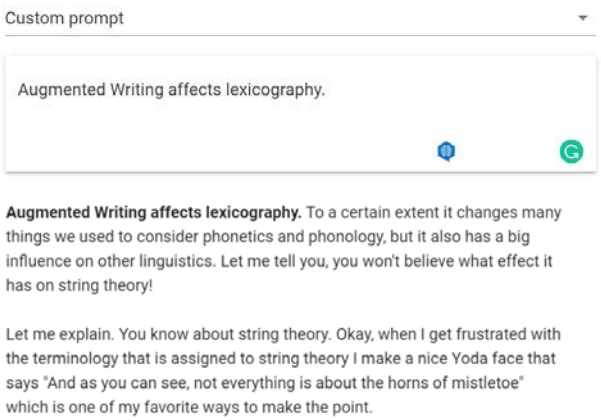


Figure 2: TalktoTransformer.

**Group 3**: L2 writing assistants such as Text Assistant. This type of tool is most often fully workspace-integrated and helps the user with context-aware recommendations in connection with L2 translation and L2 text production. The example in Figure 3 shows how Write Assistant predicts the next English word. Write Assistant is not AI-based and merely predicts the next word based on a language model and a 1:1 terminological relationship.

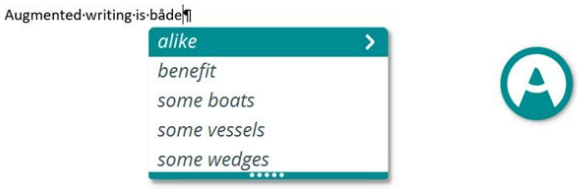


Figure 3. Write Assistant.

**Group 4:** Stylistic and tone of voice checkers such as Persado or MessagePath. This type of tool is most often workspace-integrated, particularly browser-based, and helps the user with stylistic and/or tone analysis of specific texts, for example, sales or marketing texts. Figure 4 below shows how the content and tone of voice analysis works in MessagePath.

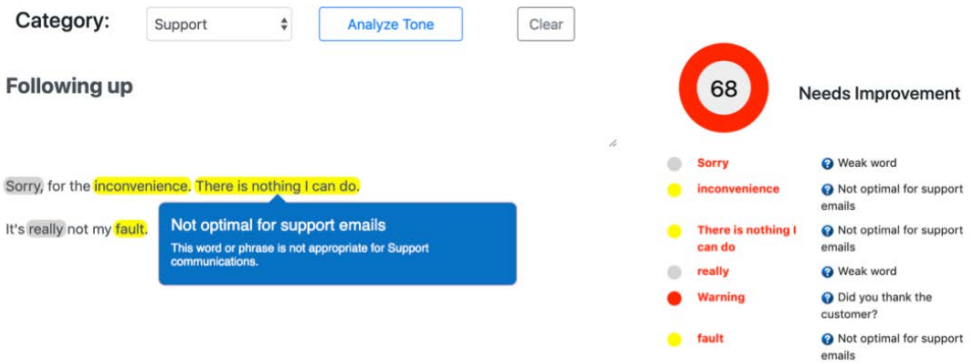


Figure 4: MessagePath.

**Group 5:** Special-purpose language pattern assistants such as Textio. This type of tool is most often browser-based and helps, for example, HR departments screening texts from candidates and producing job ads with the right sound. It also helps companies around the world produce insightful and inclusive texts based on data on age and gender bias.


Figure 5: Textio.

Finally, the structured test and analysis of the AW solutions also revealed that many news agencies have already implemented special-purpose robot journalists designed to extract and produce specific news articles, for example, financial news, soccer news or football match reports. The robot journalist tools are designed to extract data from existing news media and produce specific news articles based on text templates. In other words, AW also plays an increasing role in the news industry.

The second important finding from the structured test is that the technological maturity of many AW technologies is very high and they already seem to be a major challenge to many conventional lexicographic services such as spellchecking and grammar dictionaries. Tarp et al. (2017; 2019) reached a similar conclusion that L2 writing assistants and context-aware dictionaries seem to have much to offer to producers of L1 and L2 texts. AW really seems to challenge the type of lexicographical products, which focus exclusively on the delivery of data and information. This argument is already seen in Simonsen (2020a; 2020b), who argues that we may have to "turn lexicography upside down" dividing specific tasks between man and machine. Figure 5 below shows Liew's DIKIW model (Liew 2013) with my additions (dotted lines and vertical text).

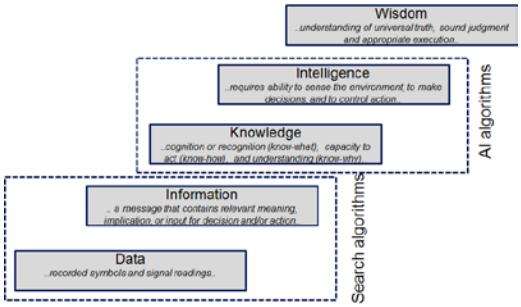


Figure 6: The DIKIW Model (my additions).

It is argued that lexicographical products, which solely focus on the first two levels in Liew's model (delivering data and information), are already being replaced by powerful search algorithms. Many people do not look up words in a dictionary anymore. They merely double-click and then right-click on a word in MS Word and perform a smart search on Bing and/or Google, cf. also de Schryver (2012:130), who observed this experimentally. Furthermore, the structured test of the 32 different AW technologies also showed that the next two levels in Liew's model (providing knowledge and intelligence) are increasingly being challenged by AI and even though existing autonomous AW solutions still leave much to be desired when it comes to quality and relevance, they are improving exponentially.

The third overall finding from the structured test was that AW platforms are moving into the lexicographical arena. This may have dramatic consequences for dictionaries providing users with knowledge and understanding as they may be in danger of being disrupted or replaced by AI (Simonsen 2020a; 2020b). AW solutions based on strong AI may very well become the next big disruptor in lexicography because the development of these AI technologies has priority in many countries. Their ease of use, ubiquity and degree of integration make them interesting for many users.

However, the test also revealed that the quality of the autonomous AW solutions such as TalktoTransformer leaves much to be desired. Most AW solutions will first try to understand the context of the input you feed into them using AI algorithms. Then they will locate the best text resources available and reconstruct it all to one coherent text through language models or NLP engines. So AW solutions are not necessarily based on curated data, but language models. To sum up, the test showed that AW needs curated lexicographical data, world knowledge and relational knowledge and thus needs to form a relationship with lexicography.

The output quality of many AW technologies can be improved significantly using curated lexicographical data. These lexicographical data should be available in special corpora and used when the AW attempts to locate the best text resources available. In other words, lexicography can help AW with curated lexicographic data and thus significantly

improve the output quality.

The output quality of many AW technologies can also be improved by adding world knowledge and relational knowledge to the actual output of the AW. Most AW technologies do not sufficiently understand context and lexicography might provide both world knowledge and relational knowledge (Simonsen 2020a; 2020b). In other words, lexicography can also help AW as condensation and description of world knowledge is central to lexicography.

Providing world knowledge and relational knowledge is not an easy task, but AW technologies could be equipped with an auxiliary post-editing window providing as much help as possible to the user when post-editing the output text. Similar arguments are found in Leroyer and Simonsen (2019), who have developed a framework for providing help to users when post-editing professional texts.

The suggested framework for the lexicographical augmentation of AW technologies takes its starting point in the division of labour between man and machine (Colson 2019), the layered understanding of data, information, knowledge and intelligence (Liew 2013) and last but not least, the idea of providing access to specially selected lexicographical data in the post-editing phases (Leroyer & Simonsen 2019). The suggested framework is based on OpenAi's Natural Language Understanding (NLU) model, which was trained to perform a single task of predicting the next word with a given set of words and a very large dataset (Rodriquez 2019). The GPT-2 model is used in TalktoTransformer and it does yield amazing output based on just a few words as it was demonstrated in Figure 2 above.

I argue that the symbiotic relationship between AW and lexicography could be consummated by building an AW where curated lexicographical data are simply part of the first priority training datasets. This would significantly improve the output quality of the AW.

When it comes to improving the output of AW technologies with world knowledge and relational knowledge it is much more complex. It is not about just inserting yet another fine-tuning layer in OpenAi's GPT-2 language model (OpenAi 2019). Human augmentation and intervention are needed. I argue that an external post-editing window is needed because human augmentation is required when adding world knowledge or relational knowledge in line with (Colson 2019), who makes the case for the division of labour between man and machine. This external post-editing window could be the final step in the output process of a lexicographically augmented AW technology. In other words, lexicographically augmented AW technologies might be what we need.

## 5 Conclusion

Building on Colson (2019), Banks (2019), Liew (2013), Marconi (2017), Simonsen (2020a, 2020b), Tarp et al. (2017) and Tarp (2019) and the empirical analysis, this article offered a discussion of selected AW services.

Based on the structured test it was first possible to develop an overall typology of AWs, which were categorized in five overall groups. The second finding based on the structured test was that the technological maturity of most AW technologies is very high. The third finding was that the output quality of most AW technologies leaves much to be desired and that what is needed is curated lexicographical data and world knowledge and relational knowledge.

The analysis and discussion of the 32 AW technologies also revealed that AW is or may develop into a major challenge to many conventional lexicographic services offering only data and information (Liew 2013). The discussion also revealed the weaknesses and lacking quality of some AW technologies and the discussion uncovered many considerations on how lexicography can augment AW or even form a symbiotic relationship with AW.

Lexicography has an important role to play in the development of new advanced text production technologies and the lexicographical augmentation of AW could be an important step in the right direction. In conclusion, lexicography has much to offer to AW especially when it comes to human augmentation of the automatic output from an AW service.

## 6 References

Abel, A. (2009). Towards a systematic classification framework for dictionaries and CALL. In S. Granger and M. Paquot (eds), eLexicography in the 21st century: New challenges, new applications. *Proceedings of eLex 2009*, Louvain-la-Neuve, 22-24 October 2009, 3-11.

Allen, L., Jacovina, M. & McNamara, D. (2016). Computer-based writing instruction. In C.A. MacArthur, S. Graham & J. Fitzgerald (eds.) *Handbook of writing research.* New York, NY: Guilford, pp. 316-329.

Arnold, N., Ducate, L., & Kost, C. (2009). Collaborative writing in wikis: Insights from culture projects in German classes. In L. Lomicka & G. Lord (Eds.), *The next generation: Social networking and online collaboration in foreign language learning* (pp. 115-144). San Marcos, TX: CALICO.

Banks, C. (2019). What is an Augmented Writing Platform? Accessed at: https://medium.com/swlh/what-is-an-augmented-writing-platform-b28fa588a1c5 [19/04/2020].

Colson, E. (2019). What AI-Driven Decision Making Looks Like. Accessed at: https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like. [19/04/2020].

de Schryver, Gilles-Maurice (2012). Lexicography in the crystal ball: Facts, trends and outlook. In: Fjeld, Ruth V. & Julie M. Torjusen (eds). *Proceedings of the 15th EURALEX International Congress*, 7-11 August, 2012, Oslo: 93–163. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

De la Colina, A. A., & García Mayo, M. d. P. (2007). Attention to form across collaborative tasks by low-proficiency learners in an EFL setting. In M. d. P. García Mayo (Ed*.), Investigating tasks in formal language learning* (pp. 91-116). Clevedon: Multilingual Matters Ltd.

De Wachter, L., Verlinde, S., D'Hertefelt, M., Peeters, G.,Tounsi, L. (2014): How to deal with students' writing problems? Process-oriented writing support with the digital Writing Aid Dutch. In Rak, Rafal (Editor) The 25th International Conference on Computational Linguistics, Date: 2014/08/23 - 2014/08/29, Location: Dublin. *Proceedings of the Conference. System Demonstrations;* 2014; pp. 20 – 25.

Elola, I., & Oskoz, A. (2010). Collaborative writing: Fostering foreign language and writing conventions development. *Language Learning and Technology*, 14(3), 51-71.

Fadel, C., Bialik, M. & Trilling, B. (2017). Fire-dimensional uddannelse: Kompetencer til at lykkes i det 21. århundrede. 1. udgave, 1. oplag. Dafolo.

Frankenberg-Garcia, A., Lew, R., Roberts, J., Rees, G & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL, 31(1)*, pp. 23-39.

Fuertes-Olivera, P.A. (2016). A Cambrian Explosion in Lexicography: Some Reflections for Designing and Constructing Specialised Online Dictionaries.In *International Journal of Lexicography 29(2)*: 226-247.

Granger, S. & Paquot, M. (2015). Electronic lexicography goes local: Designs and structures of a needs-driven online academic writing aid. *Lexicographica, 31(1)*, pp. 118–141.

G2.com (2019). Best AI Writing Assistant Software. Accessed at: https://www.g2.com/categories/ai-writing-assistant [19/04/2020].

Kay, M. (1997). The Proper Place of Men and Machines in Language Translation. In *Machine Translation*. 12 (1–2): 3–23.

Kessler, G., Bikowski, D., & Boggs, J. (2012). Collaborative writing among second language learners in academic web-based projects. In *Language Learning & Technology, 16(1)*, 91-109.

Kost, C. (2011). Investigating writing strategies and revision behaviour in collaborative wiki projects. In CALICO Journal, 28(3), 606-620.

Leroyer, P. & Simonsen, H. K. (2019). Google Translate som trussel eller redning for oversættelsesordbøger. In *LexicoNordica 26*, 2019.

Liew, A. (2013). DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships. In *Business Management Dynamics*. Vol. 2, Issue 10, April 2013, 49-62.

Marconi, F. (2017). NiemanLab. Predictions for Journalism (2017): The Year of Augmented Writing. Accessed at: <https://www.niemanlab.org/2016/12/the-year-of-augmented-writing [19/04/2020].

OpenAi. Accessed at: https://openai.com/ [19/04/2020]

Rapier, G (2020). 13 Quotes From Bosses Who Mocked Technology and Got It (Very) Wrong. Accessed at: https://www.inc.com/business-insider/boss-doesnt-understand-technology-mocks-trend-wrong.html [19/04/2020].

Rodriquez, J. (2019). One Language Model to Rule Them All. Accessed at: https://towardsdatascience.com/one-language-model-to-rule-them-all-26f802c90660 [19/04/2019]

Schwab, K. (2015). The Fourth Industrial Revolution: What It Means and How to Respond. Foreign Affairs. Accessed at: www.foreignaffairs.com/articles/2015-12-12/fourth-industrialrevolution [19/04/2020].

*Sdltrados*. Accessed at: https://www.sdl.com/software-and-services/translation-software/sdl-trados-studio/ [19/04/2020].

Simonsen, H. K. (2016). Hvor er forretningsmodellen? *En analyse af de forretningsmæssige udfordringer i forlags- og informationsindustrien med særlig fokus på opslagsværker. MBA-afhandling.* Institut for Økonomi og Ledelse. Aalborg Universitet.

Simonsen, H. K. (2020a). Augmented Writing: nye muligheder og nye teorier. In *Nordiske Studier i Leksikografi 15, 2019*, Rapport fra 15. Konference om Leksikografi i Norden – Finland 4. juni–7. juni 2019 (In press).

Simonsen, H. K. (2020b). Når Augmented Writing og leksikografi går hånd i hånd. In: *LEDA-nyt nr. 69* - april 2020, 3-13.

Storch, N. (2005). Collaborative writing: Product, process, and students´ reflections. *Journal of Second Language Writing, 14(3)*, 153-173.

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A. & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. In *Computers & Education, 131*, pp. 33-48.

Tarp, S. (2019). Connecting the Dots: Tradition and Disruption in Lexicography. In *Lexikos 29*, 224-249.

Tarp, S., Fisker, K., & Sepstrup, P. (2017). L2 writing assistants and context-aware dictionaries: New challenges to lexicography. In *Lexikos 27*(1), 494-521.

Wanner, L., Verlinde, S. & Alonso Ramos, M. (2013). Writing assistants and automatic lexical error correction: word combinatorics. In. Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M. & Tuulik, M. (eds.) Electronic lexicography in the 21st century: thinking outside the paper. *Proceedings of the eLex 2013 conference*, 17-19 October 2013, Tallinn, Estonia, pp. 472-487.

*Wordfast*. Accessed at: https://www.wordfast.com/ [19/04/2020].

*Write Assistant*. Accessed at: https://www.writeassistant.com/da/ [19/04/2020].