UNIVERSITY OF COPENHAGEN

Data-driven learning for languages other than English: the cases of French, German, Italian, and Spanish

Reka Jablonkai[1], Luciana Forti[2], Magdalena Abad Castelló[3], Isabelle Salengros Iguenane[4], Eva Schaeffer-Lacroix[5], and Nina Vyatkina[6]

Abstract. This paper summarises the contributions to EuroCALL's CorpusCALL SIG Symposium for the year 2020. In line with this year's EuroCALL conference theme, 'CALL for widening participation', the Symposium centred around the theme of *Data-driven learning for languages other than English.* This paper gives a brief overview of developments and challenges when using Data-Driven Learning (DDL) to teach French, German, Italian, and Spanish. As research suggests, a DDL approach has been effectively utilised to teach these languages. However, there are differences in available DDL resources and corpora for the respective languages that are appropriate for language teaching. The main challenges for future developments are also discussed.

Keywords: DDL, corpora, LOTEs.

1. Introduction

This paper shares developments in using DDL in teaching Languages Other Than English (LOTEs) within the wider DDL community. As literature on DDL has primarily focused on studies in the context of teaching English (Chambers, 2019), we provide brief overviews of the current state of DDL in relation to the teaching

1. University of Bath, Bath, United Kingdom; reka.jablonkai@gmail.com; https://orcid.org/0000-0001-9616-5683
2. University For Foreigners of Perugia, Perugia, Italy; luciana.forti@unistrapg.it; https://orcid.org/0000-0001-5520-7795
3. Instituto Cervantes Máncheste, Manchester, United Kingdom / Universidad Internacional Iberoamericana, Puerto Rico; malena@abad.com; https://orcid.org/0000-0002-1942-7421
4. École des Ponts ParisTech, Marne-la-Vallée, France; isasalengros@hotmail.com
5. Sorbonne Université / Inspé de l'académie de Paris, Paris, France; eva.lacroix@sorbonne-universite.fr; https://orcid.org/0000-0002-6260-9095
6. University of Kansas, Lawrence, Kansas, United States; vyatkina@ku.edu; https://orcid.org/0000-0002-2778-8016

How to cite: Jablonkai, R., Forti, L., Abad Castelló, M., Salengros Iguenane, I., Schaeffer-Lacroix, E., & Vyatkina, N. (2020). Data-driven learning for languages other than English: the cases of French, German, Italian, and Spanish. In K.-M. Frederiksen, S. Larsen, L. Bradley & S. Thouësny (Eds), *CALL for widening participation: short papers from EUROCALL 2020* (pp. 132-137). Research-publishing.net. https://doi.org/10.14705/rpnet.2020.48.1177

of French, German, Italian, and Spanish. Each overview discusses challenges and proposes solutions to realising the full potential of the DDL approach. First, an empirical study of DDL for French is reported. Next, we provide a brief overview of the range and effectiveness of corpus resources used for teaching and learning German and indicate directions for future resource development and empirical research. We then trace a brief historical overview of DDL for Italian, with an indication of the main challenges that the field faces today. Finally, challenges of DDL for teachers and learners of Spanish are discussed.

2. DDL for French: linking professional communication skills and linguistic features

Research papers from the French DDL community mainly report on indirect applications (Vyatkina, 2020a), with learner corpora analysed as error repositories (Dubois, Kamber, & Dekens, 2013) or as resources for designing learning materials (Di Vito, 2013). Direct applications are mentioned within the context of academic writing (Jacques & Rinck, 2017) and French for specific purposes (Rodgers & Chambers, 2011). Here we present the results of a study focusing on the direct use of a small, specialised corpus by a group of 12 international engineering students enrolled on a professional writing course for advanced learners of French as a foreign language (target level: B2-C1). The study aimed to determine whether guided observation of corpus data could help these students better understand recurrent language errors in their first drafts of technical specification documents, in French called 'Cahier des Clauses Techniques Particulières' (CCTP). We chose 14 CCTP samples to create a corpus accessible via Sketch Engine (Kilgarriff et al., 2014). In this corpus, we identified linguistic features corresponding to the professional communication skills targeted (see Table 1). The observed errors mainly correspond to these features.

Table 1. Professional communication skills and linguistic features

Professional communication skills	Linguistic features
Be neutral and objective	Passive voice and noun|adjective|verb agreement
Avoid mentioning agents	Impersonal structures or pronouns
Mention norms and standards	Verbs ('inform','prescribe', 'schedule', 'contain')
Describe or specify	Demonstrative pronouns (celui, celle, ceux/celui-ci, celle-ci, ceux-ci, celles-ci) [the one, those/this one, these]

During the course, the participants completed worksheets containing activities partly inspired by their own errors, and they answered two online questionnaires. The data obtained inform about the learners' progress and remaining needs. We conclude from this study that the specialised CCTP corpus offers enough data to support students who have to write a pedagogical version of a CCTP. However, more training time is needed to better explain to them the technical features of Sketch Engine. They also need to learn how to notice linguistic features and report their findings.

To boost the DDL L2 French sector, we recommend choosing a user-friendly corpus tool and concentrating on learning issues. The content of the corpus must correspond to the writing task and the query activities should focus on the observed learning needs.

3. DDL for German: available resources, learning outcomes, and future directions

The subfield of DDL for German, like the broader DDL field, can be divided into pedagogical materials, classroom reports, and empirical research. The subfield's origins go back to the turn of the 21st century (e.g. Dodd, 2000; St. John, 2001). In the most recent synthesis of DDL research, Boulton and Vyatkina (forthcoming) identify 14 empirical studies that explored the effectiveness of DDL for teaching German. Like most DDL research (ca. 90% of which has been dedicated to teaching English), studies on DDL for German primarily focus on university contexts and DDL interventions developed and administered by the researchers themselves. They report improved learner knowledge of German lexico-grammar and pragmatics as well as writing, translation, and interpreting skills and favourable learner attitudes. The geographic coverage of these studies is encouragingly broad, including seven countries and three continents, which attests to the generalizability of the findings. While more studies are needed in university contexts, promising future directions could also include an expansion of DDL for German to primary and secondary schools.

A unique feature of the German subfield is the availability of several large, well-designed, sustainable, and open-access corpora. The missing link between these rich resources and a broader German-learning and German-teaching population is teacher and learner DDL guides, written in accessible language and tethered to specific corpora. One such guide to using the DWDS corpus (http://dwds.de) and associated DDL exercises currently are being developed and gradually released

with open access at the University of Kansas (Vyatkina, 2020b). It is hoped that other DDL researchers can use this resource as a model for "bringing corpora to the masses" (Boulton, 2011, p. 69) in DDL for German and beyond.

4. DDL for Italian: studies, practices, and future prospects

The studies on DDL for Italian cover a time span of at least 27 years. A solid starting point can be traced back to 1993, when Polezzi published her pioneering work in ReCALL. Polezzi (1993) showed how a corpus of Italian for specific purposes could be built and used with beginner learners of Italian enrolled in a postgraduate course in Renaissance Studies. She supported the idea of a *didactic language corpus,* identifying the characteristics that would make such a corpus suitable for specific language learning needs.

Since then, the studies on DDL for Italian have risen steadily but not steeply. To the best of our knowledge, they are no more than 20 in total, consisting mostly of descriptive studies (e.g. Corino & Marello, 2009), and with still very few empirical studies (e.g. Forti, 2019).

The pedagogical practices adopted in the context of DDL for Italian have been closely linked to the characteristics of available corpora. While freely accessible reference corpora of Italian are available, they were primarily built by researchers for researchers. As a result, their pedagogical potential is generally restricted to the development of paper-based materials and to advanced-level learners. The first learner-friendly corpus exploration tool for Italian was developed very recently, within the SkELL platform (Baisa & Suchomel, 2014).

Bridging the teacher-researcher gap (Chambers, 2019) is one of the main challenges that DDL for Italian faces today. Integrating corpora in teacher training programmes, publishing teacher guides and developing more learner-friendly corpus exploration tools are ways to help bridge this gap.

5. DDL for Spanish: attitudes and tasks in the use of corpora

DDL did not have a name in Spanish until fairly recently. Two terms were coined (*aprendizaje basado en datos* and *aprendizaje guiado por datos)*. The field adopted

the former, likely thanks to the seminal article by Asención-Delaney et al. (2015), which reported the profusion of pedagogical articles and the shortage of empirical studies. Since then, the field has experienced a steady growth of empirical research in DDL with both native and learner corpora as sources (Benavides, 2015; Yao, 2019).

In terms of resources, there are vast open-access native corpora, such as Corpus del Español (BYU) or CORPES XXI, and also important learner corpora (such as CAES, Aprescrilov, CEDEL 2). Among the numerous pedagogical articles, the scope of learning targets has widened from lexico-grammar to pragmatics, discourse features and pronunciation (using oral corpora), and varieties of Spanish. Corpus-based tasks can also be found in recently published textbooks (e.g. Aula Internacional 4, Prisma C2), which is helping to spread DDL among practitioners and learners.

Despite this growth spurt, DDL is very far from being normalised in Spanish as a foreign language teaching practice. One main challenge lies in changing teachers' attitudes towards corpus use by training programmes and by integrating corpus use in the syllabus. As in other LOTEs, most Spanish teachers do not seem to be aware of the benefits of using corpora in language teaching. In addition, there is a need for ready-made materials and "online corpus user guides for teachers and exercises integrated with specific corpora" (Vyatkina, 2020a, p. 364) that can inspire teachers to develop their own corpora.

6. Conclusions

This brief overview on DDL research for LOTEs revealed that DDL has effectively been used for teaching the languages considered. Challenges to DDL often centre around availability of appropriate corpora and tools for practitioners. The paper concentrated on a handful of European languages. Further reviews should explore developments of DDL within a wider geographical scope, including, for example, Arabic, Mandarin, and Russian.

References

Asención-Delaney, Y., Collentine, J. G., Collentine, K., Colmenares, J., & Plonsky, L. (2015). El potencial de la enseñanza del vocabulario basada en corpus: optimismo con precaución. *Journal of Spanish Language Teaching*, *2*(2), 140-151. https://doi.org/10.1080/23247797.2015.1105516

Baisa, V., & Suchomel, V. (2014). SkELL: web interface for English language learning. In A. Horák & P. Rychlý (Eds), *Proceedings of Recent Advances in Slavonic Natural Language Processing* (pp. 63-70). https://nlp.fi.muni.cz/raslan/raslan14.pdf

Benavides, C. (2015). Using a corpus in a 300-level Spanish grammar course. *Foreign Language Annals, 48*(2), 218-235.

Boulton, A. (2011). Bringing corpora to the masses: free and easy tools for interdisciplinary language studies. In N. Kübler (Ed.), *Corpora, language, teaching, and resources* (pp. 69-96). Peter Lang.

Boulton, A., & Vyatkina, N. (forthcoming). Thirty years of data-driven learning: taking stock and charting new directions [Manuscript submitted for publication].

Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching*, *52*(4), 460-475. https://doi.org/10.1017/S0261444819000089

Corino, E., & Marello, C. (2009). Didattica con i corpora di italiano per stranieri. *Italiano LinguaDue*, *1*(1), 279-285.

Di Vito, S. (2013). L'utilisation des corpus dans l'analyse linguistique et dans l'apprentissage du FLE. *Linx*, *68-69*, 159-176. https://doi.org/10.4000/linx.1519

Dodd, B. (2000). *Working with German corpora*. University of Birmingham Press.

Dubois, M., Kamber, A., & Dekens, C. S. (2013). Être et avoir été: l'accord du participe passé par des apprenants de FLE. *Linx, 68-69*, 115-133. https://doi.org/10.4000/linx.1504

Forti, L. (2019). *Developing phraseological competence in Italian L2: a study on the effects of data-driven learning*. Unpublished PhD thesis. Università per Stranieri di Perugia.

Jacques, M.-P., & Rinck, F. (2017). Un « corpus de littéracie avancée : résultat et point de départ. *Corpus, 16*. https://doi.org/10.4000/corpus.2806

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*(1), 7-36. https://doi.org/10.1007/s40607-014-0009-9

Polezzi, L. (1993). Concordancing and the teaching of ab initio Italian language for specific purposes. *ReCALL*, *5*(9), 14-18. https://doi.org/10.1017/s0958344000004067

Rodgers, O., & Chambers, A. (2011). Corpora in the LSP classroom: a learner-centred corpus of French for biotechnologists. *International Journal of Corpus Linguistics, 16*(3), 391-411. https://doi.org/10.1075/ijcl.16.3.06rod

St. John, E. (2001). A case for using a parallel corpus and concordancer for beginners of a foreign language. *Language Learning & Technology, 5*(3), 185-203.

Vyatkina, N. (2020a). Corpora as open educational resources for language teaching. *Foreign Language Annals, 53*(2), 359-370. https://doi.org/10.1111/flan.12464

Vyatkina, N. (2020b). (Ed.). Incorporating corpora: using corpora to teach German to English-speaking learners [Online instructional materials]. KU Open Language Resource Center. https://corpora.ku.edu

Yao, G. (2019). Vocabulary learning through data-driven learning in the context of Spanish as a foreign language. *Research in Corpus Linguistics, 7*, 18-46. https://doi.org/10.32714/ricl.07.02

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

CALL for widening participation: short papers from EUROCALL 2020
Edited by Karen-Margrete Frederiksen, Sanne Larsen, Linda Bradley, and Sylvie Thouësny

Publication date: 2020/12/14

Typeset by Research-publishing.net
Cover theme by © 2020 Marie Flensborg (frw831@hum.ku.dk), based on illustration from freepik.com
Cover layout by © 2020 Raphaël Savina (raphael@savina.net)

ISBN13: 978-2-490057-81-8 (Ebook, PDF, colour)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2020.