Using learner data from Duolingo to detect micro- and macroscopic granularity through machine learning methods to capture the language learning journey

Belinda Chiera[1], Branislav Bédi[2], and Rina Zviel-Girshin[3]

Abstract. Modern language learning applications have become 'smarter' and 'intelligent' by including Artificial Intelligence (AI) and Machine Learning (ML) technologies to collect different kinds of data. This data can be used for analysis on a microscopic and/or macroscopic level to provide granulation of knowledge. We analyzed 1,213 French language learner data over a 30-day period, publicly available from Duolingo, to compare the progression of individual learners (microscopic granularity) and large groups of learners (macroscopic granularity). Using network modeling, we compared patterns of individual learners against one another and that of a learning community and determined what groups of learners typically practice across communities. Preliminary results suggest how applications for L2 learning can be designed to create an optimal path for learning.

Keywords: artificial intelligence, granularity, language learning journey, machine learning.

1. Introduction

In the past ten years, an increasing number of language learning applications have become 'smarter' and 'intelligent' by including features from both AI and ML. Both technologies enable collecting different data types for analysis on a microscopic and/or macroscopic level. Such analysis is called granularity, with microscopic

[1] University of South Australia, Adelaide, Australia; belinda.chiera@unisa.edu.au; https://orcid.org/0000-0001-9274-0943
[2] The Árni Magnússon Institute for Icelandic Studies, Reykjavik, Iceland; branislav.bedi@arnastofnun.is; https://orcid.org/0000-0001-7637-8737
[3] Ruppin Academic Center, Emek Hefer, Israel; rinazg@ruppin.ac.il; https://orcid.org/0000-0002-7926-4476

How to cite this article: Chiera, B., Bédi, B., & Zviel-Girshin, R. (2022). Using learner data from Duolingo to detect micro- and macroscopic granularity through machine learning methods to capture the language learning journey. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, & M. J. Whelpton (Eds), *Intelligent CALL, granular systems, and learner data: short papers from EUROCALL 2022* (pp. 58-64). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1435

granularity focusing on individual learners while macroscopic granularity compares large learner groups. This article explores both granularities and provides insights into a language learning journey by analyzing a sample of French language learner data. We focus on identifying and contrasting common patterns of learning behavior for insight into learners experiencing either exceptionally good or poor progress.

Many learners engage and possibly re-engage with a language learning program over time, in the pursuit of learning a target language (L2). In the majority of 'smarter' or 'intelligent' language learning applications, AI/ML technologies learn about, e.g. a learner's errors, engagement, time and program usage, mistakes in specific exercises, and so forth. Programs with AI/ML technologies collect data on individual learners performing similar actions over time yet resulting in varying levels of language competency. Insights are achieved through learning analytics, or microscopic granulation, to influence a participant's motivation and improve learner behavior to keep the learner actively engaged (Greller, Ebner, & Schön, 2014; Hai-Jew, 2014). Results from macroscopic granulation can utilize all sources of learners' actions to create an optimal path for learning (Tang, Peterson, & Pardos, 2016). Such granularity helps to design lessons in learning systems to become more digestible (Jasnani, 2013), i.e. less challenging in terms of the required competences, e.g. time management, learner's self-regulated learning, and self-organization (Lackner, Ebner, & Khalil, 2015).

2. Method

We used data generated by 1,213 French language learners over a 30-day period, publicly available from Duolingo (Settles, 2018), to demonstrate micro- and macroscopic granulation; chosen solely for the reason that this data is freely available. Each learner had a unique, anonymized ID with their session data including country of access, days in the program, time taken to complete a task, Part-Of-Speech (POS) tags, and whether the learner produced a correct answer. Three session types were recorded: *lesson*, *practice*, and *test* with three different activities, *listen*, *reverse tap*, and *reverse translate*. Reverse translate gives a learner a question in L1 to translate into L2 whereas reverse tap requires learners to construct an answer to a question in L1 using a small set of words (including distractors) in L2 (Settles, 2018). Most learners attempted all activities, however typically used two of three session types. *Reverse translate* was used by all learners while individuals selecting two activities predominantly also selected *listen*. Nearly all learners used *lesson* with several using *test*. A popular combination was *lesson* and *practice*.

We created networks of node-edge pairs from the data. For microscopic granularity we investigated co-occurrence networks of an individual learner's frequency of exposure to POS component combinations (NOUN, VERB, DET, etc.) and the vocabulary breadth encountered. For macroscopic granulation we grouped by country and created signed networks capturing correct and incorrect POS combinations, to provide insight into key lexical combinations learners found challenging or easier to grasp. The Jaccard index (Agresti, 2002) for network edge set similarity identified common elements between networks.

A mixed methodology compared individual performance against a community. The benefit of this approach is the formation of networks creating a multidimensional learning structure, even if learners do not engage directly with each other. These insights can benefit learners by gauging individual proficiency against common learning patterns of other individuals with a similar profile.

3. Results and discussion

Microscopic granulation (Figure 1) compares the typical time taken to complete learning activities of two learners, where learner (a) interacted with all activities over 25 days in the program, completing 752 tasks, whereas learner (b) used *reverse translate* only over 27 days and completed 131 tasks.

The POS graphs in Figure 2 indicate regular exposure to similar combinations (NOUN, VERB, DET, PRON) although learner (b) received greater exposure to ADJ, ADV, and INTJ components despite engaging less frequently. Vocabulary varied by learner. Figure 3 shows learner (a) predominantly practiced determinants *le* and *un* while learner (b) also practiced a third determinant, *les*.

Figure 1. Time taken to complete the learning activities for two learners

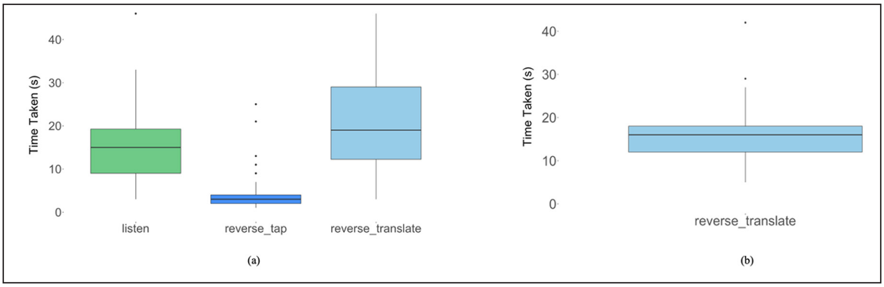
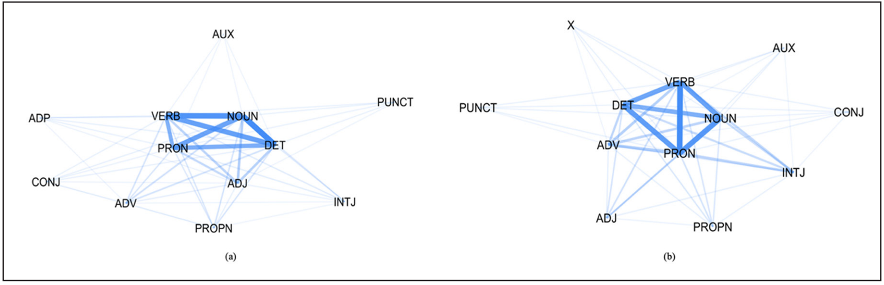


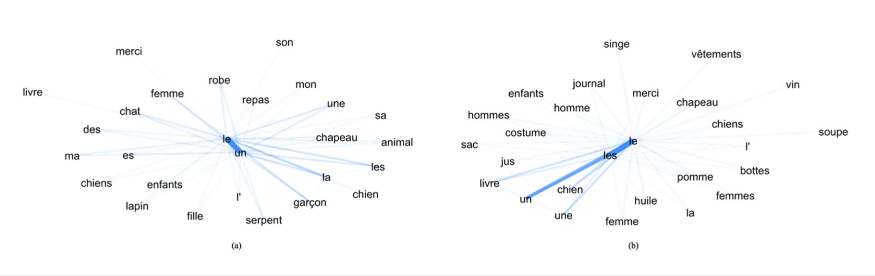
Figure 2. Exposure to POS by learner. Heavier lines indicate more frequent exposure



(a)

(b)

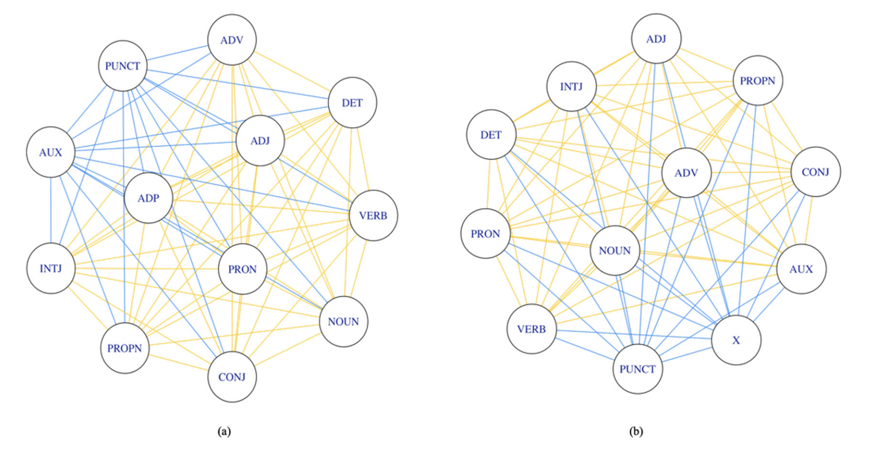
Figure 3. Vocabulary exposure by learner. Heavier lines indicate more frequent exposure



(a)

(b)

Figure 4. Signed networks indicating correct answers in blue and incorrect answers in yellow



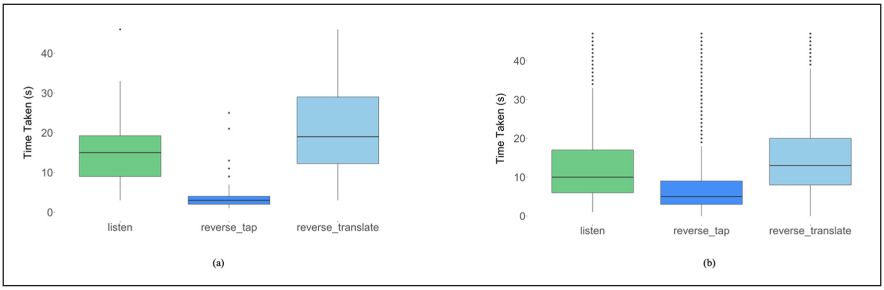
(a)

(b)

In Figure 4, learners (a) and (b) are compared using signed networks to capture POS combinations. The learners answered correctly on average (blue) and incorrectly on average (yellow). Although learners were exposed to different vocabulary (Figure 3) and training activities, the data indicated similar proficiency. It could be postulated that learner (a) had less prior language exposure, while learner (b) engaged with the learning program to refresh their knowledge. The Jaccard index reported 71% commonality between the learners, supporting similarities in learning outcomes.

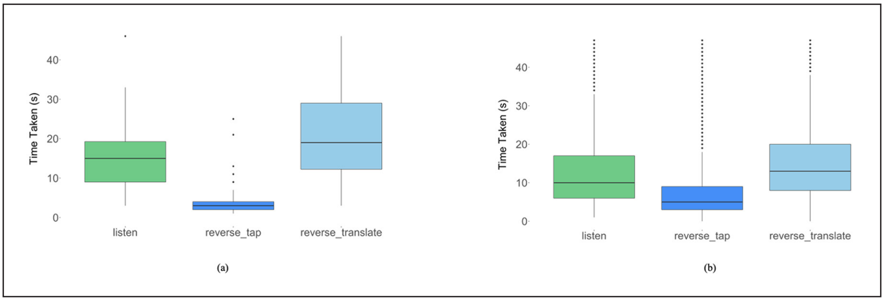
Results from macroscopic granulation can potentially be implemented as part of a language learning platform to provide a sophisticated learning mechanism for learner agency, to enable a language learner to compare their progress with that of their peers. For example, macroscopic granularity may alert a learner to limited variability in exposure to key language components when compared to their peers, which would encourage the learner to request more advanced learning modules. A mixed methods approach (Figure 5) compares learner (a) with all learners from the same country of access (b), to inform expected proficiency of an individual, based on country of access. Learner (a) is typically faster with *reverse tap* than other learners from the same country, however, takes typically longer with activities such as *reverse translate,* with extreme values comparable to those of the learner community.

Figure 5. Time taken to complete the learning activities from an individual learner (a) and all learners from the same country of access (b)



In Figure 6, learners in countries (a) and (b) are compared. Country (a) learners are typically quicker with *reverse tap* however typically slower otherwise. Such insights could inform the language learning platform how to adjust content and challenges such that learners from country (a) could be encouraged to practice *listen* and *reverse translate*, whereas learners from country (b) could be encouraged to engage with *reverse tap* tasks. These recommendations could be implemented within the language learning tool, thereby allowing learners the opportunity to take responsibility along their language learning journey by evaluating their own progression against a community of peers of interest.

Figure 6. Time taken to complete the learning activities between learners accessing the program from countries (a) and (b)



4. Conclusions

This article explored microscopic and macroscopic granularity to demonstrate that learner data accessed at individual and group levels can provide insights into assessing an individual's learning journey, to allow supportive learning advice that can be auto generated when combined with more sophisticated iterations of this analysis. Having insight at a micro-granular level allowed a comparison of learners to uncover individual progression relative to engagement with the language learning platform. Macro-granular level analysis allowed for a complementary comparison at a community level, to allow assessment of the individual's performance against a community of peers of interest, as well as the ability to compare communities with one another.

Limitations of the current study include the investigation of a single data set comparing English and French users and the inability to characterize learners beyond their country of access, which limits the potential to compare community structures. Future research already underway considers the exploration of more than one data set to provide granularity across multiple languages of interest. To support learner characterization, open-source applications and those supported by crowdsourcing are encouraged to make richer datasets publicly available to researchers who are interested in designing new, or improving existing, applications.

References

Agresti, A. (2002). *Categorical data analysis*. John Wiley and Sons.

Greller, W., Ebner, M., & Schön, M. (2014). Learning analytics: from theory to practice – data support for learning and teaching. In M. Kalz & E. Ras (Eds), *Computer assisted assessment. Research into e-assessment. CAA 2014. Communications in Computer and Information Science*, 439, 79-87. https://doi.org/10.1007/978-3-319-08657-6_8

Hai-Jew, S. (2014). Iff and other conditionals: expert perceptions of the feasibility of massive open online courses (MOOCs) – a modified e-delphi study. In S. Hai-Jew (Ed.), *Remote workforce training: effective technologies and strategies* (pp. 278-410). IGI Global. https://doi.org/10.4018/978-1-4666-5137-1.ch013

Jasnani, P. (2013). *Designing MOOCs. A white paper on instructional design for MOOCs*. Tata Interactive System. http://www.tatainteracive.com/pdf/Designing_MOOCs-A_White_Paper_on_ID_for_MOOCs.pdf

Lackner, E., Ebner, M., & Khalil, M. (2015). MOOCs as granular systems: design patterns to foster participant activity. *eLearning papers, 42*, 28-37.

Settles, B. (2018). *Data for the 2018 Duolingo shared task on second language acquisition modeling (SLAM), Harvard Dataverse (V4)* [Data file]. https://doi.org/10.7910/DVN/8SWHNO

Tang, S., Peterson, J. C., & Pardos, Z. A. (2016). *Modelling student behavior using granular large scale action data from a MOOC*. arXiv preprint arXiv:1608.04789.

Research-publishing.net

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

Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022
Edited by Birna Arnbjörnsdóttir, Branislav Bédi, Linda Bradley, Kolbrún Friðriksdóttir, Hólmfríður Garðarsdóttir, Sylvie Thouësny, and Matthew James Whelpton

Publication date: 2022/12/12

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it was not under consideration for publication elsewhere. While the information in this book is believed to be true and accurate on the date of its going to press, neither the editorial team nor the publisher can accept any legal responsibility for any errors or omissions. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover photo by © 2022 Kristinn Ingvarsson (photo is taken inside Veröld – House of Vigdís)
Cover layout by © 2022 Raphaël Savina (raphael@savina.net)

ISBN13: 978-2-38372-015-7 (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 2022.