

INDUS Meeting Tübingen 28.2.-1.3.2018:

At the Interface of Second Language Acquisition and Computational Linguistics

Tuesday, 27.2.

from 19⁰⁰ Warm-up “Wurstküche” (Am Lustnauer Tor 8, <http://www.wurstkueche.com>)

Wednesday, 28.2., Fürstenzimmer, Schloss Hohentübingen

9⁰⁰ – 9¹⁵ DETMAR MEURERS, TORSTEN ZESCH: Welcome

9¹⁵ – 10¹⁵ JAN HULSTIJN (invited): The relevance of usage-based linguistics for individualized second-language learning

10¹⁵ – 10³⁰ discussion

10³⁰ – 10⁵⁰ SIMÓN RUIZ: Aptitude-treatment effects in a web-based language learning experiment targeting English phrasal verbs

10⁵⁰ – 11⁰⁰ discussion

11⁰⁰ – 11³⁰ *Coffee break*

11³⁰ – 11⁵⁰ KATHARINA WENDEBOURG: Testing boosts grammar learning: Insights from the learning of Latin morphosyntax

11⁵⁰ – 12⁰⁰ discussion

12⁰⁰ – 12²⁰ MARIA CHINKINA, DETMAR MEURERS: Input enrichment and questions as functionally-driven input enhancement

12²⁰ – 12³⁰ discussion

12³⁰ – 13³⁰ *Lunch*

13³⁰ – 14³⁰ *joint city walk*

14³⁰ – 14⁵⁰ XIAOBIN CHEN, DETMAR MEURERS: Continuation writing shows alignment in linguistic complexity

14⁵⁰ – 15⁰⁰ discussion

15⁰⁰ – 15²⁰ AKIRA MURAKAMI: Quantification of L1 influence on L2 linguistic complexity and accuracy

15²⁰ – 15³⁰ discussion

15³⁰ – 15⁵⁰ ZARAH WEISS, XIAOBIN CHEN: Comparing the development of linguistic complexity and accuracy in L1 and L2

15⁵⁰ – 16 discussion

16⁰⁰ – 17³⁰ *Coffee break*

17³⁰ – 17⁵⁰ ANDREA HORBACH: Insights on spelling errors (working title)

17⁵⁰ – 18⁰⁰ discussion

18⁰⁰ – 18²⁰ RONJA LAARMANN-QUANTE: A prototype of an interactive tool for the automatic analysis of spelling errors in freely written children’s texts

18²⁰ – 18³⁰ discussion

18³⁰ – 18⁵⁰ HEIKO HOLZ: Tablet-Based Interventions and Assessments for German Dyslexic Primary-School Children

18⁵⁰ – 19⁰⁰ discussion

20⁰⁰ *Dinner*

Thursday, 1.3.2018, Fürstenzimmer, Schloss Hohentübingen

- 9⁰⁰ – 9⁵⁰ ELMA KERZ (invited): Recent Developments in the Language Sciences and Their Implications for Individualized Second-Language Learning
9⁵⁰ – 10⁰⁰ discussion
- 10⁰⁰ – 10²⁰ BJÖRN RUDZEWITZ, RAMON ZIAI, KORDULA DE KUTHY, VERENA MÖLLER, DETMAR MEURERS: Feedback in the interactive online workbook FeedBook
10²⁰ – 10³⁰ discussion
- 10³⁰ – 11⁰⁰ *Coffee break*
- 11⁰⁰ – 11²⁰ FABIAN ZEHNER: Unattended Consequences: How Text Responses Alter Alongside the Mode Change from PISA 2012 to 2015
11²⁰ – 11³⁰ discussion
- 11³⁰ – 11⁵⁰ KATRIN WISNIEWSKI, ISABELLE NUNBERGER, WOLFGANG LENHARD: Projekt SpraStu: “Sprache und Studienerfolg bei Bildungsausländer/-innen” (working title)
11⁵⁰ – 12⁰⁰ discussion
- 12⁰⁰ – 12²⁰ TORSTEN ZESCH: Notes from the lab
12²⁰ – 12³⁰ discussion
- 12³⁰ – 13⁴⁵ *Lunch*
- 13⁴⁵ – 14³⁵ ELENA VOLODINA (invited): Annotation of second language (L2) corpora for NLP and SLA studies: The case of SweLL
14³⁵ – 14⁴⁵ discussion
- 14⁴⁵ – 15⁰⁵ ADRIANE BOYD: Learner language normalization in reading comprehension exercises
15⁰⁵ – 15¹⁵ discussion
- 15¹⁵ – 15⁴⁵ *Coffee break*
- 15⁴⁵ – 16⁰⁵ JÜRGEN TROUVAIN: Segmental, prosodic and fluency features in phonetic learner corpora
16⁰⁵ – 16¹⁵ discussion
- 16¹⁵ – 16²⁰ RAMON ZIAI, ANDREA HORBACH: Introducing the new GSCL Arbeitskreis ”CL for Education” <http://www.gscl.org/ak-bildung.html>
16²⁰ – 16³⁰ discussion
- 16³⁰ – 17⁰⁰ DETMAR MEURERS, TORSTEN ZESCH: Wrap-up discussion & Info on next INDUS Meeting

Abstracts of invited talks

Jan Hulstijn (Universiteit van Amsterdam): The relevance of usage-based linguistics for individualized second-language learning

The linguistic and cognitive study of second-language acquisition (SLA) and bilingualism/multilingualism is a young discipline, which originated in, and quickly developed, during the first wave of the Cognitive Revolution (1960 – 1985), dominated by (1) the competence-performance distinction and Universal Grammar in generative linguistics, and by (2) notions of modularity and serial processing in psycholinguistics. In contrast, more recent work in SLA is increasingly being influenced by ideas which originated during the second wave of the Cognitive Revolution (1985 – the present), in particular usage-based linguistics (e.g., Bybee, 2013; Goldberg, 1995), Emergentism and the Competition Model (e.g., MacWhinney & O’Grady, 2015), Dynamic Systems Theory (e.g.,

De Bot, Verspoor & Lowie, 2007; Lowie & Verspoor, 2015; Larsen-Freeman, 2014), and implicit/statistical learning (e.g., Tomasello, 2003; Rebuschat, 2015; Williams & Rebuschat, 2016; Hulstijn, 2015a) and statistical learning in relation to corpus linguistics (e.g., Rebuschat, Meurers & McEnery, 2017).

These developments might be important for people conducting 'applied' work, e.g. by participants in the INDUS network. However, a complete break with top-down, form-focused, and rehearsal-oriented instructional practices in L2 instruction, based on earlier theories and research, should be avoided. In this presentation, I will give a brief overview of these more recent developments, with a particular focus on the unified conceptualization of representation and processing (referred to together with the term cognition), the notion of graded cognition (as opposed to dichotomous views of cognition), the notion of frequency and recency of linguistic elements in L2 input (with the aid of corpus linguistics), the removal of traditional barriers between lexis and grammar, and commonalities and differences among L1ers in terms of (1) degree of bilingualism, (2) literacy experiences, (3) attributes such as working memory capacity and intelligence (Hulstijn, under review). For each of these points I will give an example of how it might be relevant for people working in (computer-aided) individualized L2 instruction and learning and I will bring to the fore some desiderata for future work in this area of applied linguistics. If time permits, I will also briefly present my views on the CEFR (Hulstijn, 2011, 2015b, Chapter 10).

- Bybee, J.L. (2013). Usage-based theory and exemplar representations of constructions. In T. Hoffmann & G. Trousdale (eds.), *The Oxford Handbook of Construction Grammar* (pp. 1-24 [e-book download]). Oxford, UK: Oxford University Press.
- De Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and cognition*, 10(1), 7-21.
- Ellis, N. C. (2002). Frequency effects in language processing and acquisition. *Studies in Second Language Acquisition*, 24, 143-88. doi:10.1017/S0272263102002024
- Ellis, N.C., Römer, U., Brook O'Donnell, M. (2016). Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of Construction Grammar. *Language Learning*, 66, issue S1
- Elman, J.L. 1999: The emergence of language: a conspiracy theory. In MacWhinney, B., editor, *The emergence of language* (pp. 1-28). Mahwah, NJ: Erlbaum.
- Goldberg, A. E. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago, IL: University of Chicago Press
- Hulstijn, J.H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, 8, 229-249. doi: 10.1080/15434303.2011.565844
- Hulstijn, J.H. (2015a). Explaining phenomena of first and second language acquisition with the constructs of implicit and explicit learning: The virtues and pitfalls of a two-system view. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 25-46). Amsterdam: John Benjamins.
- Hulstijn, J. H. (2015b). *Language proficiency in native and non-native speakers: Theory and research*. Amsterdam: John Benjamins.
- Larsen-Freeman, D. (2014). 12 Complexity Theory. *Theories in Second Language Acquisition: An Introduction*, 185.
- Lowie, W., & Verspoor, M. (2015). Variability and variation in second language acquisition orders: A dynamic reevaluation. *Language Learning*, 65(1), 63-88.
- MacWhinney, B., & O'Grady, W. (Eds.) (2015). *The handbook of language emergence*. Chisester, UK: Wiley.
- Rebuschat, P. (Ed.) (2015), *Implicit and explicit learning of languages*. Amsterdam: John Benjamins.
- Rebuschat, P., Meurers, D., & McEnery, T. (2017). Language Learning Research at the Intersection of Experimental, Computational, and Corpus-Based Approaches. *Language Learning*, 67(S1), 6-13.

- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
- Williams, J., & Rebuschat, P. (2016). *Implicit learning and second language acquisition*.

Elma Kerz (RWTH Aachen): Recent Developments in the Language Sciences and Their Implications for Individualized Second-Language Learning

Due to the influential direct impact of Chomskyan (generative) perspectives on the language sciences and their indirect effects (cf. Chomsky's hidden legacy, Christiansen and Chater, 2016a), much second language (L2) acquisition research has tacitly adopted the notions of an 'ideal native speaker-listener competence' and 'ultimate attainment' and has assumed that native speaker competence is a convergent, single benchmark for L2 performance and target for L2 development. Furthermore, the preoccupation with the search for universal characteristics, processes and developmental stages of language acquisition may have inhibited the growth of individual differences (IDs) research.

In recent years, however, there has been an explosion of studies uncovering substantial IDs across multiple components of language across the lifespan in native speakers (see, e.g., Dabrowska and Street, 2006, Dabrowska, 2012; Montag and MacDonald, 2015; Brysbaert et al., 2016; see, e.g., Kidd, Donnelly and Christiansen, 2018 for a recent overview). This is in line with an increasing interest in 'emergentist' accounts of first language (L1) and L2 learning – umbrella terms for a family of models, such as usage-based, complex dynamic system, constraint-based and connectionist models – that all have highlighted the experientially adaptive nature of language knowledge (cf., e.g., Ellis & Larsen-Freeman, 2006; Beckner et al. 2009; McClelland, et al. 2010; Larsen-Freeman, 2011; MacWhinney, 2012; Ambridge & Lieven, 2015; Christiansen & Chater, 2016). These accounts put the emphasis on usage and/or experience with language input and assume a direct and immediate relationship between processing and learning, conceiving of them as inseparable rather than governed by different mechanisms. In these accounts, the nature and extent of individuals' experience with language input are key determinants of language development. Both L1 and L2 learning are conceived of as a continuous process which does not end at some discrete point in time in ontogenetic development but instead takes place across the lifespan (cf. 'learning-as-processing' assumption, e.g. Chang, Dell and Bock, 2006). In other words, language learning does not result in the establishment of a static knowledge system. Rather, as long as there is exposure to linguistic input, an individual's knowledge of a language is in constant flux. Correspondingly, these accounts render the notion of 'ultimate attainment' superfluous and view language learning as a dynamic process of updating patterns rather than a stage-like progression. While models of representation focusing on static competence ('property theories') have traditionally been kept distinct from models of acquisition and use focusing on mechanisms underlying language learning and processing ('transition theories'), emergentist approaches usefully conflate a property theory with a transition theory (cf. Ellis and Larsen-Freeman, 2006). They typically seek to account for language learning in terms of general cognitive mechanisms rather than internalized linguistic constraints (cf. Christiansen & Chater, 2008, 2016). One such mechanism that has received a lot of attention recently is that of 'statistical learning' (SL), succinctly defined as the human ability to detect and keep track of distributional regularities and patterns inherent in language input at multiple levels of linguistic representation (cf., e.g., Armstrong, Frost & Christiansen, 2016 and contributions therein). A growing body of research has

demonstrated a close relationship between IDs in SL ability and variability in native language learning in both child and adult populations (see Conway et al., 2010; Misyak and Christiansen, 2012; Siegelman and Frost, 2015; Kidd and Arciuli, 2016), and in adult L2 populations (see Frost et al., 2013; Ettliger et al., 2016; Onnis et al., 2016).

The developments in research on language learning outlined above are driven, among other things, by the existence of corpora of naturalistic data and new statistical and computational methods to analyze such data. For researchers working from an emergentist perspective, analyses of learner production data supported by natural language processing (NLP) techniques are central to assessing language proficiency, performance and development, and an important complement to experimental research (see, Christiansen and Chater, 2016; see also Ellis, Rebuschat, Meurers and McEnery, 2017).

In the first half of the talk I will use some examples from my own work to show how our understanding of language learning and language development can be advanced through a stronger focus on (1) IDs in both L1 and L2 abilities and their interrelationships and (2) shared rather than separate mechanisms underlying L1 and L2 learning. In the second part of the talk I will address how individualized language learning systems can be employed not only in pursuit of applied goals, such as tracking and accelerating the development of language skills, but also how the dense longitudinal data ('learning trajectories') generated by such systems have the potential to shed new light on language development (cf. 'input-output modeling' approach).

- Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system. Position paper. *Language Learning*, 59, 1–26.
- Brysbaert M, Stevens M, Mandera P, & Keuleers E. (2016) How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*. 7:1116.
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113(2), 234–272.
- Christiansen, M. H., & Chater, N. (2016a). *Creating language. Integrating evolution, acquisition, and processing*. MIT Press.
- Christiansen, M. H., & Chater, N. (2016b). The now-or-never bottleneck. A fundamental constraint on language. *Behavioral and Brain Sciences*, 39, 1–72.
- Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing. Word predictability is the key. *Cognition*, 114(3), 356–371.
- Dabrowska, E. (2012). Different speakers, different grammars. Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, 2(3), 219–253.
- Dabrowska, E., & Street, J. (2006). Individual differences in language attainment. Comprehension of passive sentences by native and non-native English speakers. *Language Sciences*, 28(6), 604–615.
- Ellis, N. C., & Larsen-Freeman, D. (2006). Language emergence. Implications for applied linguistics—introduction to the special issue. *Applied Linguistics*, 27(4), 558–589.
- Ettliger, M., Morgan-Short, K., Faretta-Stutenberg, M., & Wong, P. (2016). The relationship between artificial and second language learning. *Cognitive Science*, 40(4), 822–847.
- Frost, R., Siegelman, N., Narkiss, A., and Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science*, 24(7), 1243–1252.
- Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, 87(1), 184–193.
- Kidd, E., Donnelly, S. & Christiansen, M. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, 22(2), 154–169.
- MacWhinney, B. (2012). The logic of the Unified Model. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 211–227).

- McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8), 348–356.
- Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8- and 12-year-old children and adults. *Journal of Experimental Psychology: General*, 144(2), 447–68.
- Onnis, L., Frank, S. L., Yun, H., and Lou-Magnuson, M. (2016). Statistical learning bias predicts second- language reading efficiency. In Papafragou, A et al. (Ed.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pages 2105–2110). Austin, TX: Cognitive Science Society.
- Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability. Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, 81, 105–120.

Elena Volodina (Göteborgs universitet): Annotation of second language (L2) corpora for NLP and SLA studies: The case of SweLL

Metadata and annotation in L2 corpora is the place where linguistics, pedagogy (and partly sociolinguistics) "hides" in NLP nowadays. This talk will draw on the general L2 community needs of metadata, reasoning around error annotation versus "can-do" annotation, and tie those to the SweLL project. The latter part would cover annotation, metadata, tools, formats, reliability dimension, as well as legal aspects. Given the presence of linguistics and pedagogy in the corpora, machine (and deep) learning experiments can yield nice results, both with relevance to automated error detection/annotation/correction, linguistic annotation of (automatically corrected) L2 texts, as well as with regards to building applications on top of this data. However, as the practice shows, tools and algorithms decay, while data stays, so the way the data is prepared is a key to what is possible to do with is later - be it SLA research questions or NLP applications. I will demonstrate an error-annotation tool that we are developing in the project and, if time allows, showcase some tools developed in our group based on L2 data, such as for grading learner texts (TextEval) and for level-grading lexical complexity of unseen vocabulary (Siwoco).