BUILDING A LEARNER CORPUS FOR CURRICULAR AND TESTING



PURPOSES

Erdal Ayan, PhD Candidate/Academic Assistant Anglistik/Computational Linguistics, Kassel University/Herder Institute, Germany E-mail: erdal ayan@yahoo.com



Introduction

- Digital text production at schools: essays, periodic writing assignments, diaries, daily records can be re-used for curriculum and assessment designs.
- Role of learning management systems (LMS): big data production, data mining and learning analytics
- Building a learner corpus may provide support for individualized learning processes.
- Learner corpus: types of words, associations, grammatical features and
- Implementations: Student performance/skill prediction, course planning, feedback production versus annotation of particular forms/structures, network visualization, similarity analysis
- Corpus-based diagnostic analysis & content-based error analysis (linguistic transfers) and creating measuring scales for testing purposes.
- Focus 1: A data processing model about how to gather, clean and annotate digital text material produced by students.
- Focus 2: Digital tools to use corpus-based data for curriculum and assessment development purposes.

Related Work

- Corpus Linguistics: appeared as a scientific method focusing on analyzing digital text. By this method, experts and teachers may have close insights into the progress of the learners, potential components of course curricula and diagnostic measuring scales for linguistic transfer cases in accordance with the proficiency level of the learners.
- Linguistic Transfer (LT): also called as Linguistic Interference or Linguistic Influence from native language (NL) to target language (TL), is a determining factor on speed and efficacy of learning processes at schools.
- Scientific research and academic discussions about LT date back to 1950s and 1960s. Although there is not a complete consensus on finding a mutual terminology about using the terms such as transfer, interference, influence, cross-linguistic influence (Gass & Selinker, 2008; Tat, 2013)
- Since the early 1950s, there has been a close interest in LT issues by the linguists (Harris, 1954; Odlin, 2003; Rosenbaum, 1949). However, there are not many research focusing on the types and levels of LT and their potential relations to other factors in especially Turkey context in terms of English language teaching at university levels.
- Ringbom (2007) has emphasized that "similarities in cultural background will facilitate foreign language learning" (39) but "differences [may] hinder [it]" (Odlin, 2003: 441) and "social background [really] matters" as Jarvis (2000) writes (as cited in Odlin, 2003: 453).
- Learning a foreign language: closely related to prior knowledge of the learners, and students are extensively dependent on their native languages while acquiring or learning a foreign language.

Problems & Purposes

- Learners assume that their perceived linguistic structures in NL may meet what they wanted to mean in the TL in terms of these LT cases. Language patterns of NL are of negative influence on the written productions in TL.
- There occurred complications in accuracy and intelligibility of their writings.
- Learners had problems and difficulties in both production and process levels of language learning.
- They mostly tended to use LT in syntactic, lexical, grammatical as well as contextual levels as defined in the Table 1.
- Purpose 1: Creation of data processing model for building a learner corpus
- Purpose 2: Annotation of LT cases existed in writing assignments
- Purpose 3: Using appropriate tools for data analysis, data visualization, digital curriculum and assessment development

Table 1: Samples for LT levels								
Levels of LT case	LT Case Samples	Reconstruction						
Syntactic [SYNX]	a) [] "Loves reading books and watching movies." [] b) [] "Their three childern." [] c) [] "Serkan Ankara lives." [] d) [] "I very miss." [] e) [] "Yozgat was born in and live there." [] f) [] "Uzungöl I went." []	 a) S/he loves reading books and watching movies. b) They have got three children. c) Serkan lives in Ankara. d) I miss [him/her] a lot. e) S/he/I was born in Yozgat and live(s) there. f) I went to Uzungöl. 						
Lexical [LEX]	a) [] "he is long length."[] b) [] "I'm speak english arabic turkish osmanlica " [] c) [] "I writing my notebook ve tea drink." [] d) [] "Military High School?? Nu bitirdi" [] e) [] "She has got three brother and two sister." [] f) [] "She is very komik" []	a) He is tall. b) I can speak English, [] and Ottoman Turkish. c) I write on my notebook and drink tea. d) He finished the military high school. e) She has got three brothers and two sisters. f) She is very funny.						
Grammatical [GRAM]	a) [] "I'm talking to Turkish." [] b) [] "I'm doing breakfast" [] c) [] "I am playing voletboll" [] d) [] "I'm reading a amasya" [] e) [] "Because I am shopping on Thursday." [] f) [] "I'm sitting in Kayseri" []	a) I speak Turkish b) I have my breakfast. c) I play volleyball. d) I study in Amasya. e) Because I go shopping on Thursdays. f) I stay in Kayseri.						
Contextual [CONTX]	a) [] "My little brother reading elementary scholl" [] b) [] "I listened to my head." [] c) [] "I am enter the course every morning." [] d) [] "I could not get my sleep." [] e) [] "He does not like people who do no	a) My little brother is a student at an elementary school. b) I took a rest. c) I attend the course every morning. d) I could not sleep. e) He does not like people who do not make mistakes.						

f) [...] "I speak very nice Turkısh." [...]

e) I speak Turkish very well.

Methodology

Corpus-based Diagnostic Analysis & Content-based Error Analysis

- Materials and Participants -> periodic writing assignments (N=848) by university students (N=115).
- Data Collection -> 8 writing tasks by Oxford English Testing Service (LMS)
- Writing Tasks -> tasks with or without help, certain context and topics such as friendship, favorite dish, booking holiday/travel, leveled from easy to difficult.
- Corrective & Motivational Feedback (See Figure 1 & Figure 2)



Figure 1: Screen-shot of an original answer by a participant for a writing task in the LMS

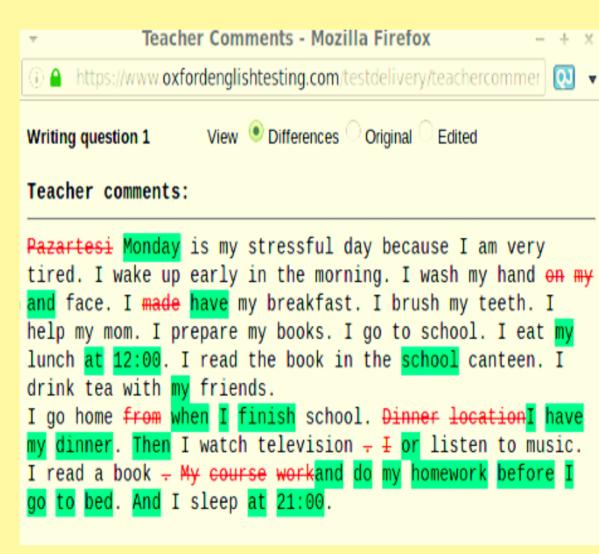


Figure 2: Screen-shot of a sample writing task with

Design and Procedure

- (i) Data (as .txt format) export from LMS,
- (ii) Data cleaning by Python NLTK and manual annotation of .txt data (see Figure 3),
- (iii) Data import into #LancsBox (Lancaster University corpus toolbox) for visualization purposes (See Figure 4),
- (iv) Creating digital curriculum & assessment environment

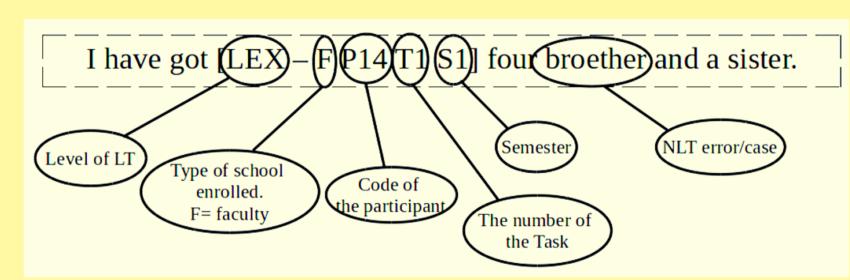
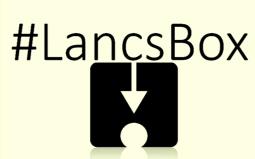


Figure 3: A sample annotated and categorized LT case





with Python NLTK



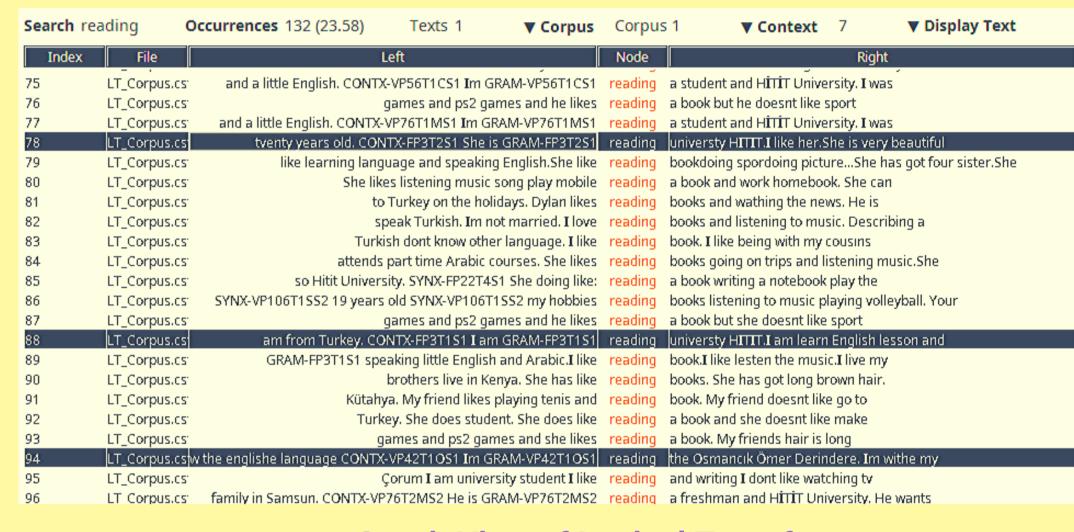


Findings

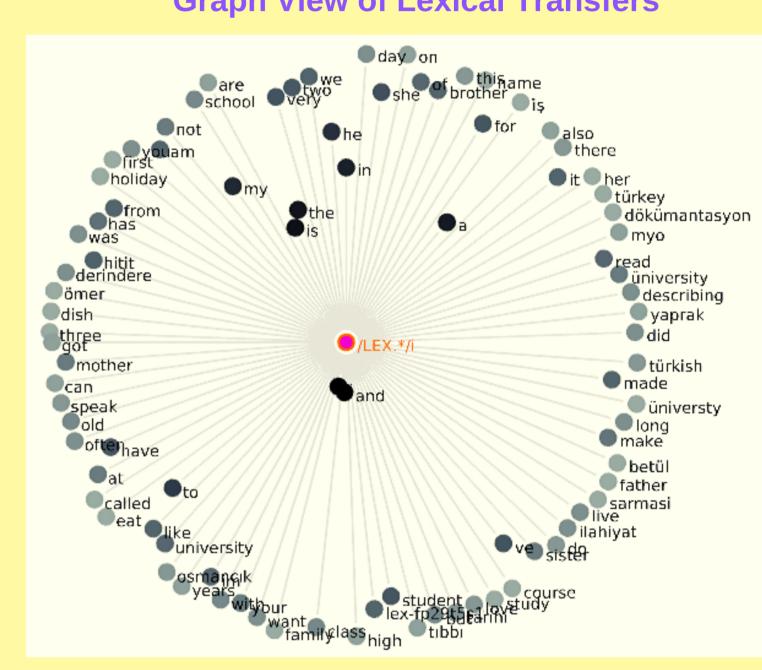
Distribution of LT Cases According to Levels

Table 2: Descriptive Statistics for LT Cases						
	Word Count	Frequency of LT Cases	Syntactic	Contextual	Grammatical	Lexical
Mean	391.7	9.157	2.617	1.322	2.148	3.070
Std. Deviation	345.5	9.437	4.139	1.770	3.234	4.565
Sum	4.505e+4	1053	301.0	152.0	247.0	353.0

Exploration of Particular Words Transferred



Graph View of Lexical Transfers



Graph-based Network View of LT Levels



Sample Cloze Test from Testing Tool



Selected References

Beel, J., Genzmehr, M., Langer, S., Nürnberger, A., & Gipp, B. (2013). A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In

Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation - RepSys '13 (pp. 7–14). Hong Kong, China: ACM Press. https://doi.org/10.1145/2532508.2532511 Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2015). Research-paper recommender systems: a literature survey. International Journal on Digital Libraries, 17(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0 Bransford, J. D., Brown, A. L. B., Cocking, R. R., Donovan, M. S., & Pellegrino, J. W. (Eds.). (2000). How People Learn: Brain, Mind, Experience, and School (Exp Sub edition). Washington, D.C: National Academies Press. Callies, Marcus, & Gotz, S. (2015). Learner Corpora in Language Testing and Assessment. Retrieved from https://benjamins.com/catalog/scl.70 Drachsler, H., Verbert, K., Santos, O. C., & Manouselis, N. (2015). Panorama of Recommender Systems to Support Learning. In F. Ricci, L. Rokach, & B. Shapira (Eds.), Recommender Systems Handbook (pp. 421–451). https://doi.org/10.1007/978-1-4899-7637-6_12 Gotz, Sandra, & Mukherjee, J. (2019). Learner Corpora and Language Teaching. Retrieved from https://benjamins.com/catalog/scl.92 Granger, S. (2011). How to use Foreign and Second Language Learner Corpora. In A. Mackey & S. M. Gass (Eds.), Research Methods in Second Language Acquisition (pp. 5–29). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781444347340.ch2/summary Fazeli, S., Drachsler, H., Bitter-Rijpkema, M., Brouns, F., Vegt, W. v d, & Sloep, P. B. (2018). User-Centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg. IEEE Transactions on Learning Technologies, 11(3), 294–306. https://doi.org/10.1109/TLT.2017.2732349 Felfernig, A. (2007). The VITA Financial Services Sales Support Environment. Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, 8. Retrieved from http://www.aaai.org/Papers/AAAI/2007/AAAI07-274.pdf Gass, S. (1996). Second Language Acquisition and Linguistic Theory: The Role of Language Transfer. In W. C. Ritchie & T. K. Bhatia (Eds.), Handbook of Second Language Acquisition (1 edition, pp. 317-345). San Diego: Academic Press. Gass, S. M., & Selinker, L. (2008). Second Language Acquisition: An Introductory Course. New York: Routledge/Taylor and Francis Group. Granger, S. (2011). How to use Foreign and Second Language Learner Corpora. In A. Mackey & S. M. Gass (Eds.), Research Methods in Second Language Acquisition (pp. 5–29). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781444347340.ch2/summary Harris, Z. S. (1954). Transfer Grammar. International Journal of American Linguistics, 20 (4), 259–270. Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 16(3), 261–273. https://doi.org/10.1016/j.eij.2015.06.005 Murphy, S. (2005). Second Language Transfer During Third Language Acquisition. Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 3 (1). Retrieved from http://journal.tc-library.org/index.php/tesol/article/viewArticle/23 Puglisi, S., Parra-Arnau, J., Forné, J., & Rebollo-Monedero, D. (2015). On content-based recommendation and user privacy in social-tagging systems. Computer Standards & Interfaces, 41, 17-27. https://doi.org/10.1016/j.csi.2015.01.004

Aguilar, J., Valdiviezo-Díaz, P., & Riofrio, G. (2017). A general framework for intelligent recommender systems. Applied Computing and Informatics, 13(2), 147–160.

https://doi.org/10.1016/j.aci.2016.08.002

O'Donnell, M. (2013). From Learner Corpora to Curriculum Design: An Empirical Approach to Staging the Teaching of Grammatical Concepts. Procedia - Social and Behavioral Sciences, 95, 571-580. https://doi.org/10.1016/j.sbspro.2013.10.684 Odlin, T. (2003). Cross-Linguistic Influence. In C. J. Doughty & M. H. Long, The Handbook of Second Language Acquisition (pp. 436-486). Malden, MA: Blackwell Pub. Ringbom, H. (2007). Cross-linguistic similarity in foreign language learning . England: Multilingual Matters Rosenbaum, E. (1949). The Application of Transfer between Foreign Languages. The Modern Language Journal, 33 (4), 287. http://doi.org/10.2307/319391 Römer, U. (2011). Corpus Research Applications in Second Language Teaching. Annual Review of Applied Linguistics, 31, 205-225. https://doi.org/10.1017/S0267190511000055 Selvi, O., & Küçüksille, E. U. (2017). Advertising recommendation system based on dynamic data analysis on Turkish speaking Twitter users. Tehnicki Vjesnik - Technical Gazette, 24(2). https://doi.org/10.17559/TV-20151020205558 Tat, G. (2013, August). Language Transfer in the Written English of Turkish Speaking Learners of English (MA Thesis). University of Westminster, London, England. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp Tono, Y. (1998). Learner corpora: Design, development and applications. 10. Vani, M. S. (2016). A Recommender System for Online Advertising. International Journal of Advanced Research in Computer and Communication Engineering, 5(2), 6. Waila, P., Singh, V. K., & Singh, M. K. (2016). A Scientometric Analysis of Research in Recommender Systems. Journal of Scientometric Research, 5(1), 71–84.

Wang, S., & Liu, C. (2013). Negative Language Transfer Reflected in ESL Learners' English Writing. In Conference proceedings. ICT for language learning (p. 492). libreriauniversitaria. it Edizioni. Retrieved from https://books.google.com/books?hl=en&lr=&id=1gZZBAAAQBAJ&oi=fnd&pg=PA492&dq=%22is+that+all+these+errors+take+place+in,+at+least, +more+than+40%25+of+the+essays,+and+misuse%22+%22two+forms+from+both+native+language+and+target+language+are+of+equivalent+translation+so+as %22+&ots=PMwNs3GUfB&sig=97ScIYZML79ZHsIiLr65VoY eJk

Wang, P. (2008). Exploring Errors in Target Language Learning and Use: Practice Meets Theory. English Language Teaching, 1 (2), 182-187.