24

Learner corpora and natural language processing

Detmar Meurers

1 Introduction

Natural Language Processing (NLP) deals with the representation and the automatic analysis and generation of human language (Jurafsky and Martin 2009). Learner corpora collect the language produced by people learning a language. The two thus overlap in the representation and automatic analysis of learner language, which constitutes the topic of this chapter.

We can distinguish three main uses of NLP 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, function and usage. The use of NLP tools for annotation can be combined with human post-editing to eliminate potential problems introduced by the automatic analysis. NLP tools can also be integrated into a manual annotation set-up to flag annotation that appears to be inconsistent across comparable corpus instances (Boyd et al. 2008), automatically identify likely error locations and refine manual annotation (Rosen et al. 2014). The use of NLP for corpus annotation will be an important focus of this chapter. A detailed discussion of the spell- and grammar-checking techniques related to error annotation can be found in Chapter 25 (this volume).

Second, NLP tools are used to provide specific analyses of the learner language in the corpus. For instance, for the task of native language identification discussed in detail in Chapter 27 (this volume), classifiers are

trained to automatically determine the native language of the second/foreign language learner who wrote a given text. In another NLP task, learner texts are analysed to determine the proficiency level of the learner who wrote a given text (Pendar and Chapelle 2008; Yannakoudakis et al. 2011; Vajjala and Lõo 2013; Hancke and Meurers 2013), a task related to the analysis of developmental sequences and criterial features of different stages of proficiency (Granfeldt et al. 2005; Rahkonen and Håkansson 2008; Alexopoulou et al. 2010; Tono 2013; Murakami 2013a, 2013b) and the popular application domain of automatic essay grading addressed in Chapter 26 (this volume).

The third type of NLP application in the context of learner corpora is related to the previous two, but, unlike them, it is not designed to provide insights into the learner corpus as such. Instead, the learner corpus is used only to train NLP tools, specifically the statistical or machine learning components. The trained NLP tools can then be applied to learner language arising in other contexts. A tool trained on a learner corpus to detect particular types of learner errors can be used to provide immediate, individualised feedback to learners who complete exercises in an intelligent tutoring system. Such Computer-Assisted Language Learning (CALL) systems to which NLP has been added are commonly referred to as Intelligent Computer-Assisted Language Learning (ICALL) – cf. Chapter 22 (this volume). While traditionally the two fields of learner corpus research and ICALL developed independently and largely unconnected (but see Granger et al. 2007), the NLP analysis of learner language for corpus annotation is essentially an offline version of the online NLP analysis of learner language in ICALL, where the learner is waiting for feedback.

Complementing the NLP analysis of learner language, the other use of NLP in the language learning context (Meurers 2013) analyses the native language to be learned. Examples of the latter include the retrieval of pedagogically appropriate reading materials (e.g. Brown and Eskenazi 2005; Ott and Meurers 2010), the generation of exercises (e.g. Aldabe 2011) or the presentation of texts to learners with visual or other enhancements supporting language learning (Meurers et al. 2010). The standard NLP tools have been developed for native language and thus are directly applicable in that domain. On the other hand, the analysis of learner language raises additional challenges, which figure prominently in the next section on the core issues, from corpus representation as such to linguistic and error annotation.

2 Core issues

2.1 Representing learner data and the relevance of target hypotheses

At the fundamental level, representing a learner corpus amounts to encoding the language produced by the learner and its metadata, such

as information about the learner and the task performed (see Granger 2008b: 264). For the spoken language constituting the primary data for research on first language acquisition (e.g. *CHILDES*,[1] MacWhinney 2000), most work on uninstructed second language acquisition (e.g. *ESF*,[2] Perdue 1993), and a small part of the instructed second language acquisition corpora (e.g. *NICT JLE*,[3] Tono et al. 2004), this involves the question of how to encode and orthographically transcribe the sound recordings (Wichmann 2008: 195ff.). Written language, such as the learner essays typically collected in instructed contexts (e.g. *ICLE*,[4] Granger et al. 2009), also requires transcription for handwritten learner texts, though essays typed by learners are increasingly common. The language typed into CALL systems is starting to be systematically collected, supporting the creation of very large learner corpora such as *EFCAMDAT*[5] (Geertzen et al. 2013). Learner corpora can also be collected from websites such as *Lang-8* (Brooke and Hirst 2013), a website where non-native writers can receive feedback from native speakers.

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 in McEnery et al. (2006) and Lüdeling and Kytö (2008, 2009), there are important aspects of representation that are specific to learner language. Researchers in second language acquisition emphasise the individual, dynamic nature of interlanguage (Selinker 1972), and focus on characterising 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 trying to analyse learner data on that basis, one encounters forms and patterns which cannot be analysed in terms of the targeted native language system.

Consider, for example, the learner sentence in (1), taken from the *Non-native Corpus of English* (*NOCE*; Díaz-Negrillo 2007), consisting of essays 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 bolded word *choiced* can be identified as a verbal suffix and interpreted

[1] *Child Language Data Exchange System*.
[2] *European Science Foundation Second Language.*
[3] *National Institute of Information and Communications Technology – Japanese Learner English.*
[4] *International Corpus of Learner English*.
[5] *EF-Cambridge Open Language Database*.

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. In this example and a systematic set of cases discussed in Díaz-Negrillo et al. (2010), it is thus not possible to assign a unique English part of speech to learner tokens.

In examples such as (1), it seems straightforward to analyse 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 be another 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 of 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: 135ff.). The sentence includes only 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}	$\text{work}_{\text{1PL/3PL/INF}}$	in	Liechtenstein

Given that agreement phenomena always involve (at least) two elements, there is a systematic ambiguity in determining grammatical target forms. If we take the second-person subject *du* at face value, the corresponding second-person verb form *arbeitest* is the likely target. If we instead interpret the verb as it was written, we have to assume that it is a finite form and thus postulate the corresponding third-person plural *sie* ('they') or first-person *wir* ('we') as subject to obtain a well-formed target language sentence. Fitzpatrick and Seegmiller (2004) present a study confirming that it is often difficult to decide on a unique target form.

Considering the difficulty of uniquely determining target forms, one needs to document the reference on which any subsequent analysis is based in order to ensure valid, sustainable interpretations of learner language. Lüdeling (2008) thus argues for explicitly specifying such target hypotheses as a representation level of learner corpora. Rosen et al. (2014: 80–1) confirm that disagreement in the analysis of learner data often arises from different target hypotheses being assumed. We will discuss their findings for the *Czech as a Second Language* (*CzeSL*) corpus in the first case study in Section 3.1.

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, what constitutes a target hypothesis and how it is obtained needs clarification. There 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 produced by the learner bottom-up in terms of a linguistic reference system, such as the targeted native-language system codified in the standard corpus annotation schemes. One can then define a target hypothesis which encodes the minimal form change that is 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 (TH1) made explicit in the annotation manual of the German learner corpus *Falko* (Reznicek et al. 2012: 42ff.). An alternative incremental operationalisation of a purely form-based target hypothesis is spelled out in Boyd (2012). Both approaches explicitly define what counts as minimal form change. 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 instead they determine the minimal number of explicitly defined form changes that is needed to turn the learner sentence into a grammatical 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. 2013; Chapter 7, this volume).

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 expectations about human communication in general. The top-down, meaning-driven and the bottom-up, form-driven interpretation processes essentially interact in any interpretation of human language. For interpreting learner language, this interaction is particularly relevant given that the interlanguage forms used by language learners cannot be fully interpreted in terms of the established linguistic reference systems developed for the native language. This is particularly evident for learner language such as the Basic Variety that is characteristic of uninstructed second language acquisition (Klein and Perdue 1997), which lacks most grammatical form marking. The fact that learner language offers limited and hard-to-interpret form-driven information bottom-up 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. Given that form-based target hypotheses are defined to consist of the minimal form change that is needed to obtain a sentence that is grammatical in the target-language system, these target hypotheses are identical to the learner sentences in both cases. 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 a second, meaning-based target hypothesis for the two sentences. On that basis, we can analyse 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 (cf. Zyzik and Azevedo 2009).

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. It is also present in the annotation guidelines used for the learner essays and summaries collected in the *Falko* corpus. The Extended Target Hypothesis (TH2) operationalised in Reznicek et al. (2012: 51ff.) 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, as such it cannot be made explicit in the same formal way as the minimal form change target hypothesis TH1. To ensure sufficient inter-annotator agreement for TH2 annotation, task design arguably requires particular attention. The importance of integrating more task and learner information into the analysis of learner data is confirmed by the prominent evidence-centred design approach in language assessment (Mislevy et al. 2003).

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 conceptualisation 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. Indeed, strategic competence is one of the components of language competence distinguished in the seminal work of Canale and Swain (1980), and Bachman and Palmer (1996) explicitly discuss planning how to approach a test task as a good example of such strategic competence. In an instructed setting, second/foreign learners know that form errors are one of the aspects they typically are evaluated on, and therefore they may strategically produce language in a way that minimises the number of form errors they produce. For example, Ott et al. (2012: 59–60) found that the learners in the *Corpus of Reading Comprehension Exercises in German* (*CREG*) simply lift material

from texts or use familiar chunks, a strategy which, for example, 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 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 learners may access chunks and represent aspects at a propositional level, cf. Kintsch and Mangalath 2011). Even when learners produce apparently grammatical target-language forms, their conceptualisation of the language forms does not necessarily coincide with the analysis in terms of the target-language system. For example, Amaral and Meurers (2009) found that learners could not interpret feedback provided by an intelligent tutoring system because they misconceptualised contracted forms.

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. It is also this targeted native language for which the linguistic annotation schemes and NLP tools have been developed. 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 language analyses. 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 may only consist of one or a couple of words instead of full sentences, and more abstract target hypothesis representations may help avoid an overcommitment that would be entailed by specifying the full surface forms of sentences, e.g. where multiple word orders are possible in a given context.

2.2 Annotating learner data

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 of doing a linear search, by reading through the phone book from beginning to end until one finds the right person, would be possible in theory but would generally not be efficient enough in practice. 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 to the typical questions one tries to address using phone books written alphabetically. For Chinese names, on the other hand, the number of strokes in the name as

written in the logographic writing system is typically used instead, with the radical-and-stroke sorting used in Chinese dictionaries being another option.

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 by the first letter of the street name. Taking this running example one important step further, consider 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 resource to first determine the names of all the butchers. 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. Any such index is an interpretation of the data giving us direct access to specific subsets of data which are relevant in a particular perspective.

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 help 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 in the corpus.

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 Second Language Acquisition (SLA) research, 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, such as parts of speech or syntactic dependencies, are useful for querying the corpus for a wide range of research questions arising in FLTL and SLA – much like annotating telephone book entries with professions allows us to search for people to address 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 generally relevant classes of data. Which type of annotations one can and should provide for learner corpora using automatic or manual annotation or a combination of the two is an important research issue at the intersection of learner corpus and NLP research.

2.2.1 Linguistic annotation

A wide range of linguistic corpus annotation schemes have been developed for written and spoken language corpora (compare, e.g., Garside et al. 1997; Leech 2005; see also Chapters 5 and 6, this volume), 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 and Schäfer 2002, 2003b; Sagae et al. 2010). 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 originally developed for native language, especially edited news text, can seemingly be applied. At closer inspection, even this requires some leeway when checking whether the definitions in the annotation schemes apply to a given learner corpus example. In NLP, such leeway 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 a particular 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 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, many of the uses of learner corpora aim to advance our understanding of language acquisition by identifying characteristics of learner language. For such research, the particularities and variability of learner language at the level being investigated are exactly what we want to identify, not gloss over robustly. In Section 2.1 we already discussed a key component for addressing this issue: target hypotheses (specifically the form-based TH1). We can see those as a way of documenting the variation that robust analysis would simply have glossed over. Target hypotheses require the researcher to make explicit where a change is required to be able to analyse the learner language using a standard linguistic annotation scheme. A learner corpus including target hypotheses and linguistic annotation thus makes it possible to identify both the places where the learner language diverges from the native language norm and the general linguistic classes needed for retrieval of relevant subsets of learner data.

At the same time, such an approach cannot be the full solution to analysing 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 emerged. Such category systems are thus difficult to develop for the individual, dynamic interlanguage of language learners. But if we instead simply use a native-language annotation scheme to characterise learner language, we run 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: 6).

Given the insight from hermeneutics[6] that every interpretation is based on a given background, it is evident that we can never perceive anything as such.[7] However, it is possible to limit the degree of the comparative fallacy entailed by the annotation scheme used. The idea is to annotate learner language as closely as possible to the specific dimensions of observable empirical properties. For example, traditional parts of speech encode a bundle of syntactic, morphological, lexical and semantic characteristics of words. For learner language, Díaz-Negrillo et al. (2010) proposed instead to employ a tripartite representation with three separate parts of speech explicitly encoding the actually observable distributional, morphological and lexical stem information. Consider the examples in (4).

(4) a. The **intrepid** girl smiled.
b. He **ambulated** home.
c. The king **of** France is bald.

In terms of the distributional evidence for the part of speech of the word *intrepid* in sentence (4a), between a determiner and a noun we are most likely to find an adjective. Illustrating the morphological evidence, in (4b) the word ending in the suffix *-ed* is most likely to be a verb. Finally, in (4c) the word *of* is lexically unambiguous so that looking up the isolated word in a dictionary is sufficient for determining that it is a preposition. For native language, the three sources of empirical evidence converge and can be encoded by one part-of-speech tag. For learner language, these three information sources may diverge, as was illustrated by example (1). To avoid some of the complexity of such multidimensional tagsets, Reznicek and Zinsmeister (2013) show that the use of underspecified tags

[6] http://plato.stanford.edu/entries/hermeneutics (last accessed on 13 April 2015).

[7] An accessible introduction to hermeneutics and radical constructivism can be found in Winograd and Flores (1986). Humans under this perspective are autopoietic systems evolving in a constant hermeneutic circle.

(leaving out some information) and portmanteau tags (providing richer tagsets, combining information) can lead to an improved part-of-speech analysis of German learner language.

In the syntactic domain, encoding classes close to the empirical observations can be realised by breaking down constituency in terms of (a) the overall topology of a sentence, i.e. the sentence-level word order, (b) chunks and chunk-internal word order and (c) lexical dependencies. What is encoded in the overall topology of a sentence depends on the language and includes grammatical functions and discourse aspects, but the prevalence of notions such as fronting or extraposition in linguistic characterisations of data illustrates the relevance of such a global, topological characterisation of sentences. For some Germanic languages, this characterisation can build on the tradition of topological fields analysis, based on which automatic NLP analyses have also been developed for German (Cheung and Penn 2009). Topological fields are also starting to be employed in the analysis of learner language (Hirschmann et al. 2007).

Chunks are widely discussed in the context of learner language (though often in need of a precise operationalisation 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. This is, for example, the case in Meaning Text Theory (Mel'čuk 1988) and it is reflected in the distinction between the analytical and the tectogrammatical[8] layer of the Prague Dependency Treebank (Böhmová et al. 2003). Against this background, 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 and Ragheb 2009; Ragheb and Dickinson 2012), such as observable case marking or agreement properties. On the other, there are approaches which essentially target a level of semantic dependencies (Rosén and Smedt 2010; Ott and Ziai 2010). The goal here is to robustly abstract away from learner-specific forms and constructions where syntax and semantics diverge (such as English case-marking prepositions or the interpretation of the subject of non-finite constructions) to encode the underlying function–argument relations from which the sentential meaning can be derived. For example, dependency parsing is used as part of a content-assessment system analysing learner responses to reading comprehension questions (Hahn and Meurers 2012). King and Dickinson (2013) report on the NLP analysis of another task-based learner corpus supporting the evaluation of meaning. For learner data

[8] The tectogrammatical layer is the underlying structure at the heart of Prague School dependency analysis. In contrast to the surface-based analytical layer, the tectogrammatical layer focuses on those aspects which contribute to the semantic and pragmatic interpretation.

from a picture-description task, they obtain very high accuracies for the extraction of the core functor–argument relations using shallow semantic analysis. For any dependency analysis of learner data to be useful for research, the essential question is which kind of dependency distinctions can reliably be identified given the information in the corpus. This is starting to be addressed in recent work (Ragheb and Dickinson 2013).

Relatedly, when using parsers to automatically assign dependency analyses for learner language, one needs to be aware that the particular parsing set-up chosen impacts the nature and quality of the dependency analysis that is obtained. Comparing two different computational approaches to dependency parsing German learner language, for example, Krivanek and Meurers (2013) show that a rule-based approach was more reliable in identifying the main argument relations, whereas a data-driven parser was more reliable in identifying adjunct relations. This is also intuitively plausible, given that statistical approaches can use the world knowledge encoded in a corpus for disambiguation, whereas the grammar-based approach can rely on high-quality subcategorisation information for the arguments.

2.2.2 Error annotation

A second type of annotation of learner corpora, error annotation, targets the nature of the difference between learner data and native language (see Chapter 7, this volume). Given the FLTL interest in identifying, diagnosing and providing feedback on learner errors, and the fact that learner corpora are commonly collected in an FLTL context, error annotation is the most common form of annotation in the context of learner corpora (Granger 2003b; Díaz-Negrillo and Fernández-Domínguez 2006). At the same time, error annotation is only starting to be subjected to the rigorous systematisation and inter-annotator 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 analyses becoming available in the NLP context confirm that the issue indeed requires scrutiny. Rozovskaya and Roth (2010a) find very low inter-annotator agreement for error classification of English as a Second Language sentences. Even for the highly focused task of annotating preposition errors, Tetreault and Chodorow (2008a) report that trained annotators failed to reach good agreement. Rosen et al. (2014) provide detailed inter-annotator agreement analyses for the *Czech as a Second Language* corpus, making concrete for which aspects of error annotation good agreement can be obtained and what this requires – a study we discuss in detail in Section 3.1. For corpus annotation to support reliable, replicable access to systematic classes of data in the way explained at the beginning of Section 2.2, it is essential to reduce annotation schemes to those categories that can reliably be assigned based on the evidence available in the corpus.

2.2.3 Automatic detection and diagnosis of learner errors

In terms of tools for detecting errors, learner corpus research has long envisaged automatic approaches (e.g. Granger and Meunier 1994), but the small community at the intersection of NLP and learner corpus research is only starting to make headway. The mentioned conceptual challenges and the unavailability of gold-standard error-annotated learner corpora hinder progress in this area. Corpora with gold-standard annotation are essential for developing and evaluating current NLP technology, which is generally built using statistical or supervised machine learning components that need to be trained on large, representative gold-standard data sets. Writing aids for native speakers, such as the standard spell- and grammar-checkers, 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 (Flor et al. in press). For example, Rimrott and Heift (2008: 73) report 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'. Examples of such multiple-edit errors include lexical competence errors such as German *Postkeutzah* → *Postleitzahl* ('postal code') or grammatical overgeneralisations as in *gegehen* → *gegangen* ('went').

The overall landscape of computational approaches for detecting and diagnosing learner errors can be systematised 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 identify specific types of errors. Language-licensing approaches attempt to analyse an entire learner utterance to diagnose its characteristics. In the following, a conceptual overview is provided, with Chapter 25 (this volume) spelling the topic out further.

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 (5) from the *Chinese Learner English Corpus* (*CLEC*).[9]

(5) **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 in (6).

(6) At class, students listen **more** careful$_{\textbf{adj}}$ **then** any other time.

[9] http://purl.org/icall/clec (last accessed on 13 April 2015).

Error pattern matching is commonly implemented in standard grammar-checkers. For example, the open source *LanguageTool*[10] provides a general implementation, in which typical learner error patterns can be specified.

While directly specifying such error patterns 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 set-up 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 can then be used to determine whether a definite, an indefinite or no determiner is required for this chunk. This general approach is a very common set-up for NLP research targeting learner language (e.g. De Felice 2008; Tetreault and Chodorow 2008b; 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 interesting connection between these questions and the need to further advance error annotation schemes based on detailed analyses of inter-annotator agreement.

Language-licensing approaches go beyond characterising local patterns and attempt to analyse 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 set-up, a grammar is a set of rules. A string is recognised 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 set-up, a grammar consists of a set of constraints. A string is grammatical if and only if 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

[10] http://languagetool.org (last accessed on 13 April 2015).

ability to generalise across categories and rules to the modern lexicalised 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). To use any of these approaches in our context, we need to consider that linguistic theories and grammars are generally designed to license well-formed native language, which raises the question of how they can license learner language (and identify errors as part of the process).

There are essentially two types of approaches for licensing learner language, corresponding to the two types of formal grammars introduced above. In a validity-based set-up using a regular parser, so-called *mal*-rules can be added to the grammar (see, e.g., Schwind 1990; Matthews 1992) to license and thereby identify ill-formed strings occurring in the learner language. For example, Schwind (1990: 575) defines a phrase-structure rule licensing German noun phrases in which the adjective follows the noun. This is ungrammatical in German, so the rule is marked as licensing an erroneous structure, i.e. it is a *mal*-rule.

A *mal*-rule approach requires each possible type of learner error to be pre-envisaged and explicitly encoded in every rule in which it may surface. For small grammar fragments, as needed for exercises effectively constraining what the learner is likely to produce (Amaral and Meurers 2011: 9ff.), this may be feasible; but for a grammar with any broader coverage of language phenomena and learner error types, writing the *mal*-rules needed would be very labour intensive and error-prone.

Some types of errors can arise in a large number of rules; for example, subject–verb agreement errors may need to be accommodated in any rule realising subjects together with a finite verbal projection. Following Weischedel and Sondheimer (1983), meta-rules can be used to express such generalisations over rules. For example, they define a meta-rule that allows subject–verb agreement to be relaxed anywhere, and one allowing articles to be omitted in different types of noun phrases.

Rule-based grammars license the infinite set of possible strings by modularising the analysis into local trees. A local tree is a tree of depth one, i.e. a 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 licenses a local tree, which in combination with other local trees licensed by other rules ultimately licenses the entire sentence. The *mal*-rule approach is conceptually simple when the nature of an error can be captured within the local domain of a single rule. For example, a *mal*-rule licensing the combination of an article and a noun disagreeing in gender (le_{MASC} $table_{\text{FEM}}$) can be added to a French grammar. 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 orders licensed by a rule *S* → *NP VP* by adding the *mal*-rule *S* → *VP NP* makes it possible to license (7a) and (7b).

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

The order in (7c), on the other hand, cannot be licensed in this way given that it involves reordering words licensed by two different rules, so that no single *mal*-rule can do the job – unless one writes an ad hoc, combined *mal*-rule for the flattened tree (*S* → *V NP NP*), which would require adding such rules for combinations with all other rules licensing VPs (intransitive, ditransitive, etc.) as well. Lexicalised 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 (cf. Heift and Schulze 2007). 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 in those extended domains.

To limit the search space explosion commonly resulting from rule interaction, the use of *mal*-rules may be limited. One option is to only include the *mal*-rules in processing when parsing a sentence with the regular grammar fails. However, this only reduces the search space for well-formed strings. If parsing fails, the question of which *mal*-rules need to be added is not addressed. An intelligent solution to this question was pursued by the *ICICLE* system (*Interactive Computer Identification and Correction of Language Errors*; Michaud and McCoy 2004). It selects groups of rules based on learner modelling. For grammar constructs that the learner has shown mastery of, it uses the native language 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. In probabilistic grammar formalisms such as probabilistic context-free grammars (PCFGs) or when using LFG grammars with optimality theoretic mark-up, one can also try to tune the licensing of grammatical and ungrammatical structures to the learner language characteristics (cf. Wagner and Foster 2009).

The second approach to licensing sentences which go beyond the native-language grammars is based on constraint relaxation (Kwasny and Sondheimer 1981). It relies on a satisfiability-based grammar set-up or a rule-based grammar formalism employing complex categories (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. One can also integrate a mechanism corresponding to the meta-rules of the *mal*-rule set-up, in which the specifications of particular features are relaxed for particular sets of rules or constraints, or everywhere in the grammar. In this context, one often finds the claim that constraint relaxation does not require learner errors to be pre-envisaged and therefore should be preferred over a *mal*-rule approach. Closer inspection makes it clear that such a broad claim is incorrect: to effectively parse a sentence with a potentially recursive structure, it is essential to distinguish those constraints which may be relaxed from those that are supposed to be hard, i.e. always enforced. Otherwise either parsing does not terminate, or any learner sentence can be licensed with any structure so that nothing is gained by parsing.

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). The as yet unsolved question raised by such approaches (and the probabilistic grammar formalisms mentioned above) is how the weights can be obtained in a way that makes it possible to identify the likely error causes underlying a given learner sentence.

The constraint-relaxation research on learner error diagnosis has generally developed handcrafted formalisms and solutions. At the same time, computer science has studied Constraint Satisfaction Problems (CSP) in general and developed general CSP solvers. In essence, licensing a learner utterance is dealt with in the same way as solving a Sudoku puzzle or finding a solution to a complex scheduling problem. 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 CSP solvers, with diagnosis of learner errors being handled by general conflict-detection approaches.

Finally, while most approaches to licensing learner language use standard parsing algorithms with extended or modified grammars to license ill-formed sentences, there is also some work modifying the algorithmic

side instead. For instance, 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 in essence moves generalisations 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 under-researched area. One trend in this domain is to avoid the problem of gold-standard error annotation as reference for testing by artificially introducing errors into native corpora (e.g. Foster 2005). While this may be a good choice to monitor progress during development, such artificially created test sets naturally only reflect the properties of learner data in a very limited sense and do not eliminate the need ultimately to evaluate an approach 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) and Chapter 25 (this volume).

3 Representative 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 information is available. The second case study provides a concrete example for error detection in the domain of word-order errors, a frequent but under-researched type of error that also allows us to exemplify how the nature of the phenomenon determines the choice of NLP analysis used for error detection.

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

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 a Second Language* (*CzeSL*) corpus based on Rosen et al. (2014). 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 370,000 words from the written portion has been manually annotated.

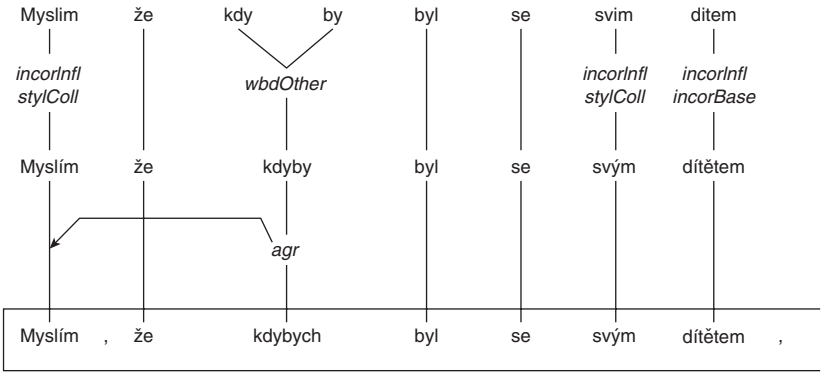

Figure 24.1 An example for the multi-tier representation of the *CzeSL* corpus (Rosen et al. 2014: 72)

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. Figure 24.1 exemplifies the *CzeSL* multi-tier representation.

Tier 0 at the top of Figure 24.1 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, shown at the bottom of the figure, further corrects an agreement error to obtain the target hypothesis, of which a glossed version is shown in (8).

(8)	Myslím,	že	kdybych	byl	se	svým	dítětem,
	think$_{SG1}$	that	if$_{SG1}$	was$_{MASC}$	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 word boundaries, inflectional and derivational morphology, incorrect word stems and invented or foreign words. The tier is thus closely related to the minimal form change target hypothesis we discussed in Section 2.1, 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 the annotation tool *feat*[11] developed for this purpose.

The annotation started with a pilot annotation of sixty-seven texts totalling almost 10,000 tokens. Fourteen 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 (κ, cf. Artstein and Poesio 2008). Since tiers 1 and 2 can differ between annotators, for computing the IAA, error tags are projected onto tier 0 tokens. The feedback from the pilot annotation 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 thirty-one annotators who analysed 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. (2014: 76), one of which is shown in Table 24.1.

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: $\kappa > 0.6$) and syntactic dependency errors (dep: κ 0.58). Such errors can thus be reliably annotated given the 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 ($\kappa < 0.4$).

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 with 0.24 when their target hypotheses differed. The study thus provides clear empirical support for the argument made in Section 2.1 that explicit target hypotheses are essential for reliable corpus annotation.

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 systematically assign 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. The authors thus obtain morphosyntactic categories and lemmas for the tier 2 target hypotheses by applying the standard Czech taggers and lemmatisers. These annotations can then be projected back onto tier 1 and the original learner forms.

[11] http://purl.org/net/feat (last accessed on 13 April 2015).

Table 24.1. Inter-annotator agreement for selected *CzeSL* error tags

Tag	Type of error	Pilot sample		All annotated texts	
		–	Avg. tags	–	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[a]	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

[a] The *styleColl* tag encodes errors specific to the diglossic language situation of Czech.
(Rosen et al. 2014: 76)

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, for example, tag it as an incorrect devoicing, a missing palatalisation or an error in capitalisation.

Third, automatic analysis is used to check the manual annotation. For example, the coverage of the automatic Czech morphological analyser used is very high. Learner forms at tier 0 that cannot be analysed by the morphological analyser are thus very likely to be ill-formed. By flagging all such forms for which no corrections have been specified at tier 1, one can systematically alert the annotators to potentially missing specifications.

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, the authors provide a comparison of the autocorrect mode of *Korektor* with the two target hypotheses spelled out in the *CzeSL* corpus. The comparison is based on a subset of almost 10,000

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 aligns neither with the purely lexical tier 1 of *CzeSL*, nor with 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. *Morče* (Votrubec 2006) prioritises morphological diagnostics and information from a lexicon over distributional context, whereas the trigram tagger *TnT* (Brants 2000) essentially uses the opposite strategy, extracting the lexicon from the training data. The authors show that the different strategies indeed lead to significantly different results. The two tagging approaches agreed on the same tag for only 28.8 per cent of the ill-formed tokens in a corpus sample of 12,681 tokens. Considered in the context of Section 2.2.1 arguing for the need to linguistically annotate learner language close to the observable evidence, this evaluation of standard part-of-speech tagging tools underlines the relevance of a tripartite encoding of parts of speech for learner language, distinguishing morphological, distributional and lexical stem information.

3.2 Metcalf, V. and Meurers, D. 2006. *When to Use Deep Processing and When not to – the Example of Word Order Errors*. Presentation at the *CALICO Workshop NLP in CALL: Computational and Linguistic Challenges, May 17*. University of Hawaii.

The second case study is intended to illustrate the issues involved in answering the question of which NLP techniques are needed for which kind of learner language analysis. The original goal of Metcalf and Meurers (2006) was to automatically identify word-order errors in different types of exercises in an ICALL context.

Language learners are known to produce a range of word-order errors. Such errors are frequent and word order differs significantly across languages so that transfer errors play a role in this context (see, e.g., Odlin 1989). It is important for learners to master word order, in particular because errors in this domain can significantly complicate comprehension. This is exemplified by (9a) from the *Hiroshima English Learners' Corpus* (*HELC*, Miura 1998), which is virtually incomprehensible, whereas the rearranged word order (9b) is already quite close to the target (9c) of this translation activity.

(9) 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. For some, the analysis can rely on lexical triggers (one of a finite set of words is known to occur) or indicative patterns such as characteristic sequences of parts of speech, whereas for others, a deeper linguistic analysis is required. Correspondingly, there are essentially 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 (10), but they must follow a pronominal object (11). For inseparable phrasal verbs (also called prepositional verbs), particles always precede the object (12).

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

Authentic learner sentences instantiating these error patterns are shown in (13), taken from the *Chinese Learner English Corpus* (*CLEC*).

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

Complementing such erroneous realisations, Metcalf and Meurers also found patterns of avoidance in the *CLEC*, such as heavy use of a pattern that is always grammatical (*verb* < *particle* < *NP*, with the operator < expressing linear precedence), but little use of patterns restricted to certain verb and object types (e.g. *verb* < *pronoun* < *particle*). Related avoidance patterns are discussed in Liao and Fukuya (2002).

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 specific subclasses of particle verbs. As a result, error patterns such as the ones in (14) and (15) (and the alternative avoidance patterns mentioned above) can be identified automatically.

(14)		* wrote down it	separable-phrasal-verb < particle < pronoun
(15)	a.	* ran my neighbour into	inseparable-phrasal-verb < NP/pronoun < particle
	b.	* ran her into	inseparable-phrasal-verb < NP/pronoun < particle

Precedence here can be specified using a regular expression allowing any number of words in between. Only matches within the same sentence are relevant here, so that a sentence-segmented corpus (Palmer 2000) is required. In terms of the general landscape of linguistic patterns that can be searched for in a corpus (Meurers 2005), we are here dealing with regular expression patterns over words and part-of-speech tags within basic domains.

The strength of such a shallow pattern-matching approach is its simplicity and efficiency. The weakness is its lack of generalisation over tokens and patterns: the words or parts of speech for which order is to be checked must be known, and all relevant word orders must be pre-envisaged 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 analysing 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 various 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*).[12]

(16) a. There have been **already** several campaigns held by 'Outdoor'.
b. while any covert action brings **rarely** such negative connotations.
c. It seems that the Earth has **still** a lot to reveal.

[12] http://purl.org/net/PICLE (last accessed on 13 April 2015).

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 characterisation 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 a language-licensing NLP approach (see Section 2.2.3), *mal*-rules can be added to a grammar so that a parser can license the additional word orders. A downside of such an approach 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-occur 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 lexicalised, constraint-based formalism such as HPSG, the position of an adverb can instead be constrained and recorded using a lexical principle governing the realisation of the entire head domain instead of in a local tree. Similarly, a dedicated level recording the topological sentence structure may be used to modularise word order.

Summing up this second case study on the detection of word-order errors, instance-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, lexicalisation or separate topological representations can be an attractive, modular alternative to phrase-structure-based encodings.

4 Critical assessment and future directions

Wrapping up the chapter with a critical analysis of the role and future directions of NLP in the context of learner corpus research, this section starts by identifying the clear opportunity NLP provides for connecting learner corpus and SLA research, before turning to trends and emerging applications relating to the intersection of NLP and learner language.

4.1 NLP supporting the use of learner corpora for SLA research

Learner corpus research has been 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 characterisation of learner language as a systematic interlanguage to be characterised in its own right as the established basis of SLA research over the past decades. With the corpus collection and representation methods largely established, learner corpus research in the future is well placed to connect to and make important contributions to the SLA mainstream (see also Chapter 14, this volume). In support of this perspective, NLP methods analysing and annotating learner data are essential for efficiently identifying the subsets of data of relevance to specific SLA research questions.

4.1.1 Facilitating the analysis of learner language in a broad range of tasks

To be able to analyse a broader, representative range of authentic learner language use, we can expect that learner corpora in the future will be designed to record learner language in a broader range of tasks. This seems crucially needed to obtain learner corpora with sufficient representation of the language constructs at issue in current SLA research. Considering where such task-based learner data can be obtained, CALL systems will play an increasingly important role – especially where the integration of NLP supports a broader range of tasks in ICALL systems, involving both form and meaning. This includes dialogue systems (Petersen 2010; Wilske 2014), where the collection of learner data becomes directly relevant to SLA research on the effectiveness of interventions and feedback.

To facilitate the interpretation and annotation of learner language across tasks, it is necessary 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 detailed analysis framework incorporating insights from task-based learning and CALL research. He uses this framework to analyse and specify the capabilities of the NLP tools that are needed to process the learner language arising in those tasks. He then applies it in the design of a CALL authoring system – software that should allow a teacher to concisely delineate the different sets of learner language expressions for which feedback is provided in the task being authored.

The relevance of analysing learner language produced in a range of explicitly controlled tasks strengthens the link between learner corpora and ICALL. In essence, the connection is threefold: (i) ICALL systems and the automatic annotation of learner corpora both need NLP for the analysis of learner language, (ii) CALL environments facilitate the collection of large, task-based learner corpora, and (iii) ICALL systems support the experimental testing of SLA hypotheses.

While more research is needed, results obtained in such ICALL experiments have been shown to generalise. Consider Petersen (2010), who studies the effectiveness of recasts providing implicit negative feedback

in the domain of English question formation. He shows that recasts in the human–computer setting using the dialogue system are as effective as the same kind of feedback delivered in dyadic, human–human interaction. Presson et al. (2013) coin the term experimental CALL (eCALL), highlighting the new opportunities arising from linking CALL environments and SLA research. The availability of very large learner corpora collected as part of CALL environments, such as *EFCAMDAT* (Geertzen et al. 2013), already makes 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 (2013a, 2013b), who raises serious doubts about the conclusion of the so-called morpheme studies, a hallmark of SLA research claiming that the order of acquisition is essentially independent of the native language of the learner.

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 – an area where NLP techniques are increasingly successful (Dzikovska et al. 2013). Tasks grounded in meaning are also attractive in making it possible to analyse the choices learners make in the linguistic system to realise the same meaning, i.e. to study the use of forms under a variationist perspective (Tagliamonte 2011), connecting linguistic and extra-linguistic variables. Corpora integrating a range of tasks may even require such analysis in order to make it possible to distinguish between those language aspects determined by the task (see Chapter 18, this volume) and those which are general characteristics of the learner language and language development.

4.2 Deeper linguistic modelling in NLP and learner language analysis

In terms of the nature of the NLP resources and algorithms employed for the analysis of learner language, the choice between surface-based and deeper linguistic analysis arguably remains the most important modelling question. For example, as argued in Meurers et al. (2014), a task such as native language identification (see also Chapter 27, this volume) provides an exciting experimental sandbox for quantitatively evaluating where a surface-based NLP analysis is successful and sufficiently general to apply 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 of which aspects of linguistic modelling 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 traditionally relying on surface-based comparisons (e.g. Latent Semantic Analysis relying solely on which words occur in a text) with essays for the same prompt for which gold-standard labels were manually assigned. The recent work on proficiency classification (Pendar and Chapelle 2008; Yannakoudakis et al. 2011; Vajjala and Lõo 2013; Hancke and Meurers 2013), on the other hand, studies lexical, syntactic, semantic and discourse aspects of the linguistic system, with the goal of learning something general about the complexity of language. The trend towards deeper linguistic modelling is particularly strong where only little language material is available. When analysing short answers to reading comprehension questions consisting of only a couple of sentences, which words occur in the answer by itself is not sufficiently informative. For the c-rater system (Leacock and 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 quantity of language in short answers. This amounts to employing a wide range of linguistic analyses for automatic meaning assessment, as exemplified by the range of approaches to the recent shared task analysing student responses (Dzikovska et al. 2013).

While most of the initial NLP work on learner corpus annotation and error detection focused on English, we are starting to see some NLP-based work on learner corpora for other languages, such as Korean (Israel et al. 2013), German (Ott and Ziai 2010; Boyd 2012; Hirschmann et al. 2013), Estonian (Vajjala and Lõo, 2013) and Czech (Rosen et al. 2014). The increased NLP interest in under-resourced and endangered languages is likely to support further growth. Corpora with a broad, representative coverage of target (and native) languages are essential for supporting general claims about SLA.

As a final point again emphasising the relevance of interdisciplinary collaboration in this domain, there is currently a surprising lack of interaction between NLP research related to first language acquisition and that on second language acquisition. Many of the representation, modelling 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 and Mitchell 2004), but there is still significantly more opportunity for synergy in future work.

Key readings

Jurafsky, D. and Martin, J. H. 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 introduction to NLP. It covers the key issues, representations and algorithms from both the theory-driven perspectives and the data-driven, statistical perspectives. It also includes a discussion of current NLP applications.

Dickinson, M., Brew, C. and Meurers, D. 2013. *Language and Computers*. Chichester: Wiley-Blackwell.

This is an introduction for the reader interested in exploring the issues in computational linguistics starting from the real-life applications, from search engines via dialogue systems to machine translation. It includes a chapter on language tutoring systems which emphasises the motivation of the linguistic modelling involved in learner language analysis and the corresponding NLP techniques.

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

The authors provide a very comprehensive historical perspective of the field of ICALL situated at the intersection of CALL and NLP. NLP analysis specifically targeting learner language was for the most part developed in this context; an understanding of the CALL needs and how NLP has tried to address them is thus of direct relevance to any NLP analysis targeting learner language. Readers interested in zooming out to NLP in the context of language learning in general can find this discussion in Meurers (2013).

Leacock, C., Chodorow, M., Gamon, M. and Tetreault, J. 2014. *Automated Grammatical Error Detection for Language Learners*, 2nd edn. Synthesis Lectures on Human Language Technologies. San Rafael, CA: Morgan and Claypool.

Starting with a characterisation 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.

Díaz-Negrillo, A., Ballier, N. and Thompson, P. (eds.) 2013. *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: Benjamins.

The book provides a broad collection of current perspectives on the automatic analysis of learner corpora. It discusses aspects of corpus representation and conceptual and methodological issues, and presents concrete studies based on a range of different written and spoken language data.

Meurers, D. 2005. 'On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German', *Lingua* 115(11): 1619–39.

This article seeks to answer the question of what exactly is involved in using corpora to address linguistic research issues. It 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 and Müller (2009), where such treebanks are the main source of data used.

Artstein, R. and Poesio, M. 2008. 'Inter-coder agreement for computational linguistics', *Computational Linguistics* 34(4): 555–96.

Corpus annotation is only useful for scientific work if it provides a reliable, replicable index into the data. Investigating the inter-coder agreement between multiple independent annotators is the most important method for determining whether the classes operationalised 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 inter-coder agreement, which in the corpus context typically is referred to as inter-annotator agreement.