Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration with Computational Linguistics

Detmar Meurers Markus Dickinson

Accepted for publication in *Language Learning, Special Issue on Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and Interpretation*, to appear 2017.
(Draft of October 3, 2016; rev. 14482)

Abstract

The paper discusses two types of opportunities for interdisciplinary collaboration between computational linguistics and language learning research. We target the connection between data and theory in second language research and highlight opportunities to i) enrich the options for obtaining data, and ii) support the identification and valid interpretation of relevant learner data.

We first characterize options, limitations, and potential for obtaining rich data on learning: from web-based intervention studies supporting the collection of experimentally controlled data to online workbooks facilitating large-scale, longitudinal corpus collection for a range of learning tasks and proficiency levels. We then turn to the question of how corpus data can systematically be used for second language research, focusing on the central role that linguistic corpus annotation plays in that regard. We show that learner language poses particular challenges to human and computational linguistic analysis and requires more interdisciplinary discussion of analysis frameworks and advances in annotation schemes.

1 Introduction

Under a computational linguistic (CL) perspective on the connection between data and theory in second language research, there are at least three types of opportunities for interdisciplinary collaboration: Firstly, the collaboration can *enrich the options for obtaining data* of interest to second language research. Secondly, computational linguistic methods can be applied to *identify relevant learner language in corpora and support its valid interpretation*. And thirdly, it can support the *computational modeling of the acquisition process*.

Before we can make those opportunities concrete, let us sketch some key aspects of data and research questions in second language research in order to establish an explicit target for CL research to connect to in order to become relevant.[1] Data in second language research includes a wide range of *online and offline measures of learner behavior*, ranging from experimental data on eye tracking to texts written by learners. To interpret the learner behavior, one also needs information on the *nature of the tasks* eliciting the data, measures of *individual differences between learners*, and the nature and presentation of the *input learners received*. For research targeting instructional interventions, the different types of *instructions*, *feedback*, and *sequencing of material* also constitute relevant parameters for the interpretation of learner behavior in second language research. On this empirical basis, research in Second Language Acquisition (SLA) investigates the *development of language competence*,

including the ability to use a range of language *forms* to express an intended *meaning* or achieve a *functional goal*. This involves linguistic means as well as extra-linguistic abilities such as *strategic competence* (in the broad sense of Bachman 1990). Research targets both the development of general language characteristics, such as Complexity, Accuracy, and Fluency (CAF, Larsen-Freeman, 1978, Housen and Kuiken, 2009), and of specific linguistic properties from all domains of linguistic modeling (lexicon, morphology, syntax, semantics, pragmatics). In language testing research, learner data is analyzed to support valid inferences about a learner's *proficiency*. The focus there is on eliciting learner data in a way that is maximally informative with respect to the targeted constructs and independent of other factors. In instructed SLA and language teaching practice, the analysis of learner data supports formative and summative feedback, advances the understanding of student abilities and needs, and helps test the effectiveness of interventions.

Keeping this range of relevant data and goals of second language research in mind, in the next two sections we consider two areas in which computational linguistics can help enrich and interpret the empirical dimension of second language research. The third area, computational modeling of the acquisition process, provides additional opportunities for interdisciplinary collaboration at the intersection of statistical learning, first and second language learning, and computational (linguistic) modeling. Complementing the longstanding research on statistical learning and language acquisition at the intersection of cognitive psychology, linguistics, and computer science (cf. Rebuschat and Williams, 2012, and references therein), there is a recent strand of work also integrating computational linguistic methods. This includes foundational research in usage-based linguistics (Ellis et al., 2016), work in the Association for Computational Linguistics (ACL) workshop series on *Cognitive Aspects of Computational Language Learning*[2], computational linguistic approaches to grammar induction (Heinz et al., 2015), as well as research integrating specific computational linguistic techniques, such as distributional semantic analysis (Alikaniotis and Williams, 2015). Modeling the learning itself raises the fundamental question of which linguistic representations need to be explicitly modeled and which fall out from the learning mechanism (cf., e.g., Baayen et al., 2016). Given space limitations, in this article we focus on the other two opportunities for collaboration supporting evidence and interpretation in language learning research: enriching the options for obtaining data and grounding valid interpretation of corpus data through corpus annotation.

1.1 Enriching the options for obtaining data

The data empirically grounding theory and practice in second language research can be obtained in a variety of ways, which can be characterized on a spectrum ranging from fully controlled experiments on the one end, via systematic collections of learner productions in carefully designed tasks, to collections of learner data arising as a side product of language teaching on the other end. These types of data differ with respect to how they can help address research questions. Data in *experimental research* typically is obtained to investigate a concrete, highly focused research hypothesis. Data collected in *instructional or naturalistic settings*, on the other hand, is more commonly collected in the hope of supporting a range of inquiries, from the practical interests of language instructors to more general research agendas in SLA.

The language data collected in corpora is shaped by a variety of factors, including: i. the *linguistic characteristics* of the language used, ii. the properties of the *speaker or writer*,

iii. the language *tasks* or purpose from which the data was collected, and iv. the nature of the *world, culture and society* talked and written about. Whether a data set can be used to investigate a particular research question depends on whether the data set is *representative* of the domain of language use to be studied (cf., e.g., Biber, 1993). No corpus, i.e., no language data set systematically collecting spoken or written language, can be universally representative of language—one can only aim for representativeness with respect to the characteristics required to support a general strand of inquiry.[3] Still, compared to data from experimental methods that by design zoom in on a specific hypothesis using few, carefully designed empirical items, corpora can be collected in a way as to make them relevant for a broader range of research questions, only indirectly constrained by the need to ensure that the data is sufficiently representative to support the general research perspective.

So how can computational linguistics help enrich the options for obtaining data for second language research? Starting from the more experimental research side, the automatic analysis of language using Natural Language Processing (NLP) can support the collection of data in full-fledged intervention studies. This includes different types of interventions, from visual input enhancement using NLP to analyze and enhance the input learners read (e.g., Meurers et al., 2010) to tutoring systems providing immediate, individualized feedback to learners based on NLP analysis of learner output (e.g., Heift, 2003; Nagata, 2009; Amaral and Meurers, 2011). While research in Intelligent Computer-Assisted Language Learning (ICALL, cf. Heift and Schulze, 2007) integrating NLP into CALL has traditionally been focused on the NLP as such and its internal evaluation, attention to authentic learner data and real-life learning is increasing (cf., e.g., the ACL workshop series on *Innovative Use of NLP for Building Educational Applications*[4]). At least in interdisciplinary efforts combining expertise in NLP and second language research, it is becoming possible to conduct tests on the methods of automatic analysis with authentic learner data and to externally evaluate the learning outcomes supported by an ICALL-mediated intervention.

Showcasing the potential, a couple of SLA studies collect and interpret learner data in ICALL systems in terms of learning outcomes. Petersen (2010) created a dialogue system offering information gap activities and used it to examine the developmental effects of recast-intensive interaction on ESL question formation and morphosyntactic accuracy. In a similar spirit, Wilske (2015) developed a dialogue system with a map task and an appointment scheduling task designed to support the acquisition of case marking and verb order in German. In these systems, computational linguistic methods are needed to analyze the learner productions in the system, both in terms of meaning to maintain the dialogue and in terms of form to provide incidental focus-on-form feedback.

Note that CALL environments without automatic NLP analysis also support some types of intervention studies and some such research provides an explicit link to SLA, such as the study reported by Zhao et al. (2013) couched in the Competition Model (MacWhinney, 1997). Leow et al. (2015) provide a current overview of studies, meta-analyses, and important methodological issues related to CALL, which in general are equally applicable to ICALL research. Without NLP analysis, however, CALL-based intervention studies must either a) be based on exercises that are constrained enough to pre-envisage all possible (correct or incorrect) learner responses and the corresponding feedback (cf., Meurers, 2012), or they must b) rely on humans to produce the enhanced input materials, provide the learner feedback, or manually code the distinctions relevant for analysis in the intervention study. The recent collection of technology-based TBLT studies by González-Lloret and Ortega (2014) shows that given the significant overhead of manual coding, studies more often than not resort to ques-

tionnaires and standardized tests in place of an analysis of the characteristics of the learner language as such.

In terms of enriching the options for obtaining data for second language research, the added value of the NLP in ICALL compared to regular CALL thus amounts to supporting data collection in a broader range of tasks (Quixal and Meurers, 2016), and to do it in a way that in principle can be scaled up to any number of learners and analyzed without human intervention or coding. The ability to scale up arguably is particularly important to obtain reliable results in empirical educational research on second language learning, where interventions supporting real-life language learning require long exposure to obtain differences in learning outcomes, individual differences are prominent, and the expected relative effect size of an intervention to be tested may be small (but cf. Plonsky and Oswald, 2014).

Independent of the type of learning environment and analysis techniques used, learning outcomes can be measured in a standard pre-/post-test setup. Given that computer-based systems support the logging of every learner interaction with the system, one can obtain a detailed record of the learner input, interaction, and output, which can be collected in a task-based learner corpus. Logging can also provide an incremental record of any other learner behavior, which has led to the rise of Educational Data Mining and Learning Analytics as new fields of study (Romero and Ventura, 2013; Baker and Inventado, 2014). We here limit ourselves to the interpretation of language data, i.e., the language input and output of learners collected in corpora. The explicit task context in which such corpora are collected facilitates the rich representation of the targeted language constructs and supports a more reliable interpretation of both form and meaning of learner data (Ott et al., 2012). The availability of task-based corpora and rich interaction logs thus nicely complements the evaluation of learning outcomes through explicit pre-/post-tests. Combining the virtues of focused experimental evaluation and broad learner data collection, task-based, incremental records of learning make it possible to tap into the full inventory of SLA techniques for analyzing learner language (Ellis and Barkhuizen, 2005), including measures of linguistic complexity and accuracy and the emergence and frequency of specific linguistic properties.

The opportunities and challenges of large-scale data on language learning can be made concrete through the example of the EFCAMDAT corpus (Geertzen et al., 2013), which is at the heart of the contribution by Alexopoulou et al. (submitted) exploring the connection between Task-based Language Teaching, learner corpora, and NLP methods. The corpus collects data from Englishtown, a web-based CALL system integrating manual human feedback to learners on a massive scale (the second release contains over 1,800,000 texts written by over 174,000 learners). Compared to dedicated experimental setups designed to elicit particular constructions, Englishtown covers a broad curriculum for learners at all stages (A1–C2), which shifts the burden from experimental study design and data collection to the identification and interpretation of the relevant data subsets and analysis of their characteristics in the vast amount of data that has been collected—to which we turn now.

1.2 Supporting the identification and interpretation of relevant data

Given a corpus of learner language and a second language research question, the essential question is how to connect the two: How can the subsets or instances of data that address the given research question be identified in the corpus?[5]

Reading through the corpus from beginning to end is only an option for relatively small corpora and when few language characteristics need to be identified. Turning from such man-

ual expert analysis to the most shallow computer-supported option, searching for concrete words or phrases can be an effective solution for some research questions. Yet most linguistic research questions are formulated in terms of abstractions and generalizations rather than the surface forms directly accessible in the corpus. So the effective identification of the relevant subsets of data requires annotation of the corpus.[6]

For corpus annotation to support research, it needs to be based on the same notions in which the research questions are formulated, or it must be possible to translate between the two (cf., e.g., Meurers, 2005), and the annotation must be discriminating and fine-grained enough for the research question at hand. Given the general nature of linguistic annotation, it can support the identification of relevant classes of data for a range of research questions—yet it will generally not answer a specific research question directly. Instead, corpus searching provides access to sets of examples, which then need to be interpreted quantitatively or qualitatively in light of the research question.

So how can the linguistic annotation crucially needed to connect research questions and corpus data be obtained for learner data? For corpora to contain a sufficient number and breadth of data that is relevant to address a given SLA research perspective, they generally need to be of a size that makes completely manual annotation impractical or impossible. In general, *the effective use of large-scale learner corpora in SLA research thus fundamentally depends on NLP methods supporting the automatic annotation of learner corpora.* We therefore consider the discussion of appropriate annotation schemes and NLP tools for annotating learner corpora to be an essential area of collaboration between SLA and computational linguistics – and, correspondingly, the related conceptual issues are at the heart of our contribution to this interdisciplinary special issue.

While it is superficially tempting, one cannot simply reuse the existing annotation schemes and NLP tools, which were developed for edited native language. The language produced by learners systematically differs from the idealized native language competence that linguistic category systems in theoretical linguistics are designed for (cf., e.g., Dickinson and Ragheb, 2015). If one directly reuses annotation schemes and NLP tools developed for native language, there is a clear danger of relying on analyses that are not valid interpretations of learner language. At the same time, given the significant form variability of learner language, answering the question of how one can obtain a systematic index into the set of learner data relevant under a particular research perspective is of even greater importance.

In other application fields, such as NLP for social media analysis, text normalization is used for "reducing linguistic noise" (Farzindar and Inkpen, 2015, p. 17). The idea there is to map unusual spellings and expressions found in social media texts to the edited native language norm, which one knows how to analyze. This is a sensible approach for social media analysis because its major goal is to extract or summarize information (opinions, events, topics, locations, ...), not to characterize the nature of social media language (though this issue is of some interest in sociolinguistics). In contrast, SLA research by definition is *investigating the characteristics of learner language*. Annotating learner corpora by naively "normalizing" learner language into native language so that annotation schemes and NLP tools developed for edited news texts can be used to annotate learner corpora thus would amount to eliminating the characteristics of learner language that SLA research needs to be able to identify.[7]

NLP tools are generally designed to be robust to irrelevant variability in the data, but robustness is only meaningful together with a clear conceptualization of what target one is robust about and why that is appropriate for learner language. Depending on the goal of

the analysis, i.e., the underlying research question, NLP tools should abstract over specific variability and ambiguity of learner language to support generalizations (e.g., determining a syntactic function even if a word is misspelled), yet make it possible to pinpoint it in other cases (e.g., identifying an interpretable but distributionally inappropriate use of a preposition in place of a conjunction). As a result, multiple levels of annotation are needed to address a broad range of research questions.

To explore which distinctions from language learning research need to be annotated, which clusters of observations emerge from the collections of learner data, and how the known and emerging, theory- and data-driven classes can be fully characterized and automatically identified in learner corpus data, taking into account task and learner information, interdisciplinary collaboration between SLA and Computational Linguistics is essential. Annotation schemes and NLP tools can then be redesigned and retrained to identify the characteristics of relevance to SLA research and, where needed, to parcel out the different types of evidence for a particular linguistic category rather than combine them into a single-best guess (e.g., Díaz Negrillo et al., 2010; Seiler, 2013).

Importantly, corpus annotation is not only needed to allow researchers to search and zoom into relevant subsets of learner data. Corpus annotation and NLP tools providing learner language analyses are equally important for determining general measures of language competence and development, such as analyses of linguistic complexity as part of the CAF triad (e.g., Housen and Kuiken, 2009). Linguistic complexity can be analyzed at all levels of linguistic modeling and language use, and NLP is increasingly being integrated to support the automatic computation of complexity measures, including aspects of the lexicon (e.g., Malvern et al., 2004; McCarthy and Jarvis, 2010; Lu, 2012; Kyle and Crossley, 2015), syntax (e.g., Covington et al., 2006; Lu, 2010), and discourse (e.g., Graesser et al., 2004). While potentially supporting a broad range of complexity analyses at an unprecedented scale, the numeric scores the tools produce make it even easier to overlook whether the underlying analyses are valid interpretations of the learner language. Even a traditional measure of lexical variation such as the type-token ratio requires a systematic annotation of orthographically normalized forms since otherwise we will not only obtain high lexical variation scores for advanced learners using a rich vocabulary but also for beginning learners producing orthographically varied realizations of the few words they know.

In addition to the fundamental questions around normalization and the nature of the categories appropriate for learner language, the quality of the NLP analyses is directly influenced by the proficiency of the learners who wrote the texts. Yet the goal of the complexity analyses is to learn about proficiency development, making it relevant to consider potential interactions between the quality of the analyses and the proficiency level of the texts analyzed. For example, NLP tools systematically rely on punctuation to segment a text into sentences. Yet punctuation in learner language can vary significantly depending on the proficiency level (and other factors, such as the task the text was written for). So, complexity analyses based on NLP tools designed for native language will mistakenly determine that beginning learners produce long sentences. A systematic discussion of how NLP preprocessing (tokenization, sentence segmentation), normalization, annotation schemes, and NLP techniques need to be adapted for learner language is crucially needed at this point – a discussion we want to foster here in this interdisciplinary issue, given that it requires expertise from both SLA and NLP.[8]

A second area providing an increasingly rich set of NLP approaches is Native Language Identification (NLI), where the goal is to identify the native language of a non-native writer for each document of a corpus. The recent NLI shared task (Tetreault et al., 2013) has in-

creased interest in the NLP community, and the best classifiers achieve accuracies over 80% accuracy in distinguishing 11 native language classes. At the same time, the quantitatively best approaches (cf. Massung and Zhai, 2015) are based on thousands of shallow features (such as word and part-of-speech n-grams), which do not lend themselves to a qualitative interpretation. Fostering more interdisciplinary discussion in this area stands to make the results of the computational linguistic research more relevant to SLA research on L1 transfer.

Having motivated the need for an interdisciplinary discussion in this area, the next sections of this paper explore how learner data can be annotated to support its effective use for second language research—and which challenges annotation schemes and NLP tools automating annotation need to tackle to be able to support this line of research. Section 2 spells out the conceptual background for using corpora before we investigate in detail the nature and role of annotation in section 3 and how the particular challenges that learner language poses can be addressed. On this basis, a central goal of this paper is to make the case that annotated corpora and the NLP techniques supporting collection and analysis can empirically enhance and strengthen second language research.

2 Linking Research Questions and Corpus Data

Given that a corpus is not a list of examples or experimental results addressing a particular research hypothesis but a general collection of data, we need to clarify how corpora can become relevant for research into language, and, more specifically here, second language research.

Corpora minimally provide the language *forms* themselves and possibly *annotation* of the data. Annotation may be on the level of the corpus contents, such as part-of-speech (POS) annotation for each word in the corpus, or the annotation may provide metadata (e.g., information concerning the task, the writer, etc.). The language forms and content-level annotation allow one to identify specific instances of language use, akin to what Saussure called *parole* (de Saussure, 1916); from this, one can attempt to generalize to the principles of the language—Saussure's *langue*.

When querying a corpus, a linguistic researcher needs to be able to access specific subsets of data that are relevant for testing or establishing generalizations (cf. langue) over particular forms (cf. parole). Consider, for example, the research of Clahsen and Muysken (1986) into the acquisition of German word order by native speakers of Romance languages. Sentences of the form shown in (1a) and (1b) are interpreted as characteristic of the Stage 2 and Stage 4 patterns in the general development of German word order, summarized in (2), and are annotated here with subscripts to indicate how they instantiate the general patterns.

(1) a. Früher ich kannte den Mann
earlier$_{AdvP}$ I$_S$ knew$_V$ [the man]$_O$

b. Früher kannte ich den Mann
earlier$_{AdvP}$ knew$_{V[+fin]}$ I$_S$ [the man]$_O$

(2) a. *Stage 1:* S (Aux) V O

b. *Stage 2:* (AdvP/PP) S (Aux) V O

c. *Stage 3:* S V[+fin] O V[-fin]

d. *Stage 4:* XP V[+fin] S O

e. *Stage 5:* S V[+fin] (Adv) O

f. *Stage 6:* *dass* S O V[+fin]

The relevant point here is that the stages are characterized in terms of specific language forms (*dass*) and linguistic abstractions, namely syntactic categories (e.g., V[+fin]) and functions (e.g., O for object).

Considering the intertwined perspectives of what researchers aim to identify in corpora and what corpora are comprised of, in unannotated corpora the success of a corpus search depends upon how well the linguistic abstraction can be approximated by the explicit word forms in the corpus.[9] Stringer (2015, p. 118) is a good example for an SLA study approximating a complex linguistic phenomenon—embedded *wh*-questions—in terms of the surface forms found in an unannotated corpus. After introducing some general notions related to corpus searches approximating linguistic patterns, we focus on learner language as the subject of second language research and the specific challenges it poses, and then proceed to annotated corpora in section 3.

2.1 Precision and recall in unannotated data

Consider studying the development of relative clauses in L2 acquisition. To identify relative clauses in corpora, we need to approximate this general, abstract linguistic characterization in terms of the specific language forms, for example by by referring to the set of forms that can introduce relative clauses, such as *that*. The success of the resulting corpus query can be evaluated in terms of precision and recall, just like other information retrieval tasks (cf., e.g., Manning et al., 2008, ch. 8). *Precision* refers to the percentage of items returned for a query that indeed are instances of what one is searching for, whereas *recall* identifies how many of the targeted items in the corpus are actually retrieved by the query.

An immediate problem for precision that arises when approximating relative clauses using queries making reference to *that* is the ambiguity of this form. As illustrated by the examples below, from the EFCAMDAT learner corpus (Geertzen et al., 2013), *that* can introduce a relative clause (3a)—yet, the precision of a query simply referring to *that* will be low given that it will also retrieve examples without relative clauses, such as uses of *that* in a complement clause (3b), as a demonstrative (3c) or as a determiner (3d).

(3) a. The first one I would like to recommend is a property **that** you will see anywhere in the world.
b. I am sorry to hear **that** you have shoping addiction.
c. Hi, I am gald to hear **that**, but I am busy ...
d. I will take my break **that** day.

Recall is affected by a) being able to pinpoint the exact set of relative pronouns (*which*, *whomever*, etc.), including: b) innovative relative pronouns used in learner language, such as the use of *what* identified by Alexopoulou et al. (2015, sec. 3.1), and, most crucially: c) zero relative clauses, ones with no relative marker (e.g., *the book [I like]*). The last kind are particularly problematic, as a form-based search cannot generally identify something not present[10]—and Alexopoulou et al. (2015) found that zero RCs made up 15% of a manually-annotated sample of data.

In short, to identify linguistic abstractions based on surface forms, the word forms and sequences thereof, means having to spell out every form comprising the generalization—an

impossible task for patterns varying in their syntax. In section 3, we discuss how annotation can help address some of these issues.

2.2 Interpreting learner data given variability and ambiguity

One of the major issues affecting precision and recall is *form variability*—an issue that becomes particularly complex for learner language since it depends on characteristics of the learner (e.g., L1, proficiency) and orthography is more variable. Well-formed variability is common in language, from alternative spellings for different dialects (e.g., *colour*/*color*) to non-standardized spelling in historical documents (see, e.g., Baron and Rayson, 2008) or for languages and varieties without standardization (e.g., in social media, cf. Baron and Rayson, 2012) and transliterated terms (e.g., there are over 100 ways to spell *Muammar Qaddafi* according to ABC News, http://goo.gl/aFMZWf), not to mention lexical choices reflecting sociolinguistic variables (e.g., *night*/*nite*). Moving beyond variability in orthography, language offers variability in morphology (e.g., *dived*/*dove*), syntax (e.g., word order variants), and meaning (e.g., synonyms).

Compounding the problem are language samples featuring novel instances of variability, where the variability in a form is either not a generally accepted part of the language in question or is of a type that does not follow from the grammatical principles of the (target) language. The ESL learner examples from the SALLE corpus (Ragheb and Dickinson, 2011) below illustrate variability in orthographic form (4a), morphological form (4b), lexical or lexico-syntactic meaning (4c), and the way the form of a sentence and its meaning relate (4d), in this case obtaining a copula meaning without the copula.

(4) a. . . . like body that will **loss** its **ballence** when one of two **organe** get **dameged**, a country also can **loss** its **ballence** when part of its people live in sadness.
b. ... **peoples** who I met is not good ...
c. So when I **admit** to korea university, I decide what i find my own way.
d. Also, the people in it very friendly.

Learners may also phrase sentences in atypical ways, e.g., learners using *pushing leaves* in place of *raking leaves* (King and Dickinson, 2013). Given the systematic presence of well-formed, ill-formed, and atypical variability in learner language, a form-based search would have to spell out all potential forms for the general pattern to be captured. While it is possible to extensionally spell out some queries in this way (cf., e.g., Stringer, 2015), the more the variability, the harder and more error-prone the task is. And where discovering the forms used by a learner and characterizing the nature of the variability actually is part of the research question (in variationist linguistic terms (Tagliamonte, 2011): discovering the variants used to express an underlying variable), spelling things out in this way is not a viable option at all. Spelling out all potential variants in the corpus query also is not an option when the set of variant forms depend on the L1 of the writer in a corpus where texts from writers with different L1s are collected; in other words, when the patterns to be searched for would have to be spelled out differently for different parts of the corpus.

Lexical (transfer) errors also illustrate how simply expanding a query to handle variants is infeasible. In (5), for example, a native Spanish speaker uses *until* to denote a time point at which at event starts (cf. *at* or *when*). If one were to expand a query to find all such cases, it would wrongly identify the uses of *until* elsewhere, even those written by speakers

of other L1s. Expanding out queries is thus not really an option when the expansion needs to be different for different parts in the corpus one is querying.

(5) ... the second [goal] is to buy an own house **until** twenty seven years old ...

Variability also makes the forms harder to interpret, increasing *ambiguity*. For example, the use of *loss* as *lose* in (4a) increases the ambiguity of the form *loss*, which can now function as a verb, in addition to its common noun uses. Ambiguity also presents issues for form-based searching in that it leads to lower precision unless more context is specified. In the above example, a search for verbs based on explicit forms would have to not only search for *lose*, but also *loss* in a verbal context (e.g., after a pronoun). Without such a defining context, a search for *loss* would result in far too many noun instances.

Variability and ambiguity as two characteristics of learner language directly affect corpus searches: variability means that one may have to spell out *many possible forms*, and ambiguity means that one may have to spell out *every possible relevant context* in which a particular form can appear.

Underlying the discussion of variable, ambiguous and innovative forms in learner data are questions about the interpretation of learner language data in general. In learner language, variability is a particular challenge given that there is less of a standard mapping between forms and meanings compared to the system established for native language, and the individual differences and interlanguage development makes the construction of a complete language system over a set of speakers and proficiency levels difficult to establish.

Consider searching for English passive participles. Given the established form-meaning mapping in English, searches for words ending in *-ed*/*-en* appearing after forms of *be* can be used to approximate searches for passives. While this search may have decent precision and recall for native language, consider an example such as (6).

(6) ... to be **choiced** for a job (Díaz Negrillo et al., 2010)

In this case, an *-ed* ending is being applied to what the native English lexicon would define as a noun, making it unclear whether one should interpret this example as a successful hit or not. Such learner innovations cannot be adequately characterized by native language categories—though we will see in section 3.2 how aspects of distribution, morphology, and lexicon can indeed be systematically characterized. Establishing annotation schemes for learner language and applying them to linguistically annotate learner corpora makes it possible to elucidate and make explicit the key questions involved in interpretation—and to develop NLP tools that automate the specific analysis to provide systematic access to larger corpora.

3 The Role of Annotation

NLP tools are of course imperfect, introducing errors into the desired annotation; in principle, however, they can provide the kinds of linguistic distinctions desirable for research (see section 1.2). Developing annotation schemes is thus a crucial step towards automating the annotation, i.e., a crucial computational linguistic question. Most importantly for our discussion, whether manually or automatically derived, corpus annotation makes it possible to search for subclasses of data in a way that overcomes some of the obstacles mentioned above, as specific subsets that speak to a given research question can reliably and efficiently be identified.

Let us consider analyzing the acquisition of relative clauses again. In section 2.1, we discussed that a search in unannotated corpora would make it necessary to spell out everything in terms of indicative lexical forms. Pursuing such research questions will generally require a richer characterization of the data, including identification of (i) the presence of a relative clause (RC), (ii) the word order of the clause containing the RC, (iii) the animacy of the modified noun, (iv) the type of RC (subject, object, ...), and possibly other properties depending on the specific research question.

Based on the version of the EFCAMDAT underlying the relative clause study of Alexopoulou et al. (2015), let us explore how one can identify the relevant aspects in the corpus, which has been syntactically annotated with dependency relations (de Marneffe et al., 2006). Take the example provided in Figure 1.

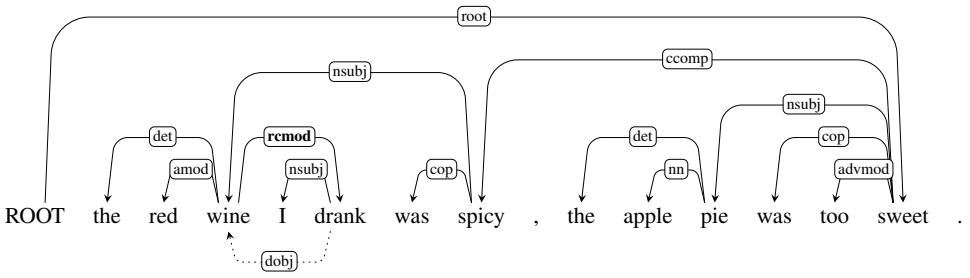

Figure 1: Dependency-annotated EFCAMDAT sentence including relative clause

While the dependency annotation is of a very general syntactic nature, it makes it possible to identify specific properties of interest for a study of relative clauses: (i) The label `rcmod` identifies relative clause modification relations, i.e., the connection between a head noun and the head of the RC, even without an explicit marker (such as *that*). (ii) The subject (*wine*) precedes the predicate (*spicy*), as indicated through the order of elements in the `nsubj` relation.

The annotation does not provide information on (iii) the animacy of the head noun (*wine*), though automatic systems are becoming more adept at providing such information (e.g., Moore et al., 2013; Alexopoulou et al., 2015). However, even if one has to analyze animacy by hand, the syntactic annotation already narrows one's focus. Instead of annotating the entire corpus of millions of words, one can focus on the relevant subset of relative clauses.

A less readily apparent challenge is that the categories and labels of the annotation scheme used in the corpus (e.g., `nsubj`) may not entirely align with one's own definitions (e.g., passive subjects, marked as `nsubjpass` in this scheme, may be relevant). Furthermore, some relations of interest may not be directly encoded in the annotation scheme at all, e.g., the `dobj` relation we indicate by a dotted line in the figure—needed to address (iv), the type of RC—is not marked in the standard dependency annotation and therefore requires more complex specifications or additional processing to be identifiable.

Overall, how one can map one's research questions and the categories and relations they rely on to the distinctions identified by the annotation schemes used in a given corpus resource is one of the central questions one needs to address to reliably make use of annotated corpora for SLA research. In answering that question, we can benefit from the insights gained in using annotated corpora for theoretical linguistics (cf., e.g., Meurers, 2005; Kübler and Zinsmeister,

2015, ch. 8); the discussion of how to link the relevant notions in the research questions to queries based on the different types of annotation schemes—from positional annotation, via syntactic constituency, to grammatical relations—carries over directly. We thus here focus on the challenges that are specific to learner language, given the space constraints. Following a discussion of the contribution of annotated corpora in the light of variability and ambiguity in section 3.1, we discuss challenges and opportunities arising in the annotation of learner language.

3.1 Annotation addressing variability and ambiguity

Systematic corpus annotation provides opportunities for dealing with some of the challenges arising from the variability and ambiguity of learner data characterized in section 2.2. The problems that form variability poses for expressing corpus queries to identify the targeted patterns in learner data can be significantly reduced, given that annotation can make explicit and integrate information from a range of sources: the forms in the corpus provide bottom-up lexical, morphological, and distributional information; tasks can provide top-down meaning and functional predictions; and learner models based on records of performance can support interpretations specific to individuals.

As a basic example illustrating different sources of information within language, consider the interpretation of the learner example (4a) again. Here, the distributional context indicates that the lexical form *loss* most likely is an orthographically incorrect realization of the verb *lose*. Note that different from ordinary contextual disambiguation of lexically ambiguous forms, there was no lexical ambiguity here (at least in the ordinary English lexicon). Annotating the distributional POS of *loss* in (4a) as a verb makes it possible to systematically retrieve this corpus instance together with the other verbs realized by the learners.

In a learner corpus annotated with target hypotheses (Lüdeling, 2008; Reznicek et al., 2013), the interpretation can be directly linked to an explicitly provided (target) form – i.e., if we annotate *loss* in the example (4a) with the target form *lose*, annotating that target form with the distributional POS tag *verb* is transparently supported. Apart from making the interpretation explicit, such target annotation also makes it possible to systematically search for cases where the context and its interpretation seems to be in conflict with the lexical forms found as surface evidence in the corpus.

To illustrate predictions going beyond the linguistic material as such, consider learner sentences such as those written by Japanese learners of English in (7), as recorded in the Hiroshima English Learners' Corpus (HELC, Miura, 1998).

(7) a. I don't know his lives.
    b. I know where he lives.

Knowing that these sentences were produced in a translation task, where learners were asked to express the Japanese sentence meaning "*I don't know where he lives.*", here provides crucial top-down information for interpreting the learner sentences in terms of the learners' capabilities to use *do* support, negation, and their difficulty distinguishing semantically related words with different parts of speech (cf. Zyzik and Azevedo, 2009). For such top-down interpretation to be explicitly supported, learner corpora must be collected in an explicit task-based setting (cf. Ott et al., 2012), which, parallel to the above form-based target hypotheses,

then supports interpretation directly linked to meaning-based target hypotheses (Meurers, 2015, sec. 2.1).

Turning to ambiguity, annotation directly addresses the challenge of ambiguous forms, given that annotation labels can be designed to disambiguate the different uses. Consider again the different uses of *that* in (3). Once *that* is annotated using a standard POS annotation scheme, the instances of *that* as a finite clause marker ((3a), (3b)) can be distinguished from the other instances, and standard syntactic annotation schemes would further delineate between relative clauses and complement clauses. Other ambiguities are not resolved by standard annotation schemes—they require integration of multiple schemes or a specifically designed one, e.g., to distinguish anaphoric and expletive uses of *it* (Boyd et al., 2005). As motivated in section 2.2, without such disambiguating annotation, one needs to specify in each query as much of the context as is necessary to disambiguate, often involving approximation. Instead, this disambiguation step can systematically be handled once, as part of the annotation process, where much more sophisticated methods (e.g., supervised machine learning) can integrate a wider range of information for disambiguation.

3.2 Challenges and opportunities for annotation

The argument from the previous section that annotation supports systematic interpretation based on different sources of information also turns out to be fruitful when considering challenges and opportunities arising in the annotation of learner language.

Starting with the lexical level, recall the occurrence of *choiced* in example (6), which poses a challenge for the traditional use and interpretation of linguistic categories. The apparent conflict between analyzing *choiced* as a past participle or as a noun can be handled if we take a closer look at the nature of the evidence that supports the different interpretations: i) the lexical stem *choice* is a noun or adjective, ii) the morphological looking-glass shows us the verbal suffix *-ed*, and iii) the distributional perspective indicates that the slot *to be __ for* is most likely filled by a participle. Following Díaz Negrillo et al. (2010), we can therefore accommodate the apparent conflict between the three sources of evidence by annotating a lexical POS, a morphological POS, and a distributional POS. This makes explicit where the evidence supporting the particular choice of annotation came from, and it makes it possible to directly identify in the corpus where language learners make choices in the different parts of the linguistic system that do not converge in the way they do in native language. Annotating and searching through three levels of POS annotation for learner language is readily supported by the multi-layer standoff annotation format of current corpora (Reznicek et al., 2013). Another option for advancing POS annotation for learner language proposed by Reznicek and Zinsmeister (2013) is to define Portmanteau tags (providing richer tagsets, combining information) and underspecified tags (leaving out some information), which they show can lead to an improved part-of-speech analysis of German learner language.

Moving on to syntax, consider the analysis of the learner utterance (8) from the SALLE corpus (Ragheb and Dickinson, 2011, 2012), where *idea* is missing a determiner, and *for do* is non-nativelike.

(8) It 's not good idea for do like that .

How to interpret the linguistic evidence in terms of the syntactic analysis depends on what exactly the syntactic annotation is intended to encode. Ragheb and Dickinson (2011, 2012)

explore this issue for the annotation of syntactic dependencies, i.e., grammatical relations between words (e.g., SUBJect). In the SALLE annotation scheme, they distinguish and annotate two syntactic aspects: a) the dependencies realized in a given sentence and b) the subcategorization frame of each word (in the native language).[11] They annotate dependencies based largely on morphological properties, as illustrated in Figure 2 for example (8).[12]

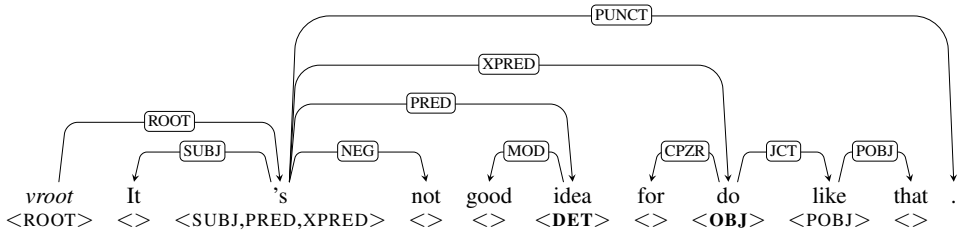

Figure 2: Morphosyntactic dependency tree

The arcs display the morphosyntactic dependency relations, encoding the relations based on the surface form of the tokens and the morphological POS, in this case annotating a case of *it* extraposition with the SUBJ, PRED, and XPRED arcs from *'s*. The subcategorization frame is shown below each word in angled brackets. Note the importance of annotating subcategorization information here for identifying learner innovations: both *idea* and *do* have selection requirements which are not fulfilled by the realized dependencies.[13]

The specifics of annotation depend upon one's goals and various practical decisions—e.g., whether to annotate multiple trees even for a single type of evidence, such as two different morphosyntactic trees for the present tense *do* vs. the baseform *do* in (8). From this example, we can see the need for multiple kinds of annotation, even just for form-based syntactic properties, and see that more work remains in developing annotation schemes that emphasize different properties.

Indeed, when we consider both form and meaning, the need for a further level of dependency annotation becomes apparent. Consider the grammatical function of *dull* in (9), discussed in Dickinson and Ragheb (2011).

(9) After the baby down, escape the **dull**.

If one considers the distributional properties of the sentence, i.e., its word order, *dull* is in the object position following the finite verb. If one instead considers the meaning, taking into account context indicating that a doll is escaping, then *dull* semantically is the agent, which grammatically would be realized as the subject.

In conclusion, parallel to the issues discussed for lexical annotation, for transparently connecting research questions in SLA with corpus data, systematic syntactic annotation of learner language needs to make explicit which source of evidence (morpho-syntax, distribution, lexical subcategorization, top-down guidance from task context, etc.) is considered in determining the different levels of annotation. Where more than one source of evidence is considered but the evidence diverges as a characteristic of learner language, each can be encoded in separate layers of annotation. The searching and interpretation of the corpus data can then systematically refer to consistently defined corpus annotation layers—an essential prerequisite for sustainable use of annotated corpora for second language research.

In consequence, joint research in SLA and computational linguistics is crucial to develop corpus annotation schemes and tools that make it possible for SLA research to systematically identify and interpret genuine characteristics of learner language. In practice, this is an incremental process that will also be informed by currently available annotation schemes and tools developed for native language. But crucially any research relying on corpus annotation should ask what evidence the annotation is based on and the goals and characteristics of the (manual or automatic) methods that were used to obtain it. For example, the Universal Dependency annotation scheme, recently used to annotate learner language (`http://esltreebank.org`), is designed to support consistent annotation across languages and is thus organized around content words and their dependencies rather than function words and morphosyntactic relations. Research using such a resource must therefore take into account the semantic rather than grammatical grounding of the annotation in order to ensure a valid interpretation.

Which kind of annotation is needed to support which kind of research questions, and which distinctions can be reliably annotated using which kind of NLP tools are central questions in making large scale learner corpus data relevant for SLA research.

4 Conclusion and Outlook

This paper has characterized two types of opportunities for interdisciplinary collaboration between computational linguistics and language learning research. Examining the connection between data and theory in second language research, we highlighted opportunities to i) enrich the options for obtaining data, and ii) support the identification and valid interpretation of relevant learner data. We highlighted the need to link research questions and corpus data, the steps involved, and the challenges that in particular the variability of forms in learner corpora and the ambiguity of their usage poses to their interpretation.

The challenges spelled out in section 3.2 present clear opportunities for future language learning research at the intersection of experimental, computational and corpus-based approaches, most notably the development and utilization of annotation schemes appropriate for capturing learner language characteristics. Annotation schemes are a link between what SLA researchers look for and what computational linguistic tools provide, and it is the domain of the SLA researchers to provide insight into interpretation that, on the one hand, shapes the interpretation of the output of linguistic searching and, on the other hand, helps refine the definitions of annotation schemes. There is very little systematically annotated learner corpus data, and now is the time for interdisciplinary work to provide annotation that supports cutting edge SLA research. This requires the contribution of various perspectives to develop annotation schemes, more precisely translating between categories found in SLA interlanguage research and categories appropriate for annotation, and to create more exemplars of work utilizing annotated corpus data to support or falsify theoretical claims or provide fodder for experimental research.

Systematic corpus annotation can help address existing research questions in new ways and to develop new investigations into the parameters of acquisition. As a step in this direction, Alexopoulou et al. (submitted) show how using linguistic annotation combined with error annotation can inform our understanding of the impact of tasks on corpus data and their interpretation as records of learning. In this context, NLP approaches designed to pinpoint annotation variability and inconsistencies (e.g., Dickinson and Meurers, 2003) could be used

to help improve the quality of annotation of large learner corpora and also provide a means to probe into the nature of the variability exhibited by the thousands of learner responses to a given task.

The opportunities and challenges connected to large scale annotated learner corpora point to a related opportunity for research collaboration stemming from new options for obtaining such data. In section 1.1, we emphasized the need to collect data in context and to record different types of learner interactions, for which tutoring systems and other (I)CALL tools provide opportunities—but also require collaboration to ensure that the setups are designed in a way satisfying the SLA research requirements. A major challenge for more SLA studies based on data collected using such systems is that a double effort is typically required: a system is designed and implemented, and an intervention study using the system is carried out. Reducing this workload would be greatly helped by: i) modular systems that can be adapted for multiple studies, ii) web-based platforms to support the testing of individual differences and pre-/post-test data specific to the targeted construct; and iii) advancing techniques to support valid inferences about language competence and proficiency. This last point involves improving NLP analyses targeting the specific needs of SLA studies (e.g., characteristics of relative clauses of Alexopoulou et al., 2015) as well as establishing sophisticated statistical analysis such as clustering techniques, linear mixed effects, and generalized additive models (Murakami, 2013; Vyatkina et al., 2015; Vajjala, 2015).

Notes

[1]For a more in-depth discussion of related background, see Norris and Ortega (2012).

[2]`https://sites.google.com/site/cognitivews2016`

[3]Representativeness is not (just) a matter of corpus size. In Brysbaert et al. (2011), for example, word frequencies drawn from the subtitle corpus SUBTLEX-US explain more of the variance in lexical decision times than those from the 7000 times larger Google book corpus. The much smaller SUBTLEX-US corpus thus seems more representative of the language experience of the subjects.

[4]`http://www.cs.rochester.edu/~tetreaul/naacl-bea11.html`

[5]While we here focus on the off-line identification and interpretation of indicative learner data in corpora, the issues discussed here are equally applicable to the on-line interpretation of learner data and the construction of learner models in tutoring systems.

[6]In this article, we focus on issues related to linguistic annotation that are crucially needed for linking SLA research questions and learner corpus evidence. For space reasons, we cannot also include a discussion specific to the annotation of learner errors, but the reader interested in NLP work on grammatical error detection can find an overview in Leacock et al. (2014). As a note of caution about the current state of the art in NLP and corpus annotation, let us mention that the best approach to grammatical error correction only reaches 39.7% precision, 30.1% recall (Ng et al., 2014) and inter-annotator agreement for manual error annotation of learner corpora, which is starting to be reported (e.g., Rosen et al., 2014) indicates that good agreement is only obtained for specific error types and when explicit target hypotheses are provided.

[7]While normalization replacing the text with normalized forms, as practised in social media analysis and other NLP application domains, eliminates distinctions relevant to SLA, the annotation of learner language with so-called target hypotheses (Hirschmann et al., 2007; Meurers, 2015) can satisfy the need for systematic access to data through normalized forms.

[8]The issues related to normalization and annotation are equally applicable to the use of corpora in historical linguistics, sociolinguistics, dialectology, and, in a somewhat different way, language typology. In historical linguistics, token normalization (Jurish, 2010; Bollmann et al., 2011, 2016; Azawi et al., 2013) and sentence segmentation (Petran, 2012) is actively discussed, which should support fruitful cross-disciplinary insight for the analysis of learner corpora.

[9]We here are assuming the corpus to be tokenized, so that words can be identified as units. For learner corpora, especially where hand-written text needs to be transcribed, this can pose additional challenges (cf., e.g., Štindlová et al., 2013, sec. 3).

[10]For some phenomena, one can detect non-presence by identifying obligatory contexts of occurrence—an option not readily available for identifying relative pronouns.

[11]Where multiple different subcategorization frames are possible for a given word, the contextually most appropriate frame is chosen.

[12]See Ragheb and Dickinson (2012) for argumentation of the bias towards morphology.

[13]The oddness of *for do* is an orthogonal issue here, which can be handled, among other ways, using morphological and distributional POS tags. We set aside here the question of which words distributionally determine the functions of which other words, e.g., whether *for* sets the context for *do* or vice versa.

References

Alexopoulou, T., Geertzen, J., Korhonen, A., and Meurers, D. (2015). Exploring large educational learner corpora for SLA research: perspectives on relative clauses. *International Journal of Learner Corpus Research*, 1(1):96–129.

Alexopoulou, T., Michel, M., Murakami, A., and Meurers, D. (submitted). Analyzing learner language in task contexts: A study case of task-based performance in EFCAMDAT. *Language Learning*. Special Issue on "Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and interpretation".

Alikaniotis, D. and Williams, J. N. (2015). A distributional semantics approach to implicit language learning. In Pirrelli, V., Marzi, C., and Ferro, M., editors, *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference, Pisa, March 30–April 1, 2015*, pages 81–84.

Amaral, L. and Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23(1):4–24.

Azawi, M. A., Afzal, M. Z., and Breuel, T. M. (2013). Normalizing historical orthography for ocr historical documents using LSTM. In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pages 80–85. ACM.

Baayen, R. H., Shaoul, C., Willits, J., and Ramscar, M. (2016). Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, 31(1):106–128.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press, Oxford, UK.

Baker, R. S. and Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics*, pages 61–75. Springer.

Baron, A. and Rayson, P. (2008). VARD2: a tool for dealing with spelling variation in historical corpora. In *Postgraduate Conference in Corpus Linguistics*, Birmingham.

Baron, A. and Rayson, P. (2012). "i didn't spel that wrong did i. oops": Analysis and normalisation of SMS spelling variation. *Lingvisticæ Investigationes*, 35(2):367–388.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4):243–257.

Bollmann, M., Dipper, S., and Petran, F. (2016). Evaluating inter-annotator agreement on historical spelling normalization. *Proceedings of LAW X – The 10th Linguistic Annotation Workshop,*, pages 89–98.

Bollmann, M., Petran, F., and Dipper, S. (2011). Rule-based normalization of historical texts. In *Proceedings of the International Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42.

Boyd, A., Gegg-Harrison, W., and Byron, D. (2005). Identifying non-referential it: A ma-

chine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 40–47, Ann Arbor, Michigan.

Brysbaert, M., Keuleers, E., and New, B. (2011). Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 2(27).

Chapelle, C. A., editor (2012). *Encyclopedia of Applied Linguistics*. Wiley, Oxford.

Clahsen, H. and Muysken, P. (1986). The availability of universal grammar to adult and child learners – a study of the acquisition of German word order. *Second Language Research*, 2(2):93–119.

Covington, M. A., He, C., Brown, C., Naçi, L., and Brown, J. (2006). How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale. Computer Analysis of Speech for Psychological Research (CASPR) Research Report 2006-01, The University of Georgia, Artificial Intelligence Center, Athens, GA.

de Marneffe, M.-C., MacCartney, B., and Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.

de Saussure, F. (1916). *Cours de Linguistique Générale [Course in General Linguistics]*. Éditions Payot & Rivages.

Dickinson, M. and Meurers, D. (2003). Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 107–114, Budapest, Hungary.

Dickinson, M. and Ragheb, M. (2011). Dependency annotation of coordination for learner language. In *Proceedings of the International Conference on Dependency Linguistics*, Barcelona, Spain.

Dickinson, M. and Ragheb, M. (2015). On grammaticality in the syntactic annotation of learner language. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 158–167, Denver, CO.

Díaz Negrillo, A., Meurers, D., Valera, S., and Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2):139–154.

Ellis, N. C., Römer, U., and O'Donnell, M. B. (2016). *Usage-Based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*. Wiley.

Ellis, R. and Barkhuizen, G. P. (2005). *Analysing learner language*. Oxford University Press.

Farzindar, A. and Inkpen, D. (2015). *Natural Language Processing for Social Media*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Geertzen, J., Alexopoulou, T., and Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum (SLRF)*. Cascadilla Press.

González-Lloret, M. and Ortega, L., editors (2014). *Technology-Mediated TBLT: Researching Technology and Tasks*, volume 6. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Graesser, A. C., McNamara, D. S., Louweerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, 36:193–202.

Heift, T. (2003). Multiple learner errors and meaningful feedback: A challenge for ICALL

systems. *CALICO Journal*, 20(3):533–548.

Heift, T. and Schulze, M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

Heinz, J., De la Higuera, C., and van Zaanen, M. (2015). *Grammatical Inference for Computational Linguistics*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Hirschmann, H., Doolittle, S., and Lüdeling, A. (2007). Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham.

Housen, A. and Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473.

Jurish, B. (2010). Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 72–77. Association for Computational Linguistics.

King, L. and Dickinson, M. (2013). Shallow semantic analysis of interactive learner sentences. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, GA USA.

Kübler, S. and Zinsmeister, H. (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury, London.

Kyle, K. and Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4):757–786.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, 12(4):439–448.

Leacock, C., Chodorow, M., Gamon, M., and Tetreault, J. (2014). *Automated Grammatical Error Detection for Language Learners*, volume 25. Morgan & Claypool Publishers, 2 edition.

Leow, R., Cerezo, L., and Baralt, M. (2015). *A Psycholinguistic Approach to Technology and Language Learning*. De Gruyter Mouton, Berlin, Boston.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Languages Journal*, pages 190–208.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In Walter, M. and Grommes, P., editors, *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweispracherwerbsforschung*, pages 119–140. Max Niemeyer Verlag, Tübingen.

MacWhinney, B. (1997). Second language acquisition and the competition model. In de Groot, A. M. B. and Kroll, Judith, F., editors, *Tutorials in Bilingualism*, pages 113–142. Lawrence Erlbaum Associates.

Malvern, D. D., J., R. B., N., C., and P., D. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Massung, S. and Zhai, C. (2015). Non-native text analysis: A survey. *Natural Language Engineering*, pages 1–24.

McCarthy, P. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Meurers, D. (2005). On the use of electronic corpora for theoretical linguistics. case studies from the syntax of German. *Lingua*, 115(11):1619–1639.

Meurers, D. (2012). Natural language processing and language learning. In Chapelle (2012), pages 4193–4205.

Meurers, D. (2015). Learner corpora and natural language processing. In Granger, S., Gilquin, G., and Meunier, F., editors, *The Cambridge Handbook of Learner Corpus Research*, pages 537–566. Cambridge University Press.

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., and Ott, N. (2010). Enhancing authentic web pages for language learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-5) at NAACL-HLT 2010*, pages 10–18, Los Angeles.

Miura, S. (1998). Hiroshima English Learners' Corpus: English learner No. 2 (English I & English II). Department of English Language Education, Hiroshima University. `http://purl.org/icall/helc`.

Moore, J., Burges, C. J., Renshaw, E., and Yih, W.-t. (2013). Animacy detection with voting models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 55–60, Seattle, Washington, USA. Association for Computational Linguistics.

Murakami, A. (2013). *Individual Variation and the Role of L1 in the L2 Development of English Grammatical Morphemes: Insights From Learner Corpora*. PhD thesis, University of Cambridge.

Nagata, N. (2009). Robo-sensei's NLP-based error detection and feedback generation. *CALICO Journal*, 26(3):562–579.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Norris, J. M. and Ortega, L. (2012). Assessing learner knowledge. In Gass, S. M. and Mackey, A., editors, *The Routledge handbook of second language acquisition*, pages 573–589. Routledge.

Ott, N., Ziai, R., and Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Schmidt, T. and Wörner, K., editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam.

Petersen, K. (2010). *Implicit Corrective Feedback in Computer-Guided Interaction: Does Mode Matter?* PhD thesis, Georgetown University.

Petran, F. (2012). Studies for segmentation of historical text: sentences or chunks? In *The Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*.

Plonsky, L. and Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64(4):878–912.

Quixal, M. and Meurers, D. (2016). How can writing tasks be characterized in a way serving pedagogical goals and automatic analysis needs? *CALICO Journal*, 33.

Ragheb, M. and Dickinson, M. (2011). Avoiding the comparative fallacy in the annotation of learner corpora. In *Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions*, pages 114–124, Somerville, MA. Cascadilla Proceedings Project.

Ragheb, M. and Dickinson, M. (2012). Defining syntax for learner language annotation. In

Proceedings of COLING 2012, pages 965–974, Mumbai, India.

Rebuschat, P. and Williams, J. N. (2012). Implicit learning in second language acquisition. In Chapelle (2012).

Reznicek, M., Lüdeling, A., and Hirschmann, H. (2013). Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In Díaz-Negrillo, A., Ballier, N., and Thompson, P., editors, *Automatic Treatment and Analysis of Learner Corpus Data*, volume 59, pages 101–123. John Benjamins.

Reznicek, M. and Zinsmeister, H. (2013). STTS-Konfusionsklassen beim Tagging von Fremdsprachlernertexten. *Journal for Language Technology and Computational Linguistics (JLCL)*.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

Rosen, A., Hana, J., Štindlová, B., and Feldman, A. (2014). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, 48(1):65–92.

Seiler, J. S. (2013). On characterizing German interlanguage part-of-speech classes: Multilevel categories as bridge between empirical observations and linguistic generalizations. Master's thesis, Universität Tübingen.

Štindlová, B., Škodová, S., Hana, J., and Rosen, A. (2013). A learner corpus of Czech: current state and future directions. In Granger, S., Gilquin, G., and Meunier, F., editors, *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*, Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve. Presses Universitaires de Louvain.

Stringer, D. (2015). Embedded *wh*-questions in L2 English in India: Inversion as a main clause phenomenon. *Studies in Second Language Acquisition*, 37(1):101–133.

Tagliamonte, S. A. (2011). *Variationist Sociolinguistics: Change, Observation, Interpretation*. John Wiley & Sons.

Tetreault, J., Blanchard, D., and Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP*, Atlanta, GA, USA. Association for Computational Linguistics.

Vajjala, S. (2015). *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. PhD thesis, University of Tübingen.

Vyatkina, N., Hirschmann, H., and Golcher, F. (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing*.

Wilske, S. (2015). *Form and meaning in dialogue-based computer-assisted language learning*. PhD thesis, Universität des Saarlandes, Saarbrücken.

Zhao, H., Koedinger, K. R., and Kowalski, J. (2013). Knowledge tracing and cue contrast: Second language English grammar instruction. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 1653–1658.

Zyzik, E. and Azevedo, C. (2009). Word class distinctions in second language acquisition. *SSLA*, 31(31):1–29.