Learner Corpora and Natural Language Processing

Detmar Meurers
Universität Tübingen

Core issues, notions, and methods

Learner corpora collect the language produced by people learning their first or a second language. Natural Language Processing (NLP) deals with the representation and the automatic analysis and generation of human language. The two thus overlap in the ***representation and automatic analysis of learner language***, which constitutes the topic of this chapter. As such, the chapter focuses on one of the two application areas for NLP in the context of language learning (cf. Meurers, 2012), the other one being the use of NLP to process the native language to be learned, for example, to generate exercises, to support retrieval of reading material at the appropriate learner level, or to present texts to learners with visual or other enhancements to support language learning.

Corpus representation At the fundamental level, representing a learner corpus amounts to encoding the language produced by the learner and its meta-data, such as information about the learner and the task performed (cf. Granger, 2008, ch. 15.3). For the ***spoken language*** constituting the primary data for research on first language acquisition (e.g., CHILDES, MacWhinney, 2000), most work on uninstructed second language acquisition (e.g., ESF, Perdue, 1993), and a small part of the instructed second language acquisition corpora (e.g., NICT-JLE, Tono et al., 2004), this involves the question how to encode and orthographically transcribe the sound recordings (Wichmann, 2008, ch. 12.3). ***Written language***, such as the learner essays typically collected in instructed second language learning contexts (e.g., ICLE, Granger et al., 2009), also require transcription for hand-written learner texts, though essays typed by learners are increasingly common and the opportunity to systematically collect learner productions entered into Computer-Assisted Language Learning systems is staring to be used, supporting the creation of very large learner corpora such as EFCAMDAT (Geerzen et al., 2013).

While the fundamental questions around how to represent spoken and written language in corpora are largely independent of the nature of the language being collected, and good general corpus linguistic discussions can be found, e.g., in McEnery et al. (2006) and Lüdeling & Kytö (2008), there are important **representation aspects specific to learner language**. Learner language by its very nature does not follow the native language norms. In fact, researchers in second language acquisition emphasize the individual, dynamic nature of interlanguage (Selinker, 1972), and focus on characterizing its properties as a language system in its own right. At the same time, the analysis of language, be it manual linguistic analysis or automatic NLP analysis, was developed for and trained on well-formed native language. When tying to analyze learner data on that basis, one encounters forms and patterns which cannot be analyzed in terms of the native language system.

Consider, for example, the learner sentence in (1), taken from the NOCE corpus (Díaz Negrillo, 2007), consisting of essays written by intermediate Spanish learners of English.

(1) People who speak another language have more opportunities to be choiced for a job because there is a lot connection between the different countries nowadays.

In line with the native English language system, the verbal *-ed* suffix of the underlined word *choiced* can be identified as a verbal suffix and interpreted as past tense, and the distributional slot between *to be* and *for a job* is syntactically appropriate for a verb. But the stem *choice* in English can only be a noun or an adjective. As discussed in Díaz Negrillo et al. (2010), in a systematic set of cases it thus is not possible to assign a unique English part-of-speech to learner tokens.

In examples such as (1), it seems straightforward to analyze the sentence as though the learner had written the appropriate native English form *chosen* in place of the interlanguage form *choiced*. Yet, even for such apparently clear non-word cases, where a learner used a word that is not part of the target language system, different native language words may be inferred as targets (e.g., *selected* could have been an alternative option for the example above), and the subsequent analysis can differ depending on which target is assumed. When we go beyond the occurrence of isolated non-words, the question which level of representation of learner corpora can form the basis for the subsequent analysis becomes more pronounced. For example, consider the sentence in (2) written by a beginning learner of German as found in the Error-Annotated German Learner Corpus (EAGLE, Boyd, 2012, ch. 7). The sentence only includes well-formed words, but the subject and the verb fail to show the subject-verb agreement required by German grammar.

(2) Du arbeiten in Liechtenstein.
you$_{2sg}$ work$_{1pl/3pl/inf}$ in Liechtenstein

Given that agreement phenomena always involve (at least) two elements, there systematically is an ambiguity in determining grammatical target forms. If we take the second person subject *du (you)* at face value, the corresponding second person verb form *arbeitest* is the likely target. Or we keep the verb, assume that it is a finite form and thus postulate the corresponding plural third person *sie (they)* or first person *wir (we)* as subject to obtain a well-formed target language sentence. Even with more linguistic context and information about the task the learner wrote this sentence for, it often is difficult to decide on a unique target form. This is empirically confirmed by the study presented in Fitzpatrick & Seegmiller (2004).

To ensure that the empirical basis on which the subsequent analysis of the learner utterance is based is explicitly present in the corpus, Lüdeling (2008, and elsewhere) has argued for overtly specifying such ***target hypotheses*** as an explicit representation level of learner corpora. To ensure valid, sustainable interpretations of learner language, one needs to document the reference on which the analysis is based. As explored in more detail in the first case study discussed below, Rosen et al. (2013, sec. 5.4) confirm that disagreement in the analysis of the learner data in their Czech as Second Language (CzeSL) corpus often arises from different target hypotheses being assumed.

While there seems to be a growing consensus that a replicable analysis of learner data requires the explicit representation of target hypotheses in learner corpora, such an approach requires a precise definition of target hypothesis and how it is obtained. There essentially are two pieces of evidence that one can take into account in determining a target hypothesis. On the one hand, one can interpret the ***forms*** the learner provided in the learner corpus bottom-up in terms of a linguistic reference system, such as the targeted native language system also codified in the standard corpus annotation schemes. One can then define a target hypothesis which encodes the minimal form-change required to turn the learner sentence into a sentence which is well-formed in terms of the target language grammar. A good example is the Minimal Target Hypothesis (ZH1) made explicit in the annotation manual of the German learner corpus FALKO (Reznicek et al., 2012, sec. 5.2). An alternative incremental operationalization of a purely form-based target hypothesis is spelled out in Boyd (2010, 2012). Both approaches explicitly define what counts as minimal form change in the derivation of the target hypothesis, i.e., they do not try to guess what the learner may have wanted to say and how this could have been expressed in a well-formed sentence, but apply the minimal number of form changes to the learner sentence needed to obtain a grammatically well-formed sentence in the target language. While this makes it possible to uniquely identify a single target hypothesis in many cases, for some cases multiple possible target hypotheses are required. This can readily be represented in corpora using multi-layer standoff annotation (Reznicek et al., forthcoming).

On the other hand, one can determine target hypotheses using top-down information about the ***function*** of the language and the ***meaning*** the learner was trying to express, based on what we know about the particular task and general expectations about human communication. The top-down, meaning-driven and the bottom-up form-driven interpretation processes essentially interact in any interpretation of human language. The interlanguage forms used by language learners cannot fully be interpreted based on the established linguistic reference systems developed for the native language. This is particularly evident for learner language varieties such as the Basic Variety that is characteristic of uninstructed second language acquisition (Klein & Perdue, 1997), which lacks most grammatical form marking altogether. Top-down guidance integrating task and context information thus is particularly important for interpreting learner language, which makes corpora with explicit task contexts particularly relevant for learner corpus research aimed at drawing valid inferences about the learners' second language knowledge and development.

Consider, for example, the learner sentences in (3) written by Japanese learners of English, as recorded in the Hiroshima English Learners' Corpus (HELC, Miura, 1998).

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

Both sentences are grammatically well-formed in English, so the form-based target hypotheses are identical to the respective learner sentences. However, if we go beyond the form of the sentence and take the context and meaning into account, we find that both sentences were produced in a translation task to express the Japanese sentence meaning *I don't know where he lives.* We can thus provide this meaning-based target hypothesis for the two sentences. On this basis, we can analyze the learner sentences and, for example, interpret them in terms of the learners' capabilities to use *do* support, negation, and to distinguish semantically related words with different parts of speech.

While the example relies on an explicit task context in which a specific sentence encodes the meaning to be expressed for this translation exercise, the idea to go beyond the forms in the sentence towards meaning and function in context is generally applicable. For example, it is also present in the annotation guidelines used for the learner essays and summaries collected in the FALKO corpus. The Extended Target Hypothesis (ZH2) operationalized in Reznicek et al. (2012, sec. 5.3) takes into account the overall text, the meaning expressed, the function and information structure, and aspects of the style. While such an extended target hypothesis provides an important reference for a more global, functional analysis, it naturally cannot be made explicit in the same formal way as the minimal form-change target hypothesis ZH1, entailing lower inter-annotator agreement for ZH2 annotation.

A global, meaning-based target hypothesis may also seem to come closer to an intuitive idea of the target hypothesis as 'what the learner wanted to say', but such a seemingly intuitive conceptualization of target hypotheses would be somewhat naive and more misleading than helpful. Learners do not simply write down language to express a specific meaning. They employ a broad range of strategies to use language in a way that achieves their communicative or task goals. Correspondingly, Bachman & Palmer (1996) discuss planning how to approach a test task as a good example of strategic competence, one of the components of language competence since Canale & Swain (1980). In an instructed second language learning setting, learners know that form errors are one of the aspects they typically are evaluated on, and therefore they strategically produce language in a way minimizing the number of form errors they produce. For example, we found in the Corpus of Reading Comprehension Exercises in German (CREG) that the learners simply lift material from texts or use familiar chunks (Ott et al., 2012, sec. 4.2.2), a strategy which, e.g., allows the learners to avoid generating the complex agreement patterns within German noun phrases. In a similar English learner corpus, Bailey (2008) found that this strategy was used more frequently by less proficient learners, who made made fewer form errors overall (but less frequently answered the question successfully). A second reason for rejecting the idea that a target hypothesis is 'what the learner wanted to say' is that learners do not plan what they want to say in terms of full-fledged target language forms (though they may access chunks and represent aspects of it at a propositional level, as included in current cognitive models of linguistic meaning (e.g., Kintsch & Mangalath, 2011)). Even for the target language forms produced by the learners, their conceptualization of the language forms used will not necessarily coincide with the analysis in terms of the target language system, as, e.g., evidenced by the learners' difficulties to interpret feedback in an intelligent tutoring system context (Amaral & Meurers, 2009).

Summing up this discussion, target hypotheses are intended to provide an explicit representation that can be interpreted in terms of an established linguistic reference system (typically that of the language being acquired). The form-based and the meaning-based target hypotheses discussed above are two systematic options that can serve as a reference for a wide range of analyses of language. Conceptually, a target hypothesis needs to make explicit the minimal commitment required to support a specific type of analysis/annotation of the corpus. As such, target hypotheses are not required to be full sentences, and more abstract target hypothesis representations may help avoid an overcommitment made by specifying the full surface forms of sentences, e.g., requiring full specification of a particular word order in a language with relatively free word order.

Three types of NLP uses for the analysis of learner language Turning from corpus representation to automatic NLP analysis, we can in principle distinguish three types of NLP uses involving learner corpora:

First, NLP tools are employed to ***annotate learner corpora*** with a wide range of general properties and to gain insights into the nature of language acquisition or typical learner needs on that basis. On the one hand, this includes general ***linguistic properties*** from part-of-speech and morphology, via syntactic structure and dependency analysis, to aspects of meaning and discourse, function, and style. On the other, there are properties specific to learner language, such as different types of ***learner errors***, again ranging from the lexical and syntactic to discourse and function. Such off-line application of NLP tools for annotation can be combined with human post-editing to eliminate some of the error introduced by the automatic analysis. NLP tools can also be integrated into a manual annotation setup to automatically identify likely error locations, refine manual annotation (e.g., Rosen et al., 2013), or flag annotation that appears to be inconsistent across comparable corpus instances (Dickinson & Meurers, 2003; Boyd et al., 2008).

Second, NLP tools are used to provide ***specific analyses of the learner language in the corpus***. For example, in Native Language Identification (NLI) classifiers are trained to automatically determine the native language of the second language learner who wrote a given essay in a learner corpus (cf. Tetreault et al., 2013). In another task, the NLP analysis of a learner essay is used to determine the proficiency level of the learner who wrote a given essay (Pendar & Chapelle, 2008; Yannakoudakis et al., 2011; Vajjala & Lõo, 2013; Hancke & Meurers, 2013), a task related to the analysis of developmental sequences and criterial features of different stages of proficiency (Granfeldt et al., 2005; Rahkonen & Håkansson, 2008; Alexopoulou et al., 2011; Tono, forthcoming; Murakami, 2013) and the popular application domain of automatic essay grading (Shermis & Burstein, 2013).

The third type of NLP application in the context of learner corpora is related to the previous two, but different from those is not designed to provide insights into the learner corpus as such. Instead, the ***learner corpus is only used to train the NLP tools***, specifically the statistical or machine learning components. The trained NLP tools can then be applied to learner language arising in other contexts. For example, using learner corpora one can develop general tools to analyze the complexity of learner language (e.g. Lu, 2009, 2010) or the readability of native language (Vajjala & Meurers, 2012; Hancke et al., 2012). A tool trained on a learner corpus to detect particular types of learner errors can also be used to provide immediate, individualized feedback to learners who complete exercises in an Intelligent Tutoring System, where it can also help determine an appropriate individual progression through the pedagogical material. So, while traditionally the two fields of Learner Corpus Research and Intelligent Computer Assisted Language Learning (ICALL) developed independently and largely unconnected (but cf. Granger et al., 2007), the automatic analysis of learner language employed to annotate learner corpora essentially is an offline variant of the online analysis of learner language in Intelligent Tutoring Systems as the most prominent application in Intelligent Computer-Assisted Language Learning (ICALL).

Annotating learner corpora The purpose of annotating learner corpora is to provide an effective and efficient index into relevant subclasses of data. As such, linguistic annotation serves essentially the same purpose as the index of a telephone book. A telephone book allows us to efficiently look up the phone number of people by the first letter of the last name – the alternative, linear search, reading through the phone book from the beginning to the end until one finds the right person, would be possible, but would not be efficient enough to be useful in real-life, at least for any phone book covering more than a small village. While indexing phone book information by the first letter of the last name is typical, it is only one possible index – one that is well-suited for the typical questions one tries to address using such phone books. For other questions which can be addressed using the same telephone book information, we need other indices. For example, consider a situation in which someone called us, we have a phone that displays the number of the caller, and we now want to find out, who called us. We would need a phone book that is indexed by phone numbers. Or to be able to efficiently look up who lives on a particular street, we would need a book that is indexed alphabetically be the first letter of the street name. Let us take this running example one important step further by considering what it takes to look up phone numbers of all the butchers in a given town. Given that a phone book typically does not list professions, we need an additional resources to first determine the names of all the butchers. And if we often want to look up people by their profession, we may decide to add that information to the phone book so that we can more readily index the data based on that information.

Each layer of annotation we add to corpora as collections of language data serves exactly that purpose of providing an efficient way to index language data to retrieve the subclasses of data that helps us answer common (research) questions. For example, to pick out occurrences of the main verb *can* as in *Dario doesn't want to can tuna for a living*, we need part-of-speech annotation that makes it possible to distinguish such occurrences of *can* from the frequent uses of *can* as an auxiliary (*Cora can dance.*) or as a noun (*What is Marius doing with that can of beer?*) which cannot readily be distinguished by only looking at surface forms.

Which subclasses are relevant depends on the research question and how corpus data is involved in addressing it. For Foreign Language Teaching and Learning (FLTL), the questions are driven by the desire to identify and exemplify typical student characteristics and needs. For First and Second Language Acquisition research (F/SLA), learner corpora are queried to inform the empirical basis on which theories of the acquisition process and its properties are developed and validated. General linguistic layers of annotation are useful for querying the corpus for a wide range of research questions arising in FLTL and F/SLA – much like annotating telephone book entries with professions allows us to search for people helping us address a wide range of different needs, from plumbers to hairdressers. On the other hand, annotating all phone entries with the particular day of the week on which they are born would not provide access to classes of data which are similarly relevant. Which type of annotations one can and should provide for learner corpora using NLP tools, manual annotation, or a combination of the two, is an important research issue at the intersection of learner corpus and NLP research.

Linguistic annotation A wide range of linguistic corpus annotation schemes have been developed for written and spoken language corpora (cf., e.g., Garside et al., 1997; Leech, 2004), and the NLP tools developed over the past two decades support the automatic identification of a number of language properties, including lexical, syntactic, semantic and pragmatic aspects of the linguistic system.

For learner corpora, the use of NLP tools for annotation is much more recent (de Haan, 2000; de Mönnink, 2000; van Rooy & Schäfer, 2002, 2003; MacWhinney, 2008; Sagae et al., 2007, 2010) and the discussion which kind of annotation schemes are relevant and useful to address which learner corpus research questions is only starting to be discussed. For advanced learner varieties, the annotation schemes and NLP tools developed for edited native language corpora can seemingly be applied, though at closer inspection even this requires some leeway when checking the definitions in the annotation schemes, which for the NLP tools is generally discussed under the topic of ***robustness***.

Real-life NLP applications such as a machine translation system should, for example, be able to translate sentences even if they contain some spelling mistakes or include words we have not encountered before, such as an unusual proper name. Robustness in corpus annotation allows the NLP tools to classify a given learner language instance as a member of a particular class (e.g., a particular part-of-speech) even when the observed properties of those instances differ from what is expected for that class (e.g., when the wrong stem is used, as in the case of *choiced* we discussed for example (1)). At a given level of analysis, robustness thus allows the NLP tools to gloss over those aspects of learner language that differ from the edited native language for which the annotation schemes and tools were developed and trained. In other words, robustness at a given level of analysis is intended to ignore the differences between the learner and the native language at that level.

In contrast, most of the uses of learner corpora are related to advancing the understanding of language acquisition or teaching by identifying characteristics of learner language. For such research, the particularities and variability of learner language at the level being investigated thus are exactly what we want to identify, not gloss over robustly. We already discussed one important component for addressing this issue: target hypotheses, which can be seen as a way to document the variation that robust analysis would simply have glossed over. The target hypotheses require the researcher to make explicit where a change is required to be able to analyze the learner language using a standard linguistic annotation scheme. A learner corpus including target hypotheses and linguistic annotation on that basis thus makes it possible to identify both the places where the learner language diverges from the native language norm as well as the general linguistic classes needed for retrieval of relevant subsets of learner data.

At the same time, this cannot be the full solution for analyzing the characteristics of learner language. It amounts to interpreting learner language in a documented way, but still in terms of the annotation schemes developed for native language instead of annotation schemes defined to systematically reflect the properties of interlanguage itself. This is natural, given that linguistic category systems arose on the basis of a long history of data observations, based on which a consensus of the relevant categories emerges – a process which highlights why such category systems are difficult to develop for the individual, dynamic interlanguage of language learners. Still, by using a native language annotation scheme to characterize learner language, one runs the danger of committing a ***comparative fallacy***, "the mistake of studying the systematic character of one language by comparing it to another" (Bley-Vroman, 1983, p. 6).

Of course, as is well-known from hermeneutics, every interpretation is based on the given background, and if one wants to push this issue to the extreme, it is evident that one can never perceive anything as such – we are autopoietic systems evolving in a constant hermeneutic circle (cf., e.g., the accessible introduction to hermeneutics and radical constructivism included in Winograd & Flores, 1986). Pursuing a more pragmatic approach, as far as we see one can effectively limit the degree of the comparative fallacy entailed by the annotation scheme used. The idea is to ensure that we annotate learner language as close as possible to the specific dimensions of observable empirical properties. For example, traditional part-of-speech encode a bundle of syntactic, morphological, lexical, and semantic characteristics of words. For learner language, we proposed in Díaz Negrillo et al. (2010) to instead use a tripartite encoding with three separate parts-of-speech to explicitly encode the actually observable distributional, morphological, and lexical stem information. For native language the three converge and can be encoded by one part-of-speech tag, whereas for learner language these three information sources may diverge (as in the example (1) discussed above).

In the syntactic domain, encoding classes close to the empirical observations can be realized by breaking down constituency in terms of a) the overall topology of a sentence (Höhle, 1986), b) chunks and chunk-internal word order (Abney, 1997), and c) lexical dependencies. Topological fields are already being employed for the analysis of learner language (Hirschmann et al., 2007), and various chunk concepts are widely discussed in the context of learner language (though often in need of a precise operationalization and corpus-based evaluation), but the dependency analysis requires more elaboration here. To pursue the envisaged analysis close to the specific empirical observations, one must carefully distinguish between morphological, syntactic, and semantic dependencies, as, e.g., realized in Meaning Text Theory, (Mel'čuk, 1988) or reflected in the distinction between the analytical and the tectogrammatical layer of the Prague Dependency Treebank (Böhmová et al., 2003). On that basis, we can distinguish two types of dependency analyses which have been developed for learner language. On the one hand, we find surface-evidence based approaches that aim at providing a fine-grained record of the morphological and syntactic evidence (Dickinson &

Ragheb, 2009; Ragheb & Dickinson, 2012). On the other, there are approaches which seem to target a level of semantic dependencies (MacWhinney, 2008; Rosén & Smedt, 2010; Ott & Ziai, 2010; Hirschmann et al., 2010). The goal here is to robustly abstract away from learner specific forms to encode the underlying function-argument relations on which the sentential meaning can be derived. For example, as part of the CoMiC system analyzing learner answers to reading comprehension questions, dependency parsing (Hahn & Meurers, 2011) serves as the basis of a meaning analysis based on formal semantics (Hahn & Meurers, 2012). King & Dickinson (2013) report on the NLP analysis of another task-based learner corpus supporting the evaluation of meaning, obtaining very high accuracies for the extraction of the core functor argument relations using shallow semantic analysis of learner data from a picture description task. The important question is which kind of dependency distinctions can reliably be identified, be it for the surface-based morphological or the underlying semantic dependencies, is also starting to be addressed in recent work (Ragheb & Dickinson, 2013).

Complementing the conceptual points relating to dependencies, the use of NLP to automatic derive dependency parses raises the question which impact the choice of parsing approach and algorithm has on the quality of the dependency analysis obtained for learner language. Comparing two different computational approaches to dependency parsing learner language for German, for example, in Krivanek & Meurers (2011) we show that the rule-based WCDG approach (Foth & Menzel, 2006) is more reliable in identifying the core functor-argument relations, whereas the data-driven MaltParser (Nivre et al., 2007) is more reliable in identifying adjunct relations. This also is intuitively plausible given that statistical approaches can use the world knowledge encoded in a corpus to disambiguate attachment and label ambiguities, whereas the grammar-based approach can rely on high quality subcategorization information forming the syntactic core of the linguistic system.

Error annotation A second type of annotation of learner corpora, error annotation, targets the nature of the difference between learner data and native language. Given the FLTL interest in identifying, diagnosing, and providing feedback on learner errors, and the fact that learner corpora are commonly collected in a foreign language teaching context, error annotation is the most commonly discussed type of annotation in the context of learner corpora (Granger, 2003; Díaz Negrillo & Fernández Domínguez, 2006). At the same time, error annotation is only starting to be subjected to the rigorous systematization and inter-agreement testing established for linguistic annotation, which will help determine which distinctions can reliably be annotated based on the evidence available in the corpus. The issue clearly requires scrutiny given that Fitzpatrick & Seegmiller (2004) report low inter-annotator agreement results for determining target forms and Rozovskaya & Roth (2010) find very low inter-annotator agreement for error classification of ESL sentences. Even for the highly focused task of annotating preposition errors, Tetreault & Chodorow (2008a) report that trained annotators failed to reach good agreement. In the second case study below, we will discuss one of the leading efforts in this domain, Rosen et al. (2013), who provide detailed inter-agreement analyses for their Czech learner corpus, making concrete for which aspects of error annotation good agreement can be obtained and what this requires.

In terms of computational tools for detecting errors, learner corpus research has long envisaged automatic approaches (e.g., Granger & Meunier, 1994), but the small community at the intersection of NLP and learner corpus research is only starting to make headway, presumably because of the mentioned conceptual difficulties and the unavailability of gold-standard error annotated learner corpora. Writer's aids such as the standard spell and grammar checkers developed for native speakers (Dickinson, 2006) may seem like a natural option to fall back on. However, such tools rely on assumptions about typical errors made by native speakers which are not necessarily applicable to language learners. For example, Rimrott & Heift (2008) find that "in contrast to most misspellings by native writers, many L2 misspellings are multiple-edit errors and are thus not corrected by a spell checker designed for native writers."

The overall landscape of computational approaches for diagnosing learner errors can be systematized in terms of the nature of data that is targeted, from single tokens via local domains to full sentences. ***Pattern-matching approaches*** target single tokens or local patterns to detect specific types of errors. ***Language-licensing approaches*** attempt to analyze entire learner utterance to diagnose its characteristics.

Pattern-matching approaches traditionally employ ***error patterns*** explicitly specifying surface forms. For example, an error pattern for English can target occurrences of *their* immediately preceding *is* or *are* to detect learner errors such as (4) from the Chinese Learner English Corpus (CLEC, Huizhong & Shichun, 2004).

(4) Their are all kinds of people around us.

Such local error patterns can also be defined in terms of annotations such as parts-of-speech to allow identification of more general patterns. For example, one can target *more* or *less* followed by an adjective or adverb, followed by *then*, an error pattern instantiated by the CLEC learner sentence (5).

(5) At class, students listen more careful$_{adj}$ then any other time.

Error pattern matching is commonly used in standard grammar checkers. For example, the open source LanguageTool (Naber, 2003; http://www.languagetool.org) provides a general implementation, in which common learner error patterns can be specified.

While specifying such error patterns directly works well for certain clear cases of errors, more advanced pattern-matching splits the error identification into two steps. First, a ***context pattern*** is defined to identify the contexts in which a particular type of error may arise. Then potentially relevant features are collected, recording all properties which may play a role in distinguishing erroneous from correct usage. A supervised machine learning setup can then be used to learn how to weigh the evidence to accurately diagnose the presence of an error and its type. For example, given that determiner usage is a well-known problem area for learners of English, a context pattern can be used to identify all noun chunks. Properties of the noun and its context then can be used to determine whether a definite, an indefinite, or no determiner is required for this chunk. This general approach probably is the most common setup of current NLP research targeting learner language (cf., e.g., De Felice, 2008; Tetreault & Chodorow, 2008b; De Felice & Pulman, 2009; Gamon et al., 2009), and it raises important general questions for future work in terms of how much context and which linguistic properties are needed to accurately diagnose which type of errors. Note the close connection between this question and the need to further advance error annotation schemes based on detailed analyses of inter-annotator agreement.

Language licensing approaches go beyond characterizing local patterns and attempt to analyze complete sentences. These so-called deep NLP approaches are based on fully explicit, formal grammars of the language to be licensed. Grammars essentially are compact representations of the wide range of lexical and syntactic possibilities of a language. To process with such grammars, efficient parsing algorithms are available to license a potentially infinite set of strings based on finite grammars. On the conceptual side, grammars can be expressed in two distinct ways (Johnson, 1994). In a ***validity-based grammar*** setup, a grammar is a set of rules. A string is recognized if and only if one can derive that string from the start symbol of the grammar. A grammar without rules licenses no strings, and essentially the more rules are added, the more different types of strings can be licensed. In a ***satisfiability-based grammar*** setup, a grammar consists of a set of constraints. A string is grammatical iff it satisfies all of the constraints in the grammar. A grammar without constraints thus licenses any string, and the more constraints are added, the fewer types of strings are licensed.

A number of linguistic formalisms have been developed for expressing such grammars, from basic context-free grammars lacking the ability to generalize across categories and rules to the modern lexicalized grammar formalisms for which efficient parsing approaches have been developed, such as Head-Driven Phrase Structure Grammar (HPSG), Lexical-Functional Grammar (LFG), Combinatory Categorial Grammar (CCG), and Tree-Adjoining Grammar (TAG) – cf. the grammar framework overviews in Brown (2006) and Müller (2013). To use any of these approaches in our context, we need to keep in mind that linguistic theories and grammars are generally designed to license well-formed native language, which raises the question how they can license learner language (and identify errors as part of the process).

There essentially are ***two types of approaches for licensing learner language***, corresponding to the two types of formal grammars introduced above. In a validity-based setup using a regular parser, so-called ***mal-rules*** can be added to the grammar (cf., e.g., Sleeman, 1982; Schwind, 1990; Covington & Weinrich, 1991; Matthews, 1992), to also license and thereby identify ill-formed strings occurring in the learner language. Particular types of errors can arise in a large number of rules, e.g., subject-verb agreement errors may need to be accommodated in any rule realizing subjects together with a finite verbal projection. Following Weischedel & Sondheimer (1983), ***meta-rules*** can be used to express such generalizations over rules.

Rule-based grammars license the infinite set of possible strings by modularizing the analysis into local trees. A local tree is a tree of depth one, i.e. a single mother node and its immediate children. Each local tree in the overall analysis of a sentence is independently licensed by a single rule in the grammar. Thus a *mal*-rule also only licenses a local tree, which then combined with other local trees licensed by other rules. The *mal*-rule approach is conceptually simple when the nature of an error can be captured within the local domain of a single rule, e.g., a *mal*-rule licensing the combination of an article and a noun disagreeing in gender (le_{masc} $table_{fem}$) can be added. Even so, the fact that rules and *mal*-rules in a grammar interact requires very careful grammar (re)writing to avoid unintended combinations. The situation is complicated further when the domain of an error is larger than a local tree. For example, extending the word order options of a rule *S* → *NP VP* by adding the *mal*-rule *S* → *VP NP* makes it possible to license (6a) and (6b). The order (6c), on the other hand, cannot be licensed in this way though, unless one writes an ad-hoc *mal*-rule licensing the entire tree (*S* → *V NP NP*), which would then have to be written for all other kind of VPs (intransitive, ditransitive, etc.) as well.

(6) a. Mary [loves cats].
b. * [loves cats] Mary.
c. * loves Mary cats.

Lexicalized grammar formalisms using richer data structures, such as the typed feature structure representation of signs used in HPSG, make it possible to encode more general types of *mal*-rules (e.g. Heift, 1998; Fortmann & Forst, 2004). Similarly, mildly context-sensitive frameworks such as TAG and CCG provide an extended domain of locality that could in principle be used to express *mal*-rules encoding errors occurring within those extended domains. For errors relating to the overall topology of the sentence (e.g., more than one unit preceding the verb in a German V2 sentence), one can also consider a separate layer of grammar encoding topological fields directly (Cheung & Penn, 2009).

To limit the search space explosion commonly resulting from rule interaction, the use of *mal*-rules can be limited. One option is to only include the *mal*-rules in processing when parsing a sentence with the regular grammar fails. This however only reduces the search space for well-formed strings. If parsing fails, the question which *mal*-rules need to be added is not addressed. An intelligent solution to this question was pursued by the ICICLE system (Michaud & McCoy, 2004). It selects groups of rules based on ***learner modeling***. For grammar constructs that the learner has shown mastery of, it uses the native rule set, but no rules are included for constructs beyond the developmental level of the learner. For structures currently being acquired, both the native rule set and the *mal*-rules relating to those phenomena are included. Finally, probabilistic grammar formalisms such as PCFGs and optimality theoretic mark-up of LFG grammars can be used to tune the licensing of grammatical and ungrammatical structures to learner language (cf. Wagner & Foster, 2009, and references therein), though sophisticated adaption to individual learners or learner groups would require a combination of careful domain adaptation and grammar (re)training with sophisticated learner modeling and activity design to support valid inferences on a learner's state of knowledge. The current state of learner modeling (Schulze, 2012) arguably could be advanced significantly by a tighter integration of expertise on FLTL task, SLA development, and sophisticated NLP techniques.

The second approach to licensing sentences which go beyond the native language grammars is based on ***constraint relaxation*** (Kwasny & Sondheimer, 1981). The approach is based on a satisfiability-based grammar setup or those rule-based grammar formalisms employing complex categories (e.g., feature structures, first order terms) for which the process of combining information (unification) and the enforcement of constraints can be relaxed. Instead of writing complete additional rules as in the *mal*-rule approach, constraint relaxation makes it possible to eliminate specific requirements of regular rules, thereby admitting additional structures normally excluded. For example, the feature specifications ensuring subject-verb agreement can be eliminated in this way to also license ungrammatical strings.

Relaxation works best when there is a natural one-to-one correspondence between a particular kind of error and a particular specification in the grammar, as in the case of subject-verb agreement errors being directly linked to the person and number specifications of finite verbs and their subject argument. In principle, one could also integrate a mechanism corresponding to the meta-rules of the *mal*-rule setup, in which the specifications of particular features are relaxed for particular sets of rules or constraints, or everywhere in the grammar. It would, however, be incorrect to claim that a constraint-relaxation approach does not require learner errors to be preenvisaged and therefore should be preferred over a *mal*-rule approach. It is necessary to distinguish those features or constraints which may be relaxed from those that are supposed to be hard, i.e., always enforced. Otherwise any learner sentence can be licensed with any structure and nothing is gained by parsing (or constraint-resolution being performed) at all.

Instead of completely eliminating constraints, constraints can also be associated with weights or probabilities with the goal of preferring or enforcing a particular analysis without ruling out ungrammatical sentences. One prominent example is the Weighted Constraint Dependency Grammar (WCDG) approach of Foth et al. (2005). Just as for the case of probabilistic grammar formalisms mentioned above, the as yet unsolved question raised by such approaches is how the weights can be obtained in a way that makes it possible to identify the likely error causes behind a given learner sentence written to complete a specific task by a given learner at a particular level of proficiency.

The constraint-relaxation research on learner error diagnosis has generally developed handcrafted formalisms and solutions. In principle, such an approach is closely related to Constraint Satisfaction Problems as a research domain in computer science. Boyd (2012) presents an approach that explores this connection and shows how learner error analysis can be compiled into a form that can be handled by general constraint resolvers, with diagnosis of learner errors being handled by approaches to detect conflicts in unconstrained constraint satisfaction problems.

While most approaches licensing learner language use standard parsing algorithms with extended or modified grammars to license ill-formed sentences, there is some work modifying the algorithmic side as well. For example, Reuer (2003) combines a constraint relaxation technique with a parsing algorithm modified to license strings in which words have been inserted or omitted, an idea which essentially moves generalizations over rules in the spirit of meta-rules into the parsing algorithm.

Let us conclude this discussion with a note on evaluation. Just like the analysis of inter-annotator agreement is an important evaluation criterion for the viability of the distinctions made by an error annotation scheme, the meaningful evaluation of grammatical error detection approaches is an important and underresearched area. One trend in this domain is to avoid the problem of gold standard error annotation as reference for testing by systematically introducing errors into native corpora (e.g. Foster, 2005). While this may be a good choice to monitor progress during development, related in spirit to the use of shallow measures such as BLEU for automatic machine translation (Papineni et al., 2001), such artificially created test sets naturally only reflect the properties of learner data in a very limited sense and do not eliminate the need to ultimately evaluate on authentic learner data with gold standard annotation. A good overview of the range of issues behind the difficulty of evaluating grammatical error detection systems is provided in (Chodorow et al., 2012).

Case studies

The following two case studies take a closer look at two representative approaches spelling out some of the general issues introduced above. The first case study focuses on a state-of-the-art learner corpus, for which detailed information on the integration of manual and automatic analysis as well as detailed inter-annotator agreement is available. The second case study provides a concrete example for error detection in the domain of word order errors, a frequent but underresearched type of error that at the same time allows us to exemplify how the nature of the phenomenon determines the choice of NLP analysis used for error detection.

Case study 1: Manual, semi-automatic, and automatic annotation of a learner corpus

To showcase the key components of a state-of-the-art learner corpus annotation project integrating insights and tools from NLP, we take a look at the Czech as Second Language (CzeSL) corpus, primarily based on Rosen et al. (2013). The corpus consists of 2.64 million words with written and transcribed spoken components, produced by foreign language learners of Czech at all levels of proficiency and by Roma acquiring Czech as a second language. So far, a sample of 370 thousand words from the written portion have been manually annotated.

The corpus is encoded in a multi-tier representation. Tier 0 encodes the learner text as such, tier 1 encodes a first target hypothesis in which all non-words are corrected, and tier 2 is a target hypothesis in which syntax, word order, and a few aspects of style are corrected. The differences between the tiers 0 and 1 and between the tiers 1 and 2 can be annotated with error tags. Depending on the nature of the error, the annotations link individual tokens across two tiers, or they can scope over multiple tokens, including discontinuous units.

The corpus setup is illustrated in Figure 1 as provided by Rosen et al. (2013).

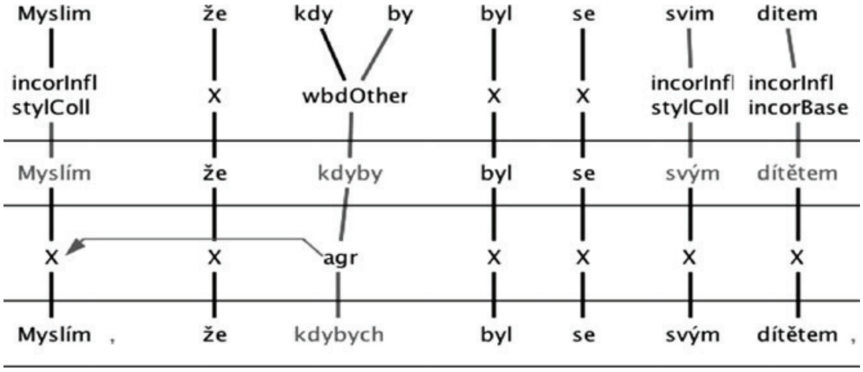

Figure 1: An example for the multi-tier representation of the CzeSL corpus

Tier 0 at the top is the sentence as written by the learner. This learner sentence includes several non-words, of which three require changes to the inflection or stem, and one requires two tokens to be merged into a single word. Tier 2 at the bottom of the figure then further corrects an agreement error to obtain the target hypothesis glossed in (7)

(7) Myslím, že kdybych byl se svým dítětem,
think$_{\mathrm{SG1}}$ that if$_{\mathrm{SG1}}$ was$_{\mathrm{MASC}}$ with my child,
‘I think that if I were with my child, . . . ’

The errors in individual word forms treated at tier 1 include misspellings, misplaced words, inflectional and derivational morphology, incorrect word stems, and invented or foreign words. The tier thus is closely related to the minimal form change target hypothesis we discussed in the first part of this article, but focuses exclusively on obtaining well-formed individual words rather than full syntactic forms. Such a dedicated tier for individual word forms is well-motivated considering the complex morphology of Czech. The target form encoded in tier 2 addresses errors in agreement, valency, analytical forms, word order, pronominal reference, negative concord, the choice of tense, aspect, lexical item or idiom. The manual annotation process is supported by a dedicated annotation tool *feat* (http://purl.org/net/feat).

The annotation started with a pilot annotation of 67 texts totalling almost 10 thousand tokens. 14 annotators were split into two groups, with each group annotating the sample independently. The Inter-Annotator Agreement (IAA) was computed using the standard Cohen’s kappa metric (cf. Artstein & Poesio, 2008). Since tier 1 and 2 can differ between annotators, for computing the IAA, error tags are projected onto tier 0 tokens. The feedback from the pilot was used to improve the annotation manual, the training of the annotators, and to modify the error taxonomy of the annotation scheme in a few cases. The annotation was then continued by 31 annotators who analyzed 1.396 texts totalling 175.234 words.

Both for the pilot and for the second annotation phase, a detailed quantitative and qualitative discussion of IAA results and confusion matrices is provided in Rosen et al. (2013), one aspect of which is shown here in Figure 2.

Tag	Type of error	Pilot sample		All annotated texts	
		κ	Avg. tags	κ	Avg. tags
*incor**	*incorBase+incorInfl*	0.84	1,038	0.88	14,380
incorBase	Incorrect stem	0.75	723	0.82	10,780
incorInfl	Incorrect inflection	0.61	398	0.71	4,679
*wbd**	*wbdPre+wbdOther+wbdComp*	0.21	37	0.56	840
wbdPre	Incorrect word boundary (prefix/preposition)	0.18	11	0.75	484
wbdOther	Incorrect word boundary	–	0	0.69	842
wbdComp	Incorrect word boundary (compound)	0.15	13	0.22	58
*fw**	*fw+fwFab+fwNc*	0.47	38	0.36	423
fwNc	Foreign/unidentified form	0.24	12	0.30	298
fwFab	Made-up/unidentified form	0.14	20	0.09	125
stylColl	Colloquial style at T1	0.25	8	0.44	1,396
agr	Agreement violation	0.54	199	0.69	2,622
dep	Syntactic dependency errors	0.44	194	0.58	3,064
rflx	Incorrect reflexive expression	0.26	11	0.42	141
lex	Lexical or phraseology error	0.37	189	0.32	1,815
neg	Incorrectly expressed negation	0.48	10	0.23	48
ref	Pronominal reference error	0.16	18	0.16	115
sec	Secondary (consequent) error	0.12	33	0.26	415
stylColl	Colloquial style at T2	0.42	24	0.39	633
use	Tense, aspect etc. error	0.22	84	0.39	696
vbx	Complex verb form error	0.13	15	0.17	233

Figure 2: Inter-Annotator Agreement for selected CzeSL error tags

We see that the annotators showed good agreement at tier 1 for incorrect morphology (incor*: $\kappa > 0.8$) and improper word boundaries (wbd*: $\kappa > 0.6$), and also for agreement errors (agr: $\kappa > 0.6$), syntactic dependency errors (dep: κ 0.58). Lexical errors and errors clearly modularized in the annotation scheme and described in the manual thus can reliably be annotated in a setup with an explicit, form-based target hypothesis.

On the other hand, pronominal reference (ref), secondary (follow-up) errors (sec), errors in analytical verb forms/complex predicates (vbx) and negation (neg) show a very low IAA level, as do tags for usage and lexical errors ($\kappa < 0.4$). Tags with a less clearly delineated interpretation, where it is difficult to establish a target hypothesis also showed low IAA, such as foreign/unidentified words or attempts to coin new Czech word (fw*: $\kappa < 0.4$).

The authors also conducted a detailed analysis of the relation between target hypotheses and error annotation agreement. They show that whether the annotators agreed on a target hypothesis or not strongly influences the IAA of the error annotation. For example, annotators agreed on agreement errors with a κ of 0.82 when their tier 1 target hypotheses agreed, but only 0.24 when their target hypotheses differed.

The confusion matrices and detailed analysis included in Rosen et al. (2013) in several places highlight generally relevant options for improving the annotation scheme and annotation manual further to avoid some of the remaining disagreement. In this context, the authors raise the question whether reducing disagreement among annotators is always desirable, suggesting that native speakers may disagree about the interpretation of a learner sentence. As far as we see this confuses two fundamentally distinct issues. The annotation of learner corpora is a scientific method for providing systematic access to well-defined classes of data.

It is not an experiment recording possible native speaker intuitions about learner language. As Rosen et al. (2013) itself highlights as the best IAA study on learner corpora to date, the goal is to be able to use annotation to replicably identify all instances of an agreed upon language property. If we, for example, leave the operationalization of the foreign word tag to the annotators intuition, as reminiscent of the early days of corpus annotation (Santorini, 1990, p. 2), the set of data this allows us to access is determined by who happened to annotate which corpus instances – instead of by a property of the language itself that the corpus user wants to systematically identify, count, and analyze all instances of. Annotations based on intuitions can essentially only form the basis of sociolinguistic studies of annotator attitudes, provided there is sufficient meta-data about who annotated what and individual properties of the annotators (age, gender, hobbies, etc.).

The ***manual annotation process is integrated with automatic tools in three ways*** in the CzeSL project. First, the availability of target hypotheses in the CzeSL corpus makes it possible to automatically obtain systematic linguistic analyses. The sentences at tier 2 of CzeSL in principle satisfy the grammatical regularities of native Czech, for which a range of NLP tools have been developed. Applying the standard taggers and lemmatizers, the authors thus obtain morphosyntactic categories and lemmas for the tier 2 target hypotheses. These annotations can then be projected back onto tier 1 and the original learner forms.

Second, some manually assigned error tags can automatically be enriched with further information. For example, for an incorrect base form or inflection that was manually annotated, an automatic process can diagnose the particular nature of the divergence and, e.g., tag it as an incorrect devoicing, a missing palatalization, or as an error in capitalization.

Third, automatic analysis is used to check the manual annotation. For example, learner forms at tier 0 that cannot be analyzed by the standard Czech morphological analyzer will generally require a correction at tier 1.

In addition to these three forms of integration of manual and automatic analysis, the authors also explored fully automatic annotation. They use tools originally developed for native Czech, the spell and grammar checker Korektor and two types of part-of-speech taggers. For the spell and grammar checker, among other aspects such as a special treatment of diacritics, the authors provide a comparison of the auto-correct mode of Korektor with the two target hypotheses spelled out in a the CzeSL corpus, based on a subset of almost 10 thousand tokens from the pilot set of annotated texts. Since Korektor integrates non-word spell checking with some grammar checking using limited context, the nature of the tool's output neither aligns with the purely lexical tier 1 of CzeSL nor tier 2 integrating everything from syntax and word order to style. Still, the authors report a precision of 74% and a recall of 71% for agreed-upon tier 1 annotations. For tier 2 the precision drops to 60% and recall to 45%. Overall, the authors interpret the results as sufficiently high to justify integrating the spell checker as a reference into the annotation workflow as an additional reference for the annotators.

For the part-of-speech taggers, they used two approaches based on different concepts, Marče (Votrubec, 2006) prioritizing lexical and morphological diagnostics over distributional context, and TnT (Brants, 2000) a trigram tagger essentially using the opposite strategy. Among a range of comparisons, the authors show that the different strategies indeed lead to significantly different results. The two tagging approaches agreed on the same tag in only 28.8% of the ill-formed tokens in a corpus sample of 12,681 tokens. Put in the context of our earlier discussion of the need to linguistically annotate learner language close to the observable evidence, this evaluation of standard part-of-speech tagging tools directly supports the relevance of a tripartite encoding of parts-of-speech for learner language, distinguishing morphological, distributional, and lexical stem information.

Case study 2: Word order errors and their different NLP requirements Our second case study is intended to illustrate the issues involved in answering the question which NLP techniques are needed for which kind of learner language analysis. It is based on work we carried out for automatically identifying word order errors in different types of exercises in an ICALL context (Metcalf & Meurers, 2006).

Language learners are known to produce a range of word order errors (cf., e.g., Odlin, 1989), and they are a frequent type of error. Word order also differs significantly across languages, so that transfer errors (cf., e.g., Selinker, 1972; Odlin, 2003) also play a role in this context. It is important for learners to master word order, which can also significantly complicate comprehension. This is exemplified by (8a) from the Hiroshima English Learners' Corpus (HELC, Miura, 1998), which is virtually incomprehensible, whereas the rearranged word order (8b) is already quite close to the target (8c) of this translation activity.

(8) a. He get to cleaned his son.
    b. He get his son to cleaned.
    c. He got his son to clean the room.

Word order errors are not uniform. Some involve lexical triggers (one of a finite set of words is known to occur) or indicative patterns, whereas others require a deeper linguistic analysis. Correspondingly, there essentially are two types of approaches for automatically identifying word order errors. On the one hand, an instance-based, shallow list and match approach, and on the other, a grammar-based, deep analysis approach. The issue can be exemplified using two aspects of English grammar with characteristic word order properties: phrasal verbs and adverbs.

For separable phrasal verbs, particles can precede or follow a full NP object (9), but they must follow a pronominal object (10). For inseparable phrasal verbs, particles always precede the object (11).

(9) a. wrote *down* the number
    b. wrote the number *down*
(10) a. * wrote *down* it
    b. wrote it *down*
(11) a. ran *into* {my neighbor, her }
    b. * ran {my neighbor, her } *into*

Authentic learner sentences instantiating these error patterns are shown in (12), taken from the Chinese Learner English Corpus (CLEC).

(12) a. * so they give up it
    b. * food which will build up him
    c. * rather than speed up it.
    d. * to pick up them.

Liao & Fukuya (2002) also show that Chinese learners of English avoid certain phrasal verb patterns and we found related patterns of avoidance in the CLEC, such as heavy use of a pattern that is always grammatical (*verb < particle < NP*), but little use of patterns restricted to certain verb and object types (e.g., *verb < pronoun < particle*).

Given that the relevant sets of particle verbs and their particles can readily be identified by the surface forms or by introducing a part-of-speech annotation including sufficiently detailed classes for the relevant subclasses of particle verbs, error patterns such as the ones in (13) and (14) (and the alternative avoidance patterns mentioned above) can easily be identified using regular expression matching.

(13) * wrote down it   separable-phrasal-verb < particle < pronoun

(14) * a. ran my neighbor into   inseparable-phrasal-verb < NP/pronoun < particle
  b. ran her into

Precedence (<) here can be specified using a regular expression allowing any number of words in-between (`.*`), though we are only interested in matches within the same sentence, so that a sentence segmented corpus (Palmer, 2000) is required. In sum, the patterns needed to identify thus are an instance of the regular expression patterns over words and part-of-speech tags within basic domains discussed in more detail in Meurers (2005, sec. 1.2).

The strength of such a shallow pattern matching approach is the simple and efficient processing. The weakness is its lack of generalization over tokens and patterns: All words (or parts-of-speech) for which order is to be checked must be known, and all relevant word orders must be preenvisaged and listed. As such, the approach works well for learner language patterns which are heavily restricted either by the targeted error pattern or through the activities in which the learner language arises, e.g., when analyzing learner language for restricted exercises such as “Build a Sentence” or “Translation” in German Tutor (Heift, 2001).

The placement of adverbs in English illustrates an error type for where such a shallow pattern approach is inadequate. English includes a range of different types of adverbs, and the word order possibilities depend on specific adverb subclass distinctions. For language learners, the rules governing adverb placement are difficult to notice and master, partly also because many adverb placements are not right or wrong, but more or less natural. As a result, students frequently misplace adverbs, as illustrated by the following examples from the Polish part of the International Corpus of Learner English (PICLE, 2004).

(15) a. they cannot already live without the dope.
  b. There have been already several campaigns held by ’Outdoor’.
  c. while any covert action brings rarely such negative connotations.
  d. It seems that the Earth has still a lot to reveal

To detect such errors, shallow pattern matching is inadequate. Many placements throughout a sentence are possible, and the possible error patterns are in principle predictable but very numerous. A compact characterization of the possible word orders and the corresponding error patterns requires reference to subclasses of adverbs and syntactic structure. A deep, grammar-based parsing approach can identify the necessary sentence structure, and the lexicon of the grammar can directly encode the relevant adverb subclasses.

Using the language licensing NLP approach introduced in the issues and methods overview of the paper, *mal*-rules can be added to a grammar so that a parser can license the additional word orders. Alternatively, one can manipulate the corresponding chart edges as in the approach of Reuer (2003). A downside of both approaches arises from the fact that phrase structure grammars express two things at once, at the level of the local syntactic tree: First, the generative potential (i.e., the combinatorics of which language items must co-occurs with which other items), and second, the word order regularities. Given that the word order possibilities are directly tied to the combinatorics, licensing more word orders significantly increases the size of the grammar and therefore the search space of parsing. In a lexicalized, constraint-based formalism such as HPSG, the position of an adverb can instead be constrained and recorded using a lexical principle governing the realization of the entire head domain instead of in a local tree. Similarly, a dedicated level recording the topological sentence structure (e.g., Cheung & Penn, 2009) may be used to modularize word order.

Summing up this second case study on the detection of word order errors, instanced-based matching is the right approach when lexical material and erroneous placements are predictable and listable and there is limited grammatical variation. Deep processing, on the other hand, is preferable when possible correct answers are predictable but not (conveniently) listable for a given activity or the predictable erroneous placements occur throughout a recursively built structure. In such a context, lexicalization and separate topological representations can be an attractive, modular alternative to phrase structure based encodings.

Future directions

Learner corpus research in the second language learning domain was heavily influenced by FLTL concerns, with limited connection to more theoretical SLA issues. One indication of this disconnect is the emphasis on learner errors in much learner corpus research, which runs counter to the characterization of learner language as a systematic interlanguage to be characterized in its own right as the established basis of SLA research over the past decades. Given that the corpus collection and representation methods have largely been settled, and the choice of NLP methods to be used for annotation and analysis depends on the nature of the research questions to be addressed, we are hopeful that learner corpus research will aim to connect and provide important contributions to the SLA mainstream. The availability of very large learner corpora collected as part of CALL environments, such as EFCAMDAT (Geerzen et al., 2013) should also make it possible to investigate important SLA issues with sufficient statistical power, including both cross-sectional and longitudinal study designs. A good example is the detailed study by Murakami (2013), who raises serious doubts about the morpheme studies as a hallmark of SLA research claiming that the order of acquisition essentially is independent of the native language of the learner.

To be able to observe a broader range of authentic learner language use, we also expect that learner corpora in the future will aim to record learner language arising include a broader range of tasks. To support interpretation of the observations of learner language across tasks, it will be important to make explicit and take into account the degree of well-formed and ill-formed variability that is supported by different tasks. Quixal (2012) provides a very detailed analysis framework incorporating insights from task-based learning and CALL research. He uses this framework to analyze and specify the capabilities of the NLP tools that are needed to process the learner language arising in those tasks (and, as the practical goal of the thesis, the capabilities needed for a CALL authoring system designed to give teachers the ability to assign feedback to particular classes of learner language). Integrating more task and learner information into the analysis of learner data also is in line with the prominent evidence-centered design approach in language assessment (Mislevy et al., 2003).

Including a wider range of tasks in learner corpora also strengthens the link between the analysis of learner corpora and that of online analyses of learner language, particularly in intelligent tutoring systems. This includes the automatic analysis of learner language to provide feedback in dialogue systems (Petersen, 2010; Wilske, 2013).

As a related trend, with the advent of a wider range of meaning-based tasks in learner corpora, when the broader question becomes how a learner uses the linguistic system to express a given meaning or function, the analysis of meaning will become an increasingly important aspect of the analysis of learner language, for which NLP techniques are increasingly successful (Dzikovska et al., 2013).

Given more grounding in meaning and function, it will increasingly become possible to analyze the use of forms under a variationist perspective, connecting linguistic variables with extra-linguistic variables, which, e.g., provide information on the learner. To be able to distinguish between language aspects determined by the task and those which are genuinely characteristic of learner language, it is relevant to observe learners where given the task they have a choice in the linguistic system. Such a variationist analysis can help us identify the static core of the linguistic system and the language variables for which we can employ the methodology of variationist sociolinguistics to study the variants correlating with the learner characteristics and other extra-linguistic variables.

In terms of the nature of the NLP resources and algorithms employed for the analysis of learner language, the exploration of the role of linguistic modeling will increasingly become important. As we, for example, discuss in Meurers et al. (2013); Bykh et al. (2013), a task such as Native Language Identification makes it possible to quantitatively evaluate where a surface-based NLP analysis is successful and sufficiently general to transfer to previously unseen data, new topics and genres, and where deeper abstractions are needed to capture aspects of the linguistic system which are not directly dependent on the topic or genre.

The question which aspects of linguistic modeling are relevant for which type of application and analysis can also be seen to be gaining ground in the overall NLP and education domain. For example, automatic essay scoring is an application where essays traditionally are classified on the basis of a holistic comparison of the new essay with essays for the same prompt for which gold-standard labels were manually assigned. The recent work on proficiency classification (Pendar & Chapelle, 2008; Yannakoudakis et al., 2011; Vajjala & Lõo, 2013; Hancke & Meurers, 2013), on the other hand, emphasizes lexical, syntactic, semantic, and discourse aspects of complexity of the linguistic system rather than task and prompt-specific training.

The trend to increase the linguistic modeling in this domain seems to be even stronger when less language material is available. In automatic essay assessment, much of the weight is carried by methods relying on the occurrence of lexical material, such as Latent Semantic Analysis. When the analysis must be performed on less language material, e.g., when analyzing short answers to reading comprehension questions consisting of one or a couple of sentences, the lexical material by itself is not informative enough. For the C-Rater system (Leacock & Chodorow, 2003) dealing with such data, very specific grading information is provided by the item designers for each item. Where no such additional, item-specific information is available, approaches must make the most of the limited amount of language data in short answers, which amounts to employing a wide range of linguistic modeling for automatic meaning assessment, as apparent in the range of approaches to the recent shared task (Dzikovska et al., 2013). Interestingly, such a task also makes it possible to identify how much the different aspects of linguistic modeling contribute to the analysis of the learner language. For example, Hahn & Meurers (2012) show that in the CoSeC system assessing German short answers in reading comprehension tasks at 86.3% accuracy, 5.6% were contributed by the dependency relations between words and 5.4% resulted from modeling information structure to distinguish focused elements in the answer.

Increasing the breadth and depth of linguistic modeling of learner data also stands to contribute to related, applied contexts. In language assessment, when items are created to test a given construct, one typically tests which items show the highest correlation and eliminates the other items from the test. The linguistic question which properties of the item contribute in which way to this high or low correlation has not traditionally been asked. Adding more linguistic modeling in the analysis of learner corpora ultimately may contribute to more linguistically controlled item design or, e.g., explicit multi-layered models of the complexity of specific gaps in C-tests, as commonly used for placement testing.

While most of the initial NLP work on corpus annotation and error detection focused on English, we also start to see some work on other languages, such as Korean (Israel et al., 2013), and the increased NLP interest in underresourced languages and endangered languages is likely to support more growth in that area in the future.

As a final point emphasizing the relevance of interdisciplinary collaboration in this domain again, there currently is a surprising lack of interaction between NLP research related to first language acquisition and that on second language acquisition. Many of the representation, modeling, and algorithmic issues are the same or closely related, both in conceptual terms and also in practical terms of jointly developing and using the same tools. There is some precedence, such as the use of CHILDES tools for SLA research (Myles & Mitchell, 2004), but significantly more opportunity for synergy in future work.

Key readings

Complementing the publications discussed as part of the specific issues discussed in the paper, the following books and article provide a useful background for the reader interested in a deeper foundation for work at the intersection of NLP and Learner Corpora.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Second edition. Upper Saddle River, NJ: Prentice Hall.

> The book provides an accessible, broad introduction to Natural Language Processing. It covers the key issues, representations, and algorithms from both the theory-driven, logic-based perspectives and the data-driven, statistical perspectives. It also includes clear discussions of current applications, such as machine translation.

Markus Dickinson, Chris Brew and Detmar Meurers. 2013. *Language and Computers*. Wiley-Blackwell.

> This is an introduction for the reader interested in exploring the issues in computational linguistics starting from the different NLP applications, from search engines via dialog systems to machine translation. It includes a chapter on Language Tutoring Systems which emphasizes the motivation of the linguistic modeling involved in learner language analysis and the corresponding NLP analysis techniques.

Trude Heift and Mathias Schulze. 2003. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

> The authors provide a very comprehensive historical perspective of the field of Intelligent Computer-Assisted Language Learning at the intersection of Computer-Assisted Language Learning and Natural Language Processing.

Claudia Leacock, Martin Chodorow, Michael Gamon and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*, Volume 3 of Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst. Morgan & Claypool Publishers.

> Starting with a characterization of the grammatical constructions that second language learners of English find most difficult to master, the book introduces the techniques used to automatically detect errors in learner writing. It discusses the research issues and approaches, how to report results to ensure sustained progress in the field, and the use of annotated learner corpora in that context.

Roger Garside, Geoffrey Leech and Tony McEnery. 1997. *Corpus annotation: linguistic information from computer text corpora*, Harlow, England: Addison Wesley Longman Ltd.

> The book provides an accessible introduction to linguistic corpus annotation. It includes a discussion of lexical, syntactic, semantic and discourse annotation as well as of methodological aspects concerning the annotation quality and domain adaptation.

Detmar Meurers. 2005. On the use of electronic corpora for theoretical linguistics. *Lingua*. 115 (11). 1619–1639. http://purl.org/dm/papers/meurers-03.html

> What exactly is involved in using corpora to address linguistic research issues? The article discusses how the linguistic terminology used in formulating research questions can be translated to the annotation found in a corpus and the conceptual and methodological issues arising in this context. Readers who are particularly interested in syntactically annotated corpora can continue with the discussion with Meurers & Müller (2009), where such treebanks are the main source of data used.

Ron Artstein and Massimo Poesio. 2008. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34 (4). 555–596. Extended version: http:/cswww.essex.ac.uk/Research/nle/arrau

> Corpus annotation is only useful for scientific work if it provides a reliable, replicable index into the data. Investigating the Inter-Code Agreement between multiple independent annotators, is the most important method for determining whether the classes operationalized and documented in the annotation scheme can reliably be distinguished solely based on the information that is available in the corpus and its meta-information. The article provides the essential methodological background for investigating the Inter-Coder Agreement, which in the corpus context typically is referred to as Inter-Annotator Agreement.

References

Abney, S. (1997). Partial Parsing via Finite-State Cascades. *Natural Language Engineering* 2, 337–344. URL http://www.vinartus.net/spa/97a.pdf.

Alexopoulou, T., H. Yannakoudakis & T. Briscoe (2011). From discriminative features to learner grammars: a data driven approach to learner corpora. Second Language Forum Research, Maryland. URL http://people.pwf.cam.ac.uk/ta259/SLRF10-dora.pdf.

Amaral, L. & D. Meurers (2009). Little Things With Big Effects: On the Identification and Interpretation of Tokens for Error Diagnosis in ICALL. *CALICO Journal* 26(3), 580–591. URL http://purl.org/dm/papers/amaral-meurers-09.html.

Artstein, R. & M. Poesio (2008). Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4), 555–596.

Bachman, L. F. & A. S. Palmer (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press.

Bailey, S. (2008). Content Assessment in Intelligent Computer-Aided Language Learning: Meaning Error Diagnosis for English as a Second Language. Ph.D. thesis, The Ohio State University. URL http://purl.org/net/Bailey-08.pdf.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33(1), 1–17.

Böhmová, A., J. Hajič, E. Hajičová & B. Hladká (2003). The Prague Dependency Treebank. In A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*, Dordrecht: Kluwer, chap. 7, pp. 103–127.

Boyd, A. (2010). EAGLE: an Error-Annotated Corpus of Beginning Learner German. In *Proceedings of LREC'10*. Valletta, Malta. URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/812.html.

Boyd, A., M. Dickinson & D. Meurers (2008). On Detecting Errors in Dependency Treebanks. *Research on Language and Computation* 6(2), 113–137. URL http://purl.org/dm/papers/boyd-et-al-08.html.

Boyd, A. A. (2012). Detecting and Diagnosing Grammatical Errors for Beginning Learners of German: From Learner Corpus Annotation to Constraint Satisfaction Problems. Ph.D. thesis, The Ohio State University. URL http://rave.ohiolink.edu/etdc/view?acc_num=osu1325170396.

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Association for Computational Linguistics, pp. 224–231. URL http://aclweb.org/anthology/A00-1031.

Brown, K. (ed.) (2006). *Encyclopedia of Language and Linguistics*. Oxford: Elsevier, 2 ed.

Bykh, S., S. Vajjala, J. Krivanek & D. Meurers (2013). Combining Shallow and Linguistically Motivated Features in Native Language Identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA. URL http://purl.org/dm/papers/Bykh.Vajjala.ea-13.html.

Canale, M. & M. Swain (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics* 1, 1–47. URL http://applij.oxfordjournals.org/cgi/reprint/I/1/1.pdf.

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

Cheung, J. C. K. & G. Penn (2009). Topological field parsing of German. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 64–72. URL http://aclweb.org/anthology/P09-1008.

Chodorow, M., M. Dickinson, R. Israel & J. Tetreault (2012). Problems in Evaluating Grammatical Error Detection Systems. In *Proceedings of COLING 2012*. Mumbai, India, pp. 611–628. URL http://cl.indiana.edu/~md7/papers/chodorow-et-al12.html.

Covington, M. A. & K. B. Weinrich (1991). Unification-based Diagnosis of Language Learners' Syntax Errors. *Literary and Linguistic Computing* 6(3), 149–154. URL http://llc.oxfordjournals.org/cgi/reprint/6/3/149.pdf.

De Felice, R. (2008). Automatic Error Detection in Non-native English. Ph.D. thesis, St Catherine's College, University of Oxford.

De Felice, R. & S. Pulman (2009). Automatic Detection of Preposition Errors in Learner Writing. *CALICO Journal* 26(3), 512–528. URL https://www.calico.org/a-758-Automatic%20Detection%20of%20Preposition%20Errors%20in%20Learner%20Writing.html.

de Haan, P. (2000). Tagging non-native English with the TOSCA-ICLE tagger. In Mair & Hundt (2000), pp. 69–79.

de Mönnink, I. (2000). Parsing a learner corpus. In Mair & Hundt (2000), pp. 81–90.

Díaz Negrillo, A., D. Meurers, S. Valera & H. Wunsch (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* 36(1–2), 139–154. URL http://purl.org/dm/papers/diaz-negrillo-et-al-09.html.

Dickinson, M. (2006). Writer's Aids. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, Oxford: Elsevier. 2 ed.

Dickinson, M. & W. D. Meurers (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)*. Budapest, Hungary, pp. 107–114. URL http://purl.org/dm/papers/dickinson-meurers-03.html.

Dickinson, M. & M. Ragheb (2009). Dependency Annotation for Learner Corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*. Milan, Italy. URL http://jones.ling.indiana.edu/~mdickinson/papers/dickinson-ragheb09.html.

Dzikovska, M., R. Nielsen et al. (2013). SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 263–274. URL http://www.aclweb.org/anthology/S13-2045.

Díaz Negrillo, A. (2007). A Fine-Grained Error Tagger for Learner Corpora. Ph.D. thesis, University of Jaén, Spain.

Díaz Negrillo, A. & J. Fernández Domínguez (2006). Error Tagging Systems for Learner Corpora. *Revista Española de Lingüística Aplicada (RESLA)* 19, 83–102. URL http://dialnet.unirioja.es/servlet/fichero_articulo?codigo=2198610&orden=72810.

Fitzpatrick, E. & M. S. Seegmiller (2004). The Montclair electronic language database project. In U. Connor & T. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, Amsterdam: Rodopi. URL http://chss.montclair.edu/linguistics/MELD/rodopipaper.pdf.

Fortmann, C. & M. Forst (2004). An LFG grammar checker for CALL. In R. Delmonte (ed.), *InSTIL/ICALL 2004 Symposium on Computer Assisted Learning, NLP and speech technologies in advanced language learning systems*. Venice, Italy: International Speech Communication Association (ISCA). URL ftp://www.ims.uni-stuttgart.de/pub/Users/forst/Fortmann:Forst-ICALL04.pdf.

Foster, J. (2005). Good Reasons for Noting Bad Grammar: Empirical Investigations into the Parsing of Ungrammatical Written English. Ph.D. thesis, Trinity College Dublin, Department of Computer Science.

Foth, K., W. Menzel & I. Schröder (2005). Robust parsing with weighted constraints. *Natural Language Engineering* 11(01), 1–25.

Foth, K. A. & W. Menzel (2006). Hybrid parsing: using probabilistic models as predictors for a symbolic parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, ACL-44, pp. 321–328. URL http://dx.doi.org/10.3115/1220175.1220216.

Gamon, M., C. Leacock, C. Brockett, W. B. Dolan, J. Gao, D. Belenko & A. Klementiev (2009). Using Statistical Techniques and Web Search to Correct ESL Errors. *CALICO Journal* 26(3), 491–511. URL http://research.microsoft.com/pubs/81312/Calico_published.pdf;https://www.calico.org/a-757-Using%20Statistical%20Techniques%20and%20Web%20Search%20to%20Correct%20ESL%20Errors.html.

Garside, R., G. Leech & T. McEnery (eds.) (1997). *Corpus annotation: linguistic information from computer text corpora*. Harlow, England: Addison Wesley Longman Limited.

Geerzen, J., T. Alexopoulou & A. Korhonen (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. URL http://corpus.mml.cam.ac.uk/efcamdat.

Granfeldt, J., P. Nugues, E. Persson, L. Persson, F. Kostadinov, M. Ågren & S. Schlyter (2005). Direkt Profil: A System for Evaluating Texts of Second Language Learners of French Based on Developmental Sequences. In J. Burstein & C. Leacock (eds.), *Proceedings of the Second Workshop on Building Educational Applications Using NLP*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 53–60. URL http://aclweb.org/anthology/W05-0209.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20(3), 465–480. URL http://purl.org/calico/granger03.pdf.

Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics. An international handbook*, Berlin, New York: Walter de Gruyter, pp. 259–275.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot (2009). *International Corpus of Learner English, Version 2*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Granger, S., O. Kraif, C. Ponton, G. Antoniadis & V. Zampa (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL* 19(3).

Granger, S. & F. Meunier (1994). Towards a grammar checker for learners of English. In U. Fries, G. Tottie & P. Schneider (eds.), *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zurich 1993*, Amsterdam: Rodopi, pp. 79–91.

Hahn, M. & D. Meurers (2011). On deriving semantic representations from dependencies: A practical approach for evaluating meaning in learner corpora. In *Proceedings of the Intern. Conference on Dependency Linguistics (DEPLING 2011)*. Barcelona, pp. 94–103. URL http://purl.org/dm/papers/hahn-meurers-11.html.

Hahn, M. & D. Meurers (2012). Evaluating the Meaning of Answers to Reading Comprehension Questions: A Semantics-Based Approach. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*. Montreal, pp. 94–103. URL http://purl.org/dm/papers/hahn-meurers-12.html.

Hancke, J. & D. Meurers (2013). Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the International Learner Corpus Research Conference (LCR-2013)*. Bergen. To appear.

Hancke, J., D. Meurers & S. Vajjala (2012). Readability Classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbay, India, pp. 1063–1080. URL http://aclweb.org/anthology-new/C/C12/C12-1065.pdf.

Heift, T. (1998). Designed Intelligence: A Language Teacher Model. Ph.D. thesis, Simon Fraser University. URL http://www.sfu.ca/~heift/pubs.html.

Heift, T. (2001). Error-Specific and Individualized Feedback in a Web-based Language Tutoring System: Do They Read It? *ReCALL* 13(2), 129–142. URL http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=82591&fulltextType=RA&fileId=S095834400100091X.

Hirschmann, H., S. Doolittle & A. Lüdeling (2007). Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*. Birmingham. URL http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/neu2/mitarbeiter-innen/anke/pdf/HirschmannDoolittleLuedelingCL2007.pdf.

Hirschmann, H., A. Lüdeling, I. Rehbein, M. Reznicek & A. Zeldes (2010). Syntactic Overuse and Underuse: A Study of a Parsed Learner Corpus and its Target Hypothesis. Presentation given at the Treebanks and Linguistic Theory Workshop.

Höhle, T. N. (1986). Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In A. Schöne (ed.), *Kontroversen alte und neue. Akten des VII. Internationalen Germanistenkongresses Göttingen 1985*, Tübingen: Niemeyer, pp. 329–340. Bd. 3.

Huizhong, Y. & G. Shichun (2004). Chinese Learner English Corpus (CLEC).

Israel, R., M. Dickinson & S.-H. Lee (2013). Detecting and Correcting Learner Korean Particle Omission Errors. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*. Nagoya, Japan. URL http://cl.indiana.edu/~md7/papers/israel-dickinson-lee13.html.

Johnson, M. (1994). Two ways of formalizing grammars. *Linguistics and Philosophy* 17(3), 221–248.

King, L. & M. Dickinson (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. URL http://cl.indiana.edu/~md7/papers/king-dickinson13.html.

Kintsch, W. & P. Mangalath (2011). The Construction of Meaning. *Topics in Cognitive Science* 3(2), 346–370. URL http://dx.doi.org/10.1111/j.1756-8765.2010.01107.x.

Klein, W. & C. Perdue (1997). The Basic Variety (or: Couldn't natural languages be much simpler?). *Second language research* 13(4), 301–347.

Krivanek, J. & D. Meurers (2011). Comparing Rule-Based and Data-Driven Dependency Parsing of Learner Language. In *Proceedings of the Intern. Conference on Dependency Linguistics (DEPLING 2011)*. Barcelona, pp. 310–317.

Kwasny, S. C. & N. K. Sondheimer (1981). Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems. *American Journal of Computational Linguistics* 7(2), 99–108. URL http://portal.acm.org/citation.cfm?id=972892.972894.

Leacock, C. & M. Chodorow (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities* 37, 389–405. URL http://www.ingentaconnect.com/content/klu/chum/2003/00000037/00000004/05144721?crawler=true.

Leech, G. (2004). Chapter 2. Adding Linguistic Annotation. In M. Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford: Oxbow Books. URL http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm.

Liao, Y. D. & Y. J. Fukuya (2002). Avoidance of phrasal verbs: The case of Chinese learners of English. *Second Language Studies* 20(2), 71–106. URL http://www.hawaii.edu/sls/uhwpesl/20(2)/Liao&Fukuya.pdf.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics* 14(1), 3–28.

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

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

Lüdeling, A. & M. Kytö (eds.) (2008). *Corpus Linguistics. An International Handbook*, vol. 1 of *Handbooks of Linguistics and Communication Science*. Berlin: Mouton de Gruyter. URL http://elpub.bib.uni-wuppertal.de/servlets/DerivateServlet/Derivate-2140/29-1.pdf.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. Vol 1: The Format and Programs, Vol 2: The Database*. Mahwah, NJ: Lawrence Erlbaum Associates, 3rd ed.

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (ed.), *Corpora in Language Acquisition Research: History, Methods, Perspectives*, Amsterdam and Philadelphia: John Benjamins, vol. 6 of *Trends in Language Acquisition Research*, pp. 165–197. URL http://childes.psy.cmu.edu/grasp/morphosyntax.doc.

Mair, C. & M. Hundt (eds.) (2000). *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi.

Matthews, C. (1992). Going AI: Foundations of ICALL. *Computer Assisted Language Learning* 5(1), 13–31.

McEnery, T., R. Xiao & Y. Tono (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge. URL http://www.ulb.tu-darmstadt.de/tocs/128706848.pdf.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press. URL http://books.google.com/books?id=diq29vrjAa4C&lpg=PR13&ots=ZcCJBmEA7g&dq=Dependency%20Syntax%3A%20Theory%20and%20Practice&lr&pg=PR13#v=onepage&q&f=false.

Metcalf, V. & D. Meurers (2006). When to Use Deep Processing and When Not To – The Example of Word Order Errors. URL http://purl.org/net/icall/handouts/calico06-metcalf-meurers.pdf. Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006. May 17, 2006. University of Hawaii.

Meurers, D. (2012). Natural Language Processing and Language Learning. In Chapelle (2012). URL http://purl.org/dm/papers/meurers-12.html.

Meurers, D., J. Krivanek & S. Bykh (2013). On the Automatic Analysis of Learner Corpora: Native Language Identification as Experimental Testbed of Language Modeling between Surface Features and Linguistic Abstraction. In *Diachrony and Synchrony in English Corpus Studies*. Frankfurt am Main: Peter Lang. To appear.

Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* 115(11), 1619–1639. URL http://purl.org/dm/papers/meurers-03.html.

Meurers, W. D. & S. Müller (2009). Corpora and Syntax (Article 42). In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics*, Berlin: Mouton de Gruyter, vol. 2 of *Handbooks of Linguistics and Communication Science*, pp. 920–933. URL http://purl.org/dm/papers/meurers-mueller-09.html.

Michaud, L. N. & K. F. McCoy (2004). Empirical Derivation of a Sequence of User Stereotypes for Language Learning. *User Modeling and User-Adapted Interaction* 14(4), 317–350. URL http://www.springerlink.com/content/lp86123772372646/.

Mislevy, R. J., R. G. Almond & J. F. Lukas (2003). *A Brief Introduction to Evidence-centered Design*. ETS Research Report RR-03-16, Educational Testing Serice (ETS), Princeton, NJ, USA. URL http://www.ets.org/Media/Research/pdf/RR-03-16.pdf.

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/eigo1.html, http://purl.org/icall/eigo2.html,.

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

Myles, F. & R. Mitchell (2004). Using information technology to support empirical SLA research. *Journal of Applied Linguistics* 1(2), 169–196. URL http://www.equinoxjournals.com/JAL/article/viewArticle/1444.

Müller, S. (2013). *Grammatiktheorie*. No. 20 in Stauffenburg Einführungen. Tübingen: Stauffenburg Verlag, second ed. URL http://hpsg.fu-berlin.de/~stefan/Pub/grammatiktheorie.html.

Naber, D. (2003). A Rule-Based Style and Grammar Checker. Master's thesis, Universität Bielefeld. URL http://www.danielnaber.de/publications.

Nivre, J., J. Nilsson, J. Hall, A. Chanev, G. Eryigit, S. Kübler, S. Marinov & E. Marsi (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering* 13(1), 1–41. URL http://w3.msi.vxu.se/~nivre/papers/nle07.pdf.

Odlin, T. (1989). *Language Transfer: Cross-linguistic influence in language learning*. New York: CUP.

Odlin, T. (2003). Cross-linguistic Influence. In C. Doughty & M. Long (eds.), *Handbook on Second Language Acquisition*, Oxford: Blackwell, pp. 436–486.

Ott, N. & R. Ziai (2010). Evaluating Dependency Parsing Performance on German Learner Language. In M. Dickinson, K. Müürisep & M. Passarotti (eds.), *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*. vol. 9 of *NEALT Proceeding Series*, pp. 175–186. URL http://hdl.handle.net/10062/15960.

Ott, N., R. Ziai & D. Meurers (2012). Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context. In T. Schmidt & K. Wörner (eds.), *Multilingual Corpora and Multilingual Corpus Analysis*, Amsterdam: Benjamins, Hamburg Studies in Multilingualism (HSM), pp. 47–69. URL http://purl.org/dm/papers/ott-ziai-meurers-12.html.

Palmer, D. D. (2000). Tokenisation and Sentence Segmentation. In R. Dale, H. Moisl & H. Somers (eds.), *Handbook of Natural Language Processing*, New York: Marcel Dekker, pp. 11–35. URL http://www.netLibrary.com/ebook_info.asp?product_id=47610.

Papineni, K., S. Roukos, T. Ward & W.-J. Zhu (2001). *BLEU: A Method for Automatic Evaluation of Machine Translation*. Tech. rep., IBM Research Division, Thomas J. Watson Research Center.

Pendar, N. & C. Chapelle (2008). Investigating the Promise of Learner Corpora: Methodological Issues. *CALICO Journal* 25(2), 189–206. URL https://calico.org/html/article_689.pdf.

Perdue, C. (ed.) (1993). *Adult Language Acquisition. Cross-Linguistic Perspectives. Volume 1: Field methods.* Cambridge, UK: Cambridge University Press.

Petersen, K. (2010). Implicit Corrective Feedback in Computer-Guided Interaction: Does Mode Matter? Ph.D. thesis, Georgetown University. URL http://apps.americancouncils.org/transfer/KP_Diss/Petersen_Final.pdf.

PICLE (2004). Polish portion of the International Corpus of Learner English. Web interface to Corpus. URL http://elex.amu.edu.pl/~przemka/concord2advr/search_adv_new.html.

Quixal, M. (2012). Language Learning Tasks and Automatic Analysis of Learner Language. Connecting FLTL and NLP in the design of ICALL materials supporting effective use in real-life instruction. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona and Eberhard-Karls-Universität Tübingen.

Ragheb, M. & M. Dickinson (2012). Defining Syntax for Learner Language Annotation. In *Proceedings of COLING 2012: Posters*. Mumbai, India, pp. 965–974. URL http://cl.indiana.edu/~md7/papers/ragheb-dickinson12.html.

Ragheb, M. & M. Dickinson (2013). Inter-annotator Agreement for Dependency Annotation of Learner Language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, GA USA. URL http://cl.indiana.edu/~md7/papers/ragheb-dickinson13.html.

Rahkonen, M. & G. Håkansson (2008). Production of Written L2-Swedish – Processability or Input Frequencies? In J.-U. Keßler (ed.), *Processability Approaches to Second Language Development and Second Language Learning*, Cambridge Scholars Publishing, pp. 135–161.

Reuer, V. (2003). Error recognition and feedback with Lexical Functional Grammar. *CALICO Journal* 20(3), 497–512. URL https://www.calico.org/a-291-Error%20Recognition%20and%20Feedback%20with%20Lexical%20Functional%20Grammar.html.

Reznicek, M., A. Lüdeling & H. Hirschmann (forthcoming). Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture. In A. Díaz-Negrillo, N. Ballier & P. Thompson (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*, John Benjamins.

Reznicek, M., A. Lüdeling, C. Krummes & F. Schwantuschke (2012). *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.0.* URL http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuchv2.0.pdf.

Rimrott, A. & T. Heift (2008). Evaluating automatic detection of misspellings in German. *Language Learning and Technology* 12(3), 73–92. URL http://llt.msu.edu/vol12num3/rimrottheift.pdf.

Rosen, A., J. Hana, B. Štindlová & A. Feldman (2013). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation* pp. 1–28. URL http://dx.doi.org/10.1007/s10579-013-9226-3.

Rosén, V. & K. D. Smedt (2010). Syntactic Annotation of Learner Corpora. In H. Johansen, A. Golden, J. E. Hagen & A.-K. Helland (eds.), *Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag [Systematic, varied, but not arbitrary. Anthology about Norwegian as a second language on the occasion of Kari Tenfjord's 60th birthday]*, Oslo: Novus forlag, pp. 120–132.

Rozovskaya, A. & D. Roth (2010). Annotating ESL errors: Challenges and rewards. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-5) at NAACL-HLT 2010*. Los Angeles: Association for Computational Linguistics.

Sagae, K., E. Davis, A. Lavie & B. M. an Shuly Wintner (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language* 37(3), 705–729.

Sagae, K., E. Davis, A. Lavie, B. MacWhinney & S. Wintner (2007). High-accuracy Annotation and Parsing of CHILDES Transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*. Prague, Czech Republic: Association for Computational Linguistics, pp. 25–32. URL http://aclweb.org/anthology/W07-0604.

Santorini, B. (1990). *Part-of-speech Tagging Guidelines for the Penn Treebank, 3rd Revision, 2nd Printing*. Tech. rep., Department of Computer Science, University of Pennsylvania. URL http://www.cs.bgu.ac.il/~nlpproj/nlp02/papers/treebank-tagset.pdf.

Schulze, M. (2012). Learner Modeling in Intelligent Computer-Assisted Language Learning. In Chapelle (2012).

Schwind, C. B. (1990). An Intelligent Language Tutoring System. *International Journal of Man-Machine Studies* 33, 557–579. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WGS-4MFTKNC-5&_user=1634476&_coverDate=11%2F30%2F1990&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000054037&_version=1&_urlVersion=0&_userid=1634476&md5=d68e9cad2ee1f9294f9fd876af9de9c4&searchtype=a.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics* 10(3), 209–231. URL http://www.reference-global.com/doi/abs/10.1515/iral.1972.10.1-4.209.

Shermis, M. D. & J. Burstein (eds.) (2013). *Handbook on Automated Essay Evaluation: Current Applications and New Directions.* London and New York: Routledge, Taylor & Francis Group.

Sleeman, D. (1982). Inferring (mal) rules from pupil's protocols. In *Proceedings of ECAI-82*. Orsay, France, pp. 160–164.

Tetreault, J., D. Blanchard & A. Cahill (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.

Tetreault, J. & M. Chodorow (2008a). Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection. In *Proceedings of the workshop on Human Judgments in Computational Linguistics at COLING-08*. Manchester, UK: Association for Computational Linguistics, pp. 24–32. URL http://www.ets.org/Media/Research/pdf/h4.pdf.

Tetreault, J. & M. Chodorow (2008b). The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*. Manchester, UK: Association for Computational Linguistics, pp. 865–872. URL http://www.ets.org/Media/Research/pdf/r3.pdf.

Tono, Y. (forthcoming). Criterial feature extraction using parallel learner corpora and machine learning. In A. Díaz-Negrillo, N. Ballier & P. Thompson (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*, John Benjamins.

Tono, Y., E. Izumi & E. Kaneko (2004). The NICT JLE Corpus: the final report. In K. Bradford-Watts, C. Ikeguchi & M. Swanson (eds.), *Proceedings of the JALT Conference*. Tokyo: JALT. URL http://jalt-publications.org/archive/proceedings/2004/E126.pdf.

Vajjala, S. & K. Lõo (2013). Role of Morpho-syntactic features in Estonian Proficiency Classification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8), Association for Computational Linguistics*.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In J. Tetreault, J. Burstein & C. Leacock (eds.), *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*. Montréal, Canada: Association for Computational Linguistics, pp. 163—-173. URL http://aclweb.org/anthology/W12-2019.pdf.

van Rooy, B. & L. Schäfer (2002). The Effect of Learner Errors on POS Tag Errors during Automatic POS Tagging. *Southern African Linguistics and Applied Language Studies* 20, 325–335.

van Rooy, B. & L. Schäfer (2003). An Evaluation of Three POS Taggers for the Tagging of the Tswana Learner English Corpus. In D. Archer, P. Rayson, A. Wilson & T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 conference Lancaster University (UK), 28 – 31 March 2003*. vol. 16 of *University Centre For Computer Corpus Research On Language Technical Papers*, pp. 835–844. URL http://www.corpus4u.org/upload/forum/2005092023174960.pdf.

Votrubec, J. (2006). Morphological tagging based on averaged perceptron. In *WDS'06 proceedings of contributed papers*. Praha, Czechia: Matfyzpress, Charles University, pp. 191—-195. URL http://www.mff.cuni.cz/veda/konference/wds/proc/pdf06/WDS06_134_i3_Votrubec.pdf.

Wagner, J. & J. Foster (2009). The effect of correcting grammatical errors on parse probabilities. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*. Paris, France: Association for Computational Linguistics, pp. 176–179. URL http://aclweb.org/anthology/W09-3827.

Weischedel, R. M. & N. K. Sondheimer (1983). Meta-rules as a Basis for Processing Ill-formed Input. *Computational Linguistics* 9(3-4), 161–177. URL http://aclweb.org/anthology/J83-3003.

Wichmann, A. (2008). Speech corpora and spoken corpora. In Lüdeling & Kytö (2008), pp. 187–207. URL http://elpub.bib.uni-wuppertal.de/servlets/DerivateServlet/Derivate-2140/29-1.pdf.

Wilske, S. (2013). Form and meaning in dialogue-based computer-assisted language learning. Ph.D. thesis, Universität des Saarlandes, Saarbrücken. To appear.

Winograd, T. & F. Flores (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Norwood, NJ: Ablex.

Yannakoudakis, H., T. Briscoe & B. Medlock (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, HLT '11, pp. 180–189. URL http://aclweb.org/anthology/P11-1019.pdf. Corpus available: http://ilexir.co.uk/applications/clc-fce-dataset.