Noname manuscript No.
(will be inserted by the editor)

Evaluating and automating the annotation of a learner corpus

Alexandr Rosen · Jirka Hana · Barbora Štindlová · Svatava Škodová · Anna Feldman

Received: date / Accepted: date

Abstract The paper describes a corpus of texts produced by non-native speakers of Czech. We discuss its annotation scheme, consisting of three interlinked levels to cope with a wide range of error types present in the input. Each level corrects different types of errors; links between the levels allow capturing errors in word order and complex discontinuous expressions. Errors are not only corrected, but also classified.The annotation scheme is tested on a doubly-annotated sample of approx. 10,000 words with fair inter-annotator agreement results. We also explore the possibility of application of automated linguistic annotation tools (taggers, spell checkers and grammar checkers) on the learner text to support or even substitute manual annotation.

Keywords learner corpus · error annotation · second language acquisition · Czech

The corpus is one of the tasks of the project Innovation of Education in the Field of Czech as a Second Language (project no. CZ.1.07/2.2.00/07.0259), a part of the operational programme Education for Competiveness, funded by the European Structural Funds (ESF) and the Czech government. The annotation tool was also partially funded by grant no. P406/10/P328 of the Grant Agency of the Czech Republic.

Alexandr Rosen
Charles University, Prague, Czech Rep.
E-mail: alexandr.rosen@ff.cuni.cz

Jirka Hana
Charles University, Prague, Czech Rep.
E-mail: jirka.hana@gmail.com

Barbora Štindlová
Technical University, Liberec, Czech Rep.
E-mail: barbora.stindlova@tul.cz

Svatava Škodová
Technical University, Liberec, Czech Rep.
E-mail: svatava.skodova@tul.cz

Anna Feldman
Montclair State University, Montclair, NJ, USA
E-mail: feldmana@mail.montclair.edu

1 Introduction

Learner corpora, i.e., databases of texts produced by non-native speakers, are a rich source of information about specific features of learners' language. They can be annotated with morphosyntactic categories or syntactic structure, but their most interesting aspect are examples of deviant use, which can be identified, corrected and classified. Annotation of this kind is a challenging task, even more so for a language such as Czech, with its rich inflection, derivation, agreement, and a largely information-structure-driven constituent order.

The present work describes a learner corpus of Czech, compiled from texts written by students of Czech as a second or foreign language. We discuss its annotation scheme, consisting of three interlinked levels to cope with a wide range of error types present in the input. The annotation scheme is tested on a doubly-annotated sample of approx. 10,000 words with fair inter-annotator agreement results. Each text is annotated twice, and several inter-annotator agreement measures are calculated.

Manual error annotation is very resource-intensive, which prompted us to explore means of obtaining error annotation and correction/emendation (semi-)automatically. For automatic correction we use a context-aware spell checker (Richter, 2010), trained on general texts produced by native speakers.

2 Learner corpora

A learner corpus, also called interlanguage or L2 corpus, is a computerised textual database of language as produced by foreign/second language (L2) learners (Leech, 1998, p. xiv). It is a very useful resource in the research of second language acquisition (SLA) and foreign language teaching (FLT). It serves as a repository of authentic data about a specific variety of natural language (Granger, 2003), namely the learner language, or interlanguage (IL).[1]

Learner corpora can be used to compare non-native and native speakers' language, or interlanguage varieties. They can be studied on the background of traditional native language corpora, which helps to track various deviations from standard usage in the language of non-native speakers, such as frequency patterns – cases of overuse or underuse or "foreign-soundingness," in comparison with the language of native speakers. Recent studies have focused primarily on the frequency of use of separate language elements (Ringbom, 1998), collocations and prefabs (Nesselhauf, 2005), lexical analysis and phrasal use (Altenberg and Tapper, 1998), etc.

An error-tagged corpus can be subjected to computer-aided error analysis (CEA), which is not restricted to errors seen as a deficiency, but understood as a means to explore the target language and to test hypotheses about the functioning of L2 grammar. CEA also helps to observe meaningful use of non-standard structures of IL. Recent studies focus on lexical errors (Leńko-Szymańska, 2004), wrong use of verbal tenses (Granger, 1999) or phrasal verbs (Waibel, 2008).

[1] Interlanguage is of highly individual and dynamic nature. It is subject to constant changes as the learner progresses through successive stages of acquiring more competence, and can be seen as an individual and dynamic continuum between one's native and target languages. See Selinker (1972).

Learner corpora can differ in many ways (for more details see, e.g., Granger, 2008, p. 260):

- Medium: Learner corpora can capture written or spoken texts, the latter much harder to compile, thus less common.
- First language (L1): The data can come from learners with the same L1 or with various L1s.
- Target language (L2): Most learner corpora cover the language of learners of English as a second or foreign language (ESL or EFL). The number of learner corpora for other languages is smaller but increasing.
- Proficiency in target language: Some corpora gather texts of students at the same level, other include texts of speakers at various levels. Most corpora focus on advanced students.
- Cross-sectional/developmental data: Most L2 corpora are cross-sectional, gathering data from various types of learners. Only few L2 corpora are developmental (longitudinal), including data acquired over time from the same learners. Several learner corpora collect balanced data from homogeneous groups of learners at different levels of L2 knowledge and are used in SLA research as quasi-longitudinal learner corpora.
- Annotation: Many learner corpora contain only raw data, some contain emendations (i.e. corrections), but only few use error tags to classify errors. Some corpora use linguistic annotation, the most common is part of-speech (POS) tagging.

Table 1 presents a representative overview of currently available learner corpora. For more details see, e.g., Pravec (2002), Nesselhauf (2005), Štindlová (2011) and Xiao (2008).

3 A learner corpus of Czech

The learner corpus of Czech as a Second Language (CzeSL) is built as a part of a larger project, the Acquisition Corpora of Czech (AKCES), a research programme pursued at Charles University in Prague since 2005 (Šebesta, 2010). AKCES includes:

- CzeSL – 1 mil. words (to be finished in 2012)
- SCHOLA 2010 and EDUCO – recordings and transcripts of classes from Czech primary schools (about 800,000 words each, finished)
- SKRIPT – written texts of Czech students (about 600,000 words so far, in development),
- ROMi – texts and speech produced by young learners with Romani background[2] (in development)

[2] It might be difficult to decide what L1 of the Czech Roma minority is, yet the students often exhibit many traits typical for the process of acquisition of Czech as a second language. Bedřichová et al (2011) assume that the social, cultural and linguistic differences between the non-Roma majority and some Roma communities may imply specific language development of Roma children.

Table 1 Some currently available learner corpora

Size (th. of words)	L1	TL	TL proficiency	Medium	Error annotation
ICLE *– International Corpus of Learner English*					
3,000	26	English	advanced	written	yes (part)
CLC *– Cambridge Learner Corpus*					
35,000	130	English	all levels	written	yes (part)
LINDSEI *– Louvain International Database of Spoken English*					
800	11	English	advanced	spoken	yes (part)
USE *– Uppsala Student English Corpus*					
1,200	Swedish	English	advanced	written	no
CYLIL *– Corpus of Young Learner Interlanguage*					
500	4	English	all levels	spoken	no
HKUST *– Hong Kong Univ. of Science and Technology Corpus of Learner English*					
25,000	Chinese	English	advanced	written	yes (part)
CHUNGDAHM *– Chungdahm English Learner Corpus*					
131,000	Korean	English	all levels	written	yes (part)
JEFLL *– Japanese EFL Learner Corpus*					
700	Japanese	English	beginners	written	yes (part)
MELD *– Montclair Electronic Language Learners' Database*					
1,000	16	English	advanced	written	no
MICASE *– Michigan Corpus of Academic Spoken English*					
1,800	various	English	advanced	spoken	no
NICT JLE *– NICT Japanese Learner English*					
2,000	Japanese	English	all levels	spoken	yes (part)
FALKO *– Fehlerannotiertes Lernerkorpus*					
300	5	German	advanced	written	yes
FRIDA *– French Interlanguage Database*					
200	various	French	intermediate	spoken	yes (part)
FLLOC *– French Learner Language Oral Corpora*					
2,000	English	French	all levels	spoken	no
PiKUST *– Poskusni korpus usvajanja slovenščine kot tujega jezika*					
40	18	Slovene	advanced	written	yes
ASU *– ASU Corpus*					
500	various	Norwegian	advanced	written	no
CEDEL 2 *– Corpus Escrito del Español como L2*					
75	various	Spanish	all levels	written	yes (part)

– IUVENT – spoken corpus of native young Czechs' language (planned)

All the corpora are collected and built under similar conditions, which allows for a wide range of linguistic comparisons.

CzeSL is focused on three main groups of non-native speakers of Czech: (1) speakers of Slavic languages, (2) speakers of other Indo-European languages, (3) speakers of distant non-Indo-European languages.

The data collected for CzeSL include:

1. Written texts, produced during all range of situations throughout the language-learning process, collected as manuscripts and transcribed into an electronic format. The transcription follows rules designed to preserve many features of handwritten texts (such as self-corrections or emoticons Štindlová, 2011, p. 106).
2. Spoken data
3. Bachelors' and Masters' theses, written in Czech by non-native students

The data cover all language levels according to the Common European Framework of Reference for Languages (CEFR), from real beginners (A1 level) to advanced learners (level B2 and higher), with a balanced mix of levels as much as possible. This spectrum of various levels and genres is unique in the context of other learner corpora.

Each text is equipped with metadata records, some of them relate to the respondent (including sociological data about the learner, such as age, gender, and language background – the first language, proficiency level in Czech, knowledge of other languages, duration and conditions of language acquisition), while other specify the character of the text and circumstances of its production (availability of reference tools, type of elicitation, temporal and size restrictions etc.).

The intended use of the Czech learner corpus is mainly pedagogical. It will be used in the education of teachers of Czech as a foreign language, it will serve as a source of examples for particular phenomena or of complete authentic texts that can be used both in the classroom and in the production of educational tools, and will help to tailor instructions and teaching materials to specific groups of learners (e.g., groups with different native languages or groups of different ages). Moreover, we expect CzeSL to become a resource for an extensive research of Czech as a second language and the second language acquisition in general (Štindlová, 2011).

4 Annotation of learner corpora

In the context of second/foreign language acquisition, the learners' language is seen as an independent system, which should be analyzed in its entirety, with incorrect structures as an important part. Texts produced by non-native speakers can be annotated in two different ways:

- Linguistic mark-up (e.g., part-of-speech tagging, morphological or syntactic annotation, lemmatization etc.). In most learner corpora, at least some parts are POS-tagged by tools and tagsets originally developed for the analysis of the national language, cf., e.g., Van Rooy and Schäfer (2003). However, it is often far from obvious what kind of annotation an incorrect expression should receive.
- Error annotation, cf., e.g., Díaz-Negrillo and Fernández-Domínguez (2006). There are two different kinds of error annotation:
  - emendation: correction of erroneous text – establishing one or more *target hypotheses* about the author's intention and its expression
  - error categorization: annotation of errors with tags from a predefined error taxonomy

  Investigating learners' language is easier when deviant forms are annotated at least by their correct counterparts, or, even better, by tags making the nature of the

error explicit.[3] Although learner corpora tagged this way exist, the two decades of research in this field have shown that designing a tagset for the annotation of errors is a task highly sensitive to the intended use of the corpus and the results are not easily transferable from one language to another.

5 Error annotation of CzeSL

5.1 Annotation scheme as a compromise

Building an error-annotated learner corpus of Czech is a unique enterprise. In comparison with Czech, languages of the existing annotated learner corpora have simpler morphology and/or a more fixed word order. Therefore, many of the problems we have encountered have not been addressed before.[4] Moreover, although the annotation scheme should be sufficiently informative and extensible, it should also be manageable and easily applicable, i.e. not too extensive. The resulting scheme and error typology is a compromise between the limitations of the annotation process and our research goals. Some of the issues involved, such as interference, interpretation, word order or style, do not have straightforward solutions:

Interference: Being no experts in L2 acquisition, the annotators cannot be expected to spot cases of linguistic interference of L1 or some other language known to the learner. A sentence such as *Tokio je pěkný hrad* 'Tokio is a nice castle' is grammatically correct, but its author, a native speaker of Russian, was misled by 'false friends' and assumed *hrad* 'castle' as the Czech equivalent of Russian *gorod* 'town, city'.

Interpretation: For some types of errors, the problem is to define the limits of interpretation. The clause *kdyby citila na tebe zlobna* is grammatically incorrect, yet roughly understandable as 'if she felt angry at you'. In such cases the task of the annotator is interpretation rather than correction. The clause can be rewritten as *kdyby se na tebe cítila rozzlobená* 'if she felt angry at you', or *kdyby se na tebe zlobila* 'if she were angry at you'; the former being less natural but closer to the original. It is difficult to provide clear guidelines.

Word order: Czech constituent order reflects information structure. It may be hard to decide (even in a context) whether an error is present. The sentence *Rádio je taky na skříni* 'A radio is also on the wardrobe' suggests that there are at least two radios in the room, although the more likely interpretation is that among other things which happen to sit on the wardrobe, there is also a radio. The latter interpretation requires a different word order: *Na skříni je taky rádio*.

Style: Students often use colloquial expressions, usually without being aware of their status and the appropriate context for their use. Even though these expressions

[3] However, some authors intentionally avoid categorizing errors. They see categorization as an interpretation model, influencing access to the data. Instead, they use emendation as an implicit explanation for the errors (Fitzpatrick and Seegmiller, 2004).

[4] To the best of our knowledge, there is only one learner corpus built for a Slavic language (Stritar, 2009) – see Table 1. However, it is of a modest size of 35,000 words, and its error annotation is adopted from a Norwegian project ASK.

might be grammatical, we emend them with their standard counterparts under the rationale that the intention of the student was to use a register that is perceived as unmarked.

Our error annotation is primarily concerned with the acceptability of the grammatical and lexical aspects of the learner's language in a narrow sense. However, we anticipate that future projects would annotate the corpus with less formal properties of speech, such as the degree of achievement of a communicative goal.

5.2 Multi-level annotation

The optimal error annotation strategy is determined both by the goals and resources of the project and by the type of the language. A single-level scheme could be used for a specific narrowly defined purpose, such as the investigation of morphological properties of the learner language. However, in our case, to apply the single-level scheme would be problematic. First of all, our corpus should be open to multiple research goals. Thus a restricted set of linguistic phenomena or a single level of analysis is not satisfactory. As a result, it is necessary to register successive emendations and to maintain links between the original and the emended form even when the word order changes or in cases of dropped or added expressions. Another reason is the need to annotate errors spanning multiple forms, often in discontinuous positions.

In the ideal case, the annotator should be free to use an arbitrary number of levels to suit the needs of successive emendations, choosing from a set of linguistically motivated levels or introduce annotation tiers ad hoc. On the other hand, the annotator should not be burdened with theoretical dilemmas and the result should be as consistent as possible, which somewhat disqualifies a scheme using a flexible number of tiers. This is why we adopted a compromise solution with two levels of annotation, distinguished by formal but linguistically founded criteria to make the annotator's decisions easy. Thus the scheme consists of three interconnected levels (see Fig. 1, glossed in (1)):

- Level 0 – anonymized transcript of the hand-written original with some properties of the manuscript preserved (variants, illegible strings, self-corrections).
- Level 1 – forms wrong in isolation are corrected. The result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole.
- Level 2 – handles all other types of errors (valency, agreement, word order, etc.).

The correspondences between successively emended forms are explicitly expressed. Nodes at neighbouring levels are usually linked 1:1, but words can be joined (*kdy by* in Fig. 1) or split, deleted or added. These relations can interlink any number of potentially non-contiguous words across the neighbouring levels. Multiple words can thus be identified as a single unit, while any of the participating word forms can retain their 1:1 links with their counterparts at other levels.

Whenever a word form is emended, the type of error can be specified as a label at the link connecting the incorrect form at a lower level with its emended form at a higher level (such as *incorInfl* or *incorBase* for morphological errors in inflectional

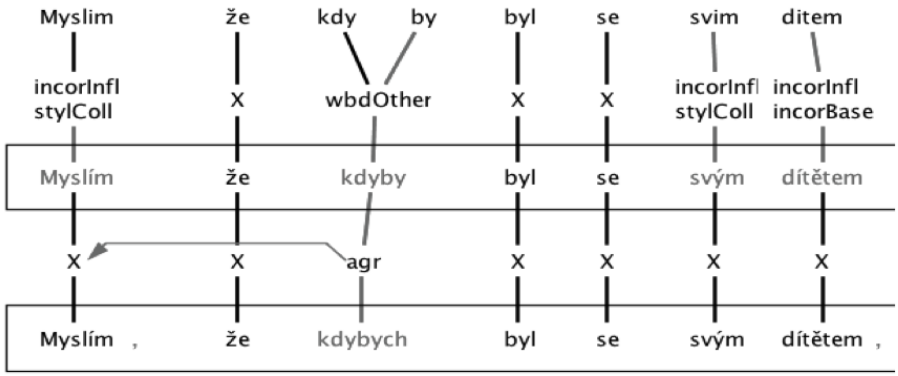

Fig. 1 Example of the three-level error annotation scheme

endings and stems, *stylColl* as a stylistic marker, *wbdOther* as a word boundary error, and *agr* as an error in agreement).

Each node may be assigned information in addition to the form of the word, such as lemma, morphosyntactic category or syntactic function.

(1) 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, . . . '

Manual annotation is supported by the purpose-built annotation tool *feat*[5] and followed by automatic post-processing (see §5.5).

5.3 Error categorization

A typical learner of Czech makes errors all along the hierarchy of theoretically motivated linguistic levels, starting from the level of graphemics up to the level of pragmatics. For practical reasons we emend the input conservatively to arrive at a coherent and well-formed result, without any ambition to produce a stylistically optimal solution, refraining from too loose interpretation. Where a part of the input is not comprehensible, it is marked as such and left without emendation. The taxonomy of errors is based on linguistic categories, complemented by a classification of superficial alternations of the source text, such as missing, redundant, faulty or incorrectly ordered element.

5.3.1 Errors at Level 1

Errors in individual word forms, treated at Level 1, include misspellings (also diacritics and capitalisation), misplaced word boundaries but also errors in inflectional and derivational morphology and unknown stems – fabricated or foreign words. Except

[5] See http://purl.org/net/feat.

Table 2 Errors at Level 1

Error type	Description	Example
incorInfl	incorrect inflection	*pracovají* v továrně; bydlím s *matkoj*
incorBase	incorrect word base	lidé jsou moc *mérný*; musíš to *posvětlit*
fwFab	non-emendable, „fabricated“ word	pokud nechceš slyšet *smášky*
fwNC	foreign word	váza je na *Tisch*; jsem v *truong*
flex	supplementary flag used with fwFab and fwNC marking the presence of inflection	jdu do *shopa*
wbdPre	prefix separated by a space or preposition without space	musím to *při pravit*; *veškole*
wbdComp	wrongly separated compound	*český anglický* slovník
wbdOther	other word boundary error	*mocdobře*; *atak*; *kdy koli*
stylColl	colloquial form	*dobrej* film
stylOther	bookish, dialectal, slang, hyper-correct	holka s *hnědými očimi*
problem	supplementary label for problematic cases	

Table 3 Errors at Level 2

Error type	Description	Example
agr	violated agreement rules	to jsou *hezké* chlapci; Jana *čtu*
dep	error in valency	bojí se *pes*; otázka *čas*
ref	error in pronominal reference	dal jsem to jemu i *jejího* bratrovi
vbx	error in analytical verb form or compound predicate	musíš *přijdeš*; kluci *jsou* běhali
rflx	error in reflexive expression	dívá na televizi; Pavel *si* raduje
neg	error in negation	žádný to *ví*; *půjdu ne* do školy
lex	error in lexicon or phraseology	jsem *ruská*; dopadlo to *přírodně*
use	error in the use of a grammar category	pošta je *nejvíc blízko*
sec	secondary error (supplementary flag)	stará se o *našich holčičkách*
stylColl	colloquial expression	viděli jsme *hezký* holky
stylOther	bookish, dialectal, slang, hyper-correct expression	rozbil se mi *hadr*
stylMark	redundant discourse marker	*no*; *teda*; *jo*
disr	disrupted construction	*kratka jakost vyborné ženy*
problem	supplementary label for problematic cases	

for misspellings, all these errors are annotated manually. The result of emendation is the closest correct form, which can be further modified at Level 2 according to context, e.g., due to an error in agreement or semantic incompatibility of the lexeme. See Table 2 for a list of errors manually annotated at Level 1. The last three error types (*stylColl*, *stylOther* and *problem*) are used also at Level 2.

The rule of “correct forms only” at Level 1 has a few exceptions: a faulty form is retained if no correct form could be used in the context or if the annotator cannot decipher the author’s intention. On the other hand, a correct form may be replaced by another correct form if the author clearly misspelled the latter, creating an unintended homograph with another form.

5.3.2 Errors at Level 2

Emendations at Level 2 concern errors in agreement, valency, analytical forms, pronominal reference, negative concord, the choice of aspect, tense, lexical item or id-

iom, and also in word order. For the agreement, valency, analytical forms, pronominal reference and negative concord cases, there is usually a correct form, which determines some properties (morphological categories) of the faulty form. Table 3 gives a list of error types manually annotated at Level 2. The automatically identified errors include word order errors and subtypes of the analytical forms error *vbx*.

5.4 Evaluation of the error mark-up

There is no widely accepted metric evaluating the consistency of annotation of learner corpora. In the current annotation practice of non-native speakers' corpora, it is common to have ill-formed texts tagged by a single annotator, despite problems in reliability and evaluation. A general shift towards multiple annotation of learner corpora is imminent.

The issue of singly annotated learner texts, used as application training data, was raised for the first time by Tetreault and Chodorow (2008), who investigated native-speakers' classification of prepositions usage. They concluded that two native annotators performing the task of tagging errors in prepositions on the same text reach at best an agreement level on the border between moderate and substantial (their kappa value was $\kappa = 0.63$ – the metric is explained in Section 4.1 below). Meurers (2009) also discusses the issue of verification of error annotation validity, viewing the lack of studies investigating inter-annotator agreement in the manual annotation of non-native speakers texts as a serious barrier for the development of annotation tools.

5.4.1 Inter-annotator agreement (IAA)

The manual annotation of *CzeSL* was evaluated using the metric κ (kappa, Cohen, 1960), widely accepted as the standard measure of inter-annotator agreement, especially for tagged corpora. The values of κ are within the interval $[-1, 1]$, where $\kappa = 1$ means perfect agreement, $\kappa = 0$ agreement equal to chance, and $\kappa = -1$ "perfect" disagreement.

5.4.2 Sample data

The data for the annotation were selected from a database compiled for *CzeSL*. The sample consists of 74 texts totalling 9848 tokens. Most of them were written by native speakers of Russian; the texts are classified according to the CEFRL scale as A2 or B1.

5.4.3 Method

The sample was annotated by 14 annotators. They were split into two groups: Annotators A and Annotators B. Each group annotated the whole sample independently. On average each annotator processed 1475 words in 11 texts.

Table 4 Inter-annotator agreement on selected tags

Tag	Type of error	Only A	Only B	A & B	κ
incorSum	incorStem+incorInfl	168	130	894	0.84
incorStem	Incorrect stem	167	165	559	0.61
incorInfl	Incorrect inflection	173	130	250	0.75
wbdSum	Incorrect word boundary	14	21	45	0.72
fwSum	fw+fwFab+fwNc	25	17	18	0.46
fw	"Non-Czech" expression	4	6	1	0.17
fwFab	Author' coinage	23	13	3	0.14
fwNc	Foreign/unidentified form	10	9	3	0.24
stylColl L1	Colloquial style at L1	10	2	2	0.25
agr	Agreement violation	82	99	110	0.54
dep	Errors in expressing syntactic dependency	99	118	87	0.43
neg	Incorrectly expressed negation	11	9	9	0.47
stylColl L2	Colloquial style at L2	14	14	10	0.42
lex	Lexical or phraseology error	107	131	74	0.37
rflx	Incorrect reflexive expression	6	11	3	0.26
use	Improper use of tense, aspect etc.	60	74	19	0.21
vbx	Ill-formed complex verb forms	20	9	3	0.17
ref	Incorrect pronominal references	14	17	3	0.16
sec	Secondary (consequent) error	45	18	4	0.11

At L1, the annotators chose from 8 tags for morphological, orthographical and word-boundary errors, and "non-Czech" expressions.[6] At L2, syntactic, morphosyntactic, lexical and stylistic errors were captured by 15 tags. Stylistically marked expressions could be assigned additional tags at both levels.[7] The results of inter-annotator agreement for the domain categories (*incor*, *wbd*, *fw* and *styl*)[8] were summed up, without distinguishing the particular error subtypes.

5.4.4 Results

Table 4 summarizes the distribution of selected error tags. The column "Only A" shows counts for each tag used by annotators in group A but not by those in group B. Similarly for the next column. The following column shows cases when both groups agreed. The last column gives the agreement measure κ.

The table shows that on L1 the annotators tend to agree in the domain categories *incorSum*, *wbdSum* and *fwSum*, i.e., for incorrect morphology, for improper word boundaries and for "non-Czech" expressions in general ($\kappa > 0.8$, $\kappa > 0.6$, and $\kappa > 0.4$, respectively). IAA was lower ($\kappa < 0.4$) for categories with a fuzzy interpretation, where a target hypothesis is difficult to establish, such as subcategories of *fw*, used

[6] Please note that here and below the abbreviations L1 and L2 refer to the levels of the annotation scheme, rather than to the first or second language.

[7] The numbers of tags given here correspond to a slightly outdated taxonomy and differ from the current state, presented in Section 5.3.

[8] The error taxonomy is hierarchical – error types are partitioned into domains, which are further divided into more specific subcategories, tagged manually or automatically. For example, the domain of complex verb form errors on L2 can be further specified as errors in analytical verb forms (*cvf*), modal verbs (*mod*), verbo-nominal predicates, passive or resultative form (*vnp*).

to tag attempts to coin a new Czech lexeme (*fwFab*), or foreign/unidentified strings of words (*fwNc*). Even the choice between the two subcategories was problematic (accounting for 26% of the total number of cases where the two annotators differed in the use of these two tags). This is true especially in cases where an annotator is not able identify the origin of the lexeme.

At L2 the annotators agree ($\kappa > 0.4$) in agreement errors (*agr*) and errors in expressing syntactic dependency (*dep*), and also in the well-defined category of errors in negation (*neg*). However, pronominal references (*ref*), secondary (consequent) errors (*sec*) and surprisingly also analytical verb forms and complex predicates (*vbx*), show a very low level of IAA, even though they are identifiable by formal linguistic criteria. In all these three cases, the distribution of tags and the annotators' feedback suggest that the annotation manual fails to provide enough guidance and formal criteria in distinguishing between the error types *ref* vs. *agr* and *ref* vs. *dep* (in either case the disagreement represents 19% of all the inconsistent uses of the tag *ref*). The use of tags for lexical and usage errors is highly dependent on the annotator's judgment, and the results are low as expected.

IAA in the distribution of tags for lexical and usage errors is within the range $0.2 < \kappa < 0.4$. The usage of these tags is highly dependent on the annotator's judgment, and the results are low as expected. An analysis has revealed that the tag *lex* has a systematic distribution: if the lexemes differ in their meaning distinctly, the annotators agree in their emendations in most of the cases; if the lexemes show semantic proximity, the annotators highly disagree in the emendation and therefore also in the consequent annotation. See the following examples (2) and (3).

(2) **R0:** *v pekařství kupuju* ***housenky***
'I buy **caterpillars** in the baker's shop'
R2: A1: … ***housky***$_{\text{LEX}}$ 'buns'
A2: … ***housky***$_{\text{LEX}}$ 'buns'

(3) **R0:** *když se dívá na* ***druhý*** *kultury*
'when one looks at **other** cultures'
R2: A1: *když se dívá na* ***druhé***$_{\text{AGR+STYLCOLL}}$ *kultury* 'other'
A2: *když se dívá na* ***jiné***$_{\text{LEX+AGR+STYLCOLL}}$ *kultury* 'different'

5.4.5 Error tags depend on emendation

Analysis of the tagged data (see Table 5) shows that the disagreement in using error tags is not necessarily caused by an annotator's fault, but could rather be dependent on the choice of the emended form (the target hypothesis), both on the current and the preceding level. For example, from the 181 cases of different use of the tag *agr*, 70 cases (39%) have a different L2 emendation. See (4) for an example.

(4) **R0:** *a kdyz stratil manzel*
R2: A1: *a když ztratí*$_{\text{AGR}}$ *manžela*$_{\text{DEP}}$
'and when she loses her husband'
A2: *a když se*$_{\text{RFLX}}$ *ztratil manžel*
'and when the husband got lost'

Table 5 Disagreement on the emendation

tag	diff. tags distr.	diff. emend. at L2		diff. emend. at L1		diff. emend. total
agr	181	70	39%	28	15%	54%
dep	218	76	35%	32	15%	50%
vbx	30	17	57%	6	20%	77%
rflx	18	12	67%	0	0	67%
lex	239	147	62%	10	4%	66%
use	135	66	49%	10	7%	56%

From the remaining 111 disagreements in the use of the tag *agr*, 28 cases (15%) differ in the emendation already on L1, as in (5).

(5) **R0:** *tezki období*
'a difficult period'
A1: **R1:** *těžký*$_{\text{INCORSTEM+INCOINFL}}$ *období*
R2: *těžké*$_{\text{AGR+STYLCOLL}}$ *období*
A2: **R1:** *těžké*$_{\text{INCORSTEM+INCOINFL}}$ *období*
R2: *těžké období*

In all these cases, tagging is correct vis-à-vis the selected emendation. Currently, we investigate the impact of emendation on error annotation at the individual levels, but we can already support the requirement of explicit target interpretation in the annotation scheme (Lüdeling, 2008). The scheme can thus be verified by the calculation of IAA in the distribution of the tags, depending on the final hypothesis (cf, i.a., Meurers, 2009).

5.4.6 Outline of the possible causes of the annotators disagreement

We can identify the following causes of the annotators' disagreements:

1. Invalid or imprecise annotation scheme: Generally, the annotators' disagreement can be caused by the annotation scheme itself: if it includes invalid tags or misses some necessary tags, or if the definition of a tag misleads the annotator. In the case of trial tagging of a sample of *CzeSL* data, it was problematic in several points, such as poorly distinguished subtypes of word boundary error (*wbd*), fuzzy definition of the error in pronominal reference (*ref*), also in contrast to the *agr* and *dep* types, or an imprecise boundary between the error due to a wrong choice of verbal tense (*use*) and the error in the analytical verb form (*vbx*).

2. Insufficient screening and training of the annotators: The level of screening and training process has a significant effect on the IAA rate. Higher IAA was demonstrated for annotators exposed to extensive and detailed pre-annotation training. It would be interesting to test what kind of impact the annotators' exposure to Czech as a foreign language has on the consistency of their annotation.

3. Different target hypotheses: Some annotations require a considerable amount of interpretation, while each annotator can have her/his own interpretation because of age, gender, education, etc. Moreover, in the case of multilevel annotation, annotators can differ also on intermediate levels, even though their target hypothesis might be

Table 6 Manually and automatically assigned error tags at L1 and L2

Error tags	L1 only	L2 only	L1 and L2	Total
Manual	8	11	3	22
Automatic	1	6	0	7
Total	9	17	3	29

identical. However, the annotation scheme of *CzeSL*, supporting emendation on both levels, makes reasons for possible disagreements explicit.

5.5 Towards Automatic annotation

So far, the annotation is largely a manual enterprise, quite demanding in terms of annotators' time and expertise. In an effort to automate some tasks of the annotation process, taggers will be used to assign POS and morphological categories to word forms, possibly even at L0 (although it is not always obvious what the correct answer should be).

Some error tags (e.g., type of a spelling alternation, missing/redundant expression, inappropriate word order, see Table 6) are assigned automatically by comparing the original and the corrected versions of the forms and their morphosyntactic tags.

As a further boost to the annotation process and the annotators' productivity, the use of a spelling correction tool was considered to suggest corrections or even to provide a less reliable but cheaper emendation of a larger quantity of learner texts.

5.5.1 Automatic methods of emendation

One of the options to (partially) automate the task of emendation is to use a proofreading tool – a spell checker or a grammar checker. So far, we have experimented with *Korektor* (Richter, 2010), a spell checker that has some functionalities of a grammar checker, using a combination of lexicon, morphology and a syntax model.[9]

The tool was tested on texts produced by learners at intermediate or higher levels of proficiency, yet among the total 9,372 tokens (7,995 tokens excluding punctuation) 918 (10%) were not recognized by the morphological analyzer included in a Czech POS tagger (see *Morče* in Spoustová et al, 2007). Even more forms were judged as faulty by the annotators: 1,189 (13%) were corrected in the same way by both annotators at L1 and 1,519 (16%) at L2.

Results of the spell checker were compared with those of the morphological analyzer and with forms at L1 and L2, provided both annotators were in agreement. The spell checker was run in three (batch) modes: (i) "autocorrect" (as proofreader), (ii) "remove-diacritics" followed by "diacritics" (as diacritics assigner), and (iii) same as in (ii), followed by "autocorrect", the latter two to test the hypothesis that diacritics is a frequent source of errors.

[9] Flor and Futagi (2011) report similar results for *ConSpel*, a tool used to detect and correct non-word misspellings in English, using n-gram statistics based on the *Google Web1T* database.

Table 7 Comparison with morphological analyzer, which identified the total of 918 unknown forms

mode	c'cted	unkn	prec	recall	F-msr
cor	1151	888	0.77	0.97	0.86
dia	1176	795	0.68	0.87	0.76
c+d	1315	906	0.69	0.99	0.81

Although the morphological analyzer includes a guesser, it makes no attempt to correct an unknown word form, only guesses its morphosyntactic tag and lemma. The spell checker is deemed to be successful for a given form if the morphological analyzer treats it as unknown and the spell checker suggests a correction, or if the analyzer treats the form as known and the spell checker leaves it intact.

Table 7 shows figures for the morphological analyzer. The rows give results for the three modes: **cor** for "autocorrect" (i), **dia** for "remove-diacritics" followed by "diacritics" (ii), and **c+d** for the full sequence (iii). The column **c'cted** gives the counts for forms corrected by the spell checker run in the relevant mode. The column **unkn** gives the number of cases where the morphological analyzer happens to flag a form corrected by the spell checker as unknown. The results of the analyzer are assumed as truth for the purpose of calculating precision (**unkn/c'cted**) and recall (**unkn**/918, the latter figure representing all forms unknown to analyzer).

Precision is not really a fair measure here, because the analyzer never flags forms which are correct in isolation but faulty in a context, while the spell checker often manages to use local context to replace a form X with an orthographically close but morphosyntactically quite different for Y: *podlé→podle, jejích→jejich, žit→žít, libí→líbí, ze→že, divá→dívá, drahy→drahý, mel→měl, jích→jich, čine→číně*. Interestingly, diacritics seem to represent a substantial share of problems in learners' writings, and the preprocessing of the input by the diacritics remover and assigner (iii) means a significant improvement.

Corrections made by the annotators can be compared verbatim with those proposed by the spell checker. The spell checker scores whenever the form proposed by the relevant mode matches the form at L1 or L2, respectively. The two annotators must agree about the corrected form, only then it is seen as fit for comparison.

At L1 the total number of corrections (1189) is higher than the number of forms unknown to the morphological analyzer (918) because the annotators correct also misspellings which look like homographs with an existing form. The result is a lower recall. Precision stays roughly the same as in the previous comparison because in one aspect L1 is similar to the analyzer: it still largely abstracts from context. E.g., annotators are instructed to leave errors due to missed grammatical concord for L2. The data are shown in Table 8 – the column **c'cted** is identical to that in Table 7, but the **wrong** column shows the number of cases where the two annotators agree about an emended form, identical with the suggestion of the spell checker.

It is interesting to investigate cases where the spell checker does not agree with the annotators, but both the spell checker and the annotators indicate an error (170 such cases at L1 for mode c+d). In some of these cases, the simple "autocorrect" mode without the diacritics component fares better (in 30 cases out of 170). It seems that

Table 8 Comparison with corrections at L1, where annotators agreed on the total of 1189 wrong forms

mode	c'cted	wrong	prec	recall	F-msr
cor	1151	846	0.74	0.71	0.72
dia	1176	780	0.66	0.66	0.66
c+d	1315	904	0.69	0.76	0.72

Table 9 Where simple autocorrect mode is better

R0	c+d	R1=cor	R2	R2 gloss
plaži	*pláží*	*pláži*	*pláž*	'beach'
tydnů	*týdnů*	*týdnu*	*týdnu*	'week$_{dat/loc}$'
lepšé	*lepše*	*lepší*	*lépe*	'better'
jide	*lidé*	*jde*	*jde*	'goes'
vždicky	*vodičky*	*vždycky*	*vždycky*	'always'

Table 10 Comparison with corrections at L2, where annotators agreed on the total of 1519 wrong forms

mode	c'cted	wrong	prec	recall	F-msr
cor	1151	687	0.60	0.45	0.51
dia	1176	640	0.54	0.42	0.47
c+d	1315	745	0.57	0.49	0.53

removing and reassigning diacritics takes the spell-checker too far (Table 9). In some cases the L1 and L2 versions differ and none of the methods matches the contextually correct version of L2 (*pláž, lépe*).

In 150 cases the spell checker suggests a correction when L1 prefers the original, but in 37 cases the spell checker agrees with an annotator at L2 (in 16 cases with both), which means that the real precision is higher. The rest of the cases are mostly inflectional issues, often due to misassigned diacritics, but also quite a few errors in the annotation (shared by both annotators).

L2 is problematic for evaluation in its own right. Some error types handled here are due to wrong word order, style, phraseology and a few other that go beyond simple spell checking, even in a broader sense of some degree of contextual sensitivity. The figures in Table 10, otherwise similar to Table 8, should be interpreted accordingly.

The two-stage annotation scheme suggests the option to distinguish corrections of forms that are wrong in any context, from those that could be correct in isolation, or in a different context, i.e. to test the grammar-checking capabilities of the spell checker. However, *Korektor* does not quite match the annotation scheme. It is only possible to find a few individual cases of successful corrections of missed agreement or case government (in the order of tens). Again, as in all the previous cases, the mode combining diacritics remover, assigner and proofreader is the best scenario.

The results seem to justify the option to integrate the spell checker into the annotation workflow, even though its suggestions may not quite match the two distinct levels without tuning to the specific task and annotation scheme.

5.6 Tagging by taggers

Despite the benefits of annotators' insight and judgment, manual annotation is tedious and costly. On the other hand, automatic tools are more error-prone and cannot produce the sort of sophisticated annotation envisaged in the present project. Aware of these pros and cons, we are still interested in how far we could get without manual annotation. Due to the lack of methods targeting learner texts, we confronted some 'native Czech' tools (two taggers and a spell checker) with ill-formed input.

The two taggers are based on different concepts: *Morče* (Votrubec, 2006) uses a morphological analyser, preferring lexical and morphological diagnostics over syntactic context, while *TnT* (Brants, 2000) has the opposite strategy, relying on a lexicon extracted from training data. Both taggers were trained on the same tagset and include a method to handle unknown words. Because of the different strategies the taggers use to tag correct input, they respond differently to various types of deviations. A mutual comparison of their results is thus as interesting as their evaluation against a golden standard, which – in the case of ill-formed input – is a difficult concept anyway.

Identifying all errors would involve comparing manual annotations at L2 form-by-form with the original text at L0. In the current absence of such data, we used data obtained from the easier task of comparing L0 to L1, where all erroneous forms are emended to a closest correct version, disregarding context.

Table 11 presents data extracted from a sample of 93 texts including 12,681 word tokens, with 1,323 tokens (8.9%) identified as ill-formed by the morphological analyser. The two taggers agreed on the same tag in 405 cases, i.e. in 28.8% of the total of ill-formed tokens, and disagreed in 918 cases (71.2%). The figures are additionally split by 12 morphological categories constituting the tag. Column 1 (L0m x L0t) shows in which categories the two taggers disagree at L0 for the 918 tokens, where their tags do not match at least in one category. Agreement is significantly lower between categories largely determined by syntactic context (POS, Gender, Number, Case) as opposed to those determined lexically. Columns 2 (L0m x L1) and 3 (L0t x L1) show agreement rates of tags assigned by *Morče* and *TnT*, respectively, to all tokens at L0[10] in comparison with tags assigned by *Morče* to the corresponding tokens at L1.[11] *Morče* shows better results overall and in most categories. Columns 4 and 5 show agreement rates for an ill-formed subset of the sample used in Columns 2 and 3. Interestingly, *TnT* shows significantly better results, except in the categories of Person and Tense.

The difference between the two taggers is also reflected in the share of different POS categories assigned to ill-formed words. Table 12 shows that *Morče* has a more even distribution, but strongly disprefers all verbal categories.

To sum up, the comparison of the two taggers confirms the assumption that the differences in their strategies will have a significant effect on the interpretation of faulty forms. A more general observation concerns the comparison of the success

[10] The size of the sample is smaller than in the previous comparison at L0 only due to a more demanding procedure to obtain the data at L1.

[11] The reason why *Morče* was used to tag L1 is because it is currently the best tagger of Czech and we were only interested in the cross-tagger comparison on the ill-formed input at L0.

Table 11 Tags on L0 and L1 – percentages of agreement

	L0m x L0t	L0m x L1	L0t x L1	L0m x L1	L0t x L1
No. of tokens	918	2589	2589	314	314
Entire tag	0	84.1	79.0	19.1	26.1
POS	39.2	89.6	88.7	43.9	52.5
SubPOS	37.1	89.2	87.9	42.0	49.7
Gender	23.9	88.8	88.2	36.0	46.5
Number	36.9	91.1	91.2	49.0	63.1
Case	31.2	89.0	86.5	43.0	51.3
Possessive Gender	98.6	99.8	99.9	98.4	99.7
Possessive Number	99.5	99.8	99.7	99.0	99.7
Person	68.1	96.3	94.2	81.8	76.1
Tense	70.6	96.7	95.3	83.1	77.4
Grade	78.3	96.4	96.9	75.2	81.5
Negation	74.4	95.3	93.8	73.9	74.2
Voice	70.6	96.7	95.5	83.1	78.7

Table 12 Numbers of tags assigned to ill-formed words

POS	*Morče*	*TnT*	POS	*Morče*	*TnT*	POS	*Morče*	*TnT*
adjective	158	94	noun	499	441	finite verb	32	129
adverb	118	21	preposition	10	–	particle	8	–
gradable adverb	31	11	infinitive	7	41	l-participle	10	119
						passive pcple	1	29

rate of the two taggers on the ill-formed input: *TnT* loses ground in a context with many errors but outperforms *Morče* on faulty forms, while *Morče* strongly disprefers verbs and works better in general.

6 Conclusion

It is no simple task to design an annotation scheme for a learner corpus and to maintain consistency in the annotated texts, both in a way that would reflect most demands of the corpus users. One of the main reasons is that annotating learner texts tends to be a highly specific enterprise, and even seemingly similar projects do not offer enough guidance – solutions are often too specific to a language or to the project concept and user requirements. On the other hand, annotation itself is quite rewarding due to the plentiful feedback from the annotators about all aspects of the task and, of course, about the learners' interlanguage.

More specifically, our experience shows that the rules for tagging morphosyntactic errors are relatively easy to formalize and it is thus possible to obtain a high inter-annotator agreement for them. However, we were unable to obtain a similarly robust annotation of semantic errors, which are much more dependent on subjective judgement. It is even unclear whether it is desirable to aim to standardize their annotation.

The pilot study, where two POS taggers and a spell checker were applied to ill-formed input, confirmed the viability of a partially or even fully automatic annotation as an alternative to manual-only annotation, especially when the demand for large data is higher than concerns about the error rate. It remains to be seen to what extent the comparison of results of multiple taggers, based on different tagging strategies, can lead to usable interpretations of faulty forms.

Acknowledgements The authors wish to express thanks to Michal Richter for his proof-reading tool and related support, and to other members of the project team, namely Karel Šebesta, Milena Hnátková, Tomáš Jelínek, Vladimír Petkevič, and to Hana Skoumalová, also for her generous help in the preparation of data.

References

Altenberg B, Tapper M (1998) The use of adverbial connectors in advanced Swedish learner's written English. In: Granger S (ed) Learner English on Computer, Longman, London, p 80–93

Bedřichová Z, Šebesta K, Škodová S, Šormová K (2011) Podoba a využití korpusu jinojazyčných a romských mluvčích češtiny: CZESL a ROMi [Form and utilization of a corpus of non-native and Romany speakers of Czech: CZESL and ROMi]. In: Čermák F (ed) Korpusová lingvistika Praha 2011, Ústav Českého národního korpusu, Nakladatelství Lidové noviny, Praha, Studie z korpusové lingvistiky, vol 2, pp 93–104

Brants T (2000) TnT – a statistical part-of-speech tagger. In: Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46

Díaz-Negrillo A, Fernández-Domínguez J (2006) Error tagging systems for learner corpora. Resla 19:83–102

Fitzpatrick E, Seegmiller S (2004) The montclair electronic language database project. In: Upton UCTA (ed) Applied Corpus Linguistics: A Multidimensional Perspective, Rodopi, p 223–238

Flor M, Futagi Y (2011) Automatic correction of non-word misspellings and generation of learner language corpora. In: Learner Corpus Research 2011 - 20 years earner Corpus Research 2011 - 20 years of learner corpus research: learner corpus research: Looking back, moving ahead, Centre for English Corpus Linguistics, Université catholique de Louvain, Louvain-la-Neuve

Granger S (1999) Use of tenses by advanced EFL learners: Evidence from error-tagged computer corpus. In: Hasselgård H, Oksefjell S (eds) Out of Corpora - Studies in Honour of Stig Johansson, Atlanta, Amsterdam, URL http://hdl.handle.net/2078.1/76322

Granger S (2003) Error–tagged learner corpora and call: A promising synergy. CALICO journal 20:465–480

Granger S (2008) Learner corpora. In: Lüdeling A, Kytö M (eds) Corpus Linguistics. An International Handbook, HSK 29. 1., vol 1, Mouton De Gruyter, Berlin/New York, pp 259–274

Leech G (1998) Preface. In: Granger S (ed) Learner English on Computer, Addison Wesley Longman, London, p xiv–xx

Leńko-Szymańska A (2004) Demonstratives as anaphora markers in advanced learners' English. In: G Aston SBDS (ed) Corpora and Language Learners, John Benjamins, Amsterdam, p 89–107

Lüdeling A (2008) Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Grommes P, Walter M (eds) Fortgeschrittene Lernervarietäten, Niemeyer, Tübingen, p 119–140

Meurers D (2009) On the automatic analysis of learner language: Introduction to the special issue. CALICO Journal 26(3):469–473, URL http://purl.org/dm/papers/meurers-09.html

Nesselhauf N (2005) Collocations in a Learner Corpus. John Benjamins, Amsterdam

Pravec NA (2002) Survey of learner corpora. ICAME Journal 26:81–114

Richter M (2010) Pokročilý korektor češtiny [An advanced spell checker of Czech]. Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague

Ringbom H (1998) Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In: Granger S (ed) Learner English on Computer, Longman, Harlow, p 41–52

Selinker L (1972) Interlanguage. IRAL 10:209–231

Spoustová D, Hajič J, Votrubec J, Krbec P, Květoň P (2007) The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, Association for Computational Linguistics, Praha, Czechia, pp 67–74

Stritar M (2009) Slovene as a foreign language: The pilot learner corpus perspective. Slovenski jezik – Slovene Linguistic Studies 7:135–152

Šebesta K (2010) Korpusy češtiny a osvojování jazyka [Corpora of Czech and language acquistion]. Studie z aplikované lingvistiky/Studies in Applied Linguistics 1:11–34

Štindlová B (2011) Evaluace chybové anotace v žákovském korpusu češtiny [Evaluation of error mark-up in a learner corpus of Czech]. PhD thesis, Charles University, Faculty of Arts, Prague

Van Rooy B, Schäfer L (2003) An evaluation of three POS taggers for the tagging of the Tswana Learner english corpus. In: D Archer AWTM R Rayson (ed) Proceedings of the Corpus Linguistics 2003 Conference Lancaster University (UK), UCREL, Lancaster University, Lancaster, p 835–844

Votrubec J (2006) Morphological tagging based on averaged perceptron. In: WDS'06 Proceedings of Contributed Papers, Matfyzpress, Charles University, Praha, Czechia, pp 191–195

Waibel B (2008) Phrasal verbs. German and Italian learners of English compared. VDM, Saarbrücken

Xiao R (2008) Well-known and influential corpora. In: Lüdeling A, Kytö M (eds) Corpus Linguistics. An International Handbook, Handbooks of Linguistics and Communication Science [HSK] 29.1, vol 1, Mouton de Gruyter, Berlin and New York, pp 383–457