Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems

Martí Quixal[1,2,3] **Björn Rudzewitz**[1,2] **Elizabeth Bear**[1,4] **Detmar Meurers**[1,2]

[1]Department of Linguistics, University of Tübingen, Germany
[2]LEAD Graduate School and Research Network, University of Tübingen, Germany
[3]Department of School Psychology, University of Tübingen, Germany
[4]Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany

marti.quixal@psycho.uni-tuebingen.de,br@sfs.uni-tuebingen.de,
elizabeth.bear@uni-tuebingen.de,dm@sfs.uni-tuebingen.de

Abstract

Integrating an adaptive Intelligent Tutoring System (ITS) in real-life school contexts requires coverage of the official curricula, which necessitates a broad range and number of activities to practice the official set of language phenomena. In the context of developing an adaptive ITS for English as a Foreign Language, we propose a method to automatically derive rich activity models from ordinary exercise specifications. The method identifies the language means being covered from the curriculum by processing the language used in the exercise and exemplary answers.

The analysis serves two purposes: First, it informs material developers about the extent to which the materials appropriately cover the language means to be practiced according to the curriculum. Second, it helps establish a direct link between rich activity and learner models, as needed for adaptively sequencing activities.

The approach includes (1) an NLP-based information extraction module annotating language means using a pedagogically-informed categorization, and (2) a tool to generate activity models offering information on the language properties of each activity in quantitative, qualitative, specific or aggregated terms. We exemplify the benefits of the method proposed in the design of materials for an ITS for language learning used in school.

1 Introduction

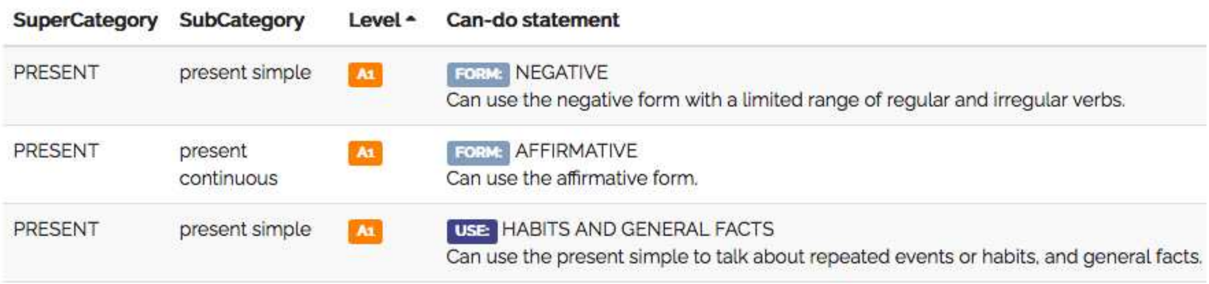
Foreign language teaching and learning in schools is typically regulated by education policy makers in state or national curricula that define which language aspects should be mastered in which grade. The curricula guide the creation of learning materials and textbooks, with publishing houses developing the materials for each grade, often followed by a government authority confirming whether the material appropriately covers the curriculum.

While the curriculum characterizes the envisioned language learning goals, teachers know that every student learns and makes progress in different ways. The substantial heterogeneity of classes in principle requires differentiation strategies that cater to the diverse learning paces and processes (Tomlinson, 2015), a highly non-trivial task (Martin-Beltrán et al., 2017). Instruction strategies supported by Intelligent Tutoring Systems (ITSs) have been shown to be effective, with most approaches targeting STEM subjects (Ma et al., 2014; VanLehn, 2011), but some recent work also focusing on foreign language learning (Choi, 2016; Meurers et al., 2019).

Complementing face-to-face instruction with ITSs makes it possible to support individual language learners by allowing them to practice with scaffolding feedback (Meurers et al., 2019). In addition, adaptive ITSs can select and sequence activities based on their difficulty in relation to the learner's knowledge and the learning goal, which presupposes the existence of both an activity and a learner model.

In this paper, we introduce an approach that facilitates the automatic derivation of activity models that can be used to assess curriculum compliance and support individual learning sequences in line with the principles of instructed Second Language Acquisition (Loewen and Sato, 2017).

Martí Quixal, Björn Rudzewitz, Elizabeth Bear and Detmar Meurers 2021. Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems. *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*. Linköping Electronic Conference Proceedings 177: 15–27.



SuperCategory	SubCategory	Level	Can-do statement
PRESENT	present simple	A1	FORM: NEGATIVE Can use the negative form with a limited range of regular and irregular verbs.
PRESENT	present continuous	A1	FORM: AFFIRMATIVE Can use the affirmative form.
PRESENT	present simple	A1	USE: HABITS AND GENERAL FACTS Can use the present simple to talk about repeated events or habits, and general facts.

Figure 1: EGP example descriptors for grammatical accuracy for CEFR A1 level

After introducing related work on identifying language phenomena in learner language and in activity models in section 2, we describe the implementation context of our approach and the resources developed in section 3. We then present and exemplify the process of generating activity models in section 4, showcase the application of the approach in the educational context in terms of curriculum coverage and ITS development in section 5, and conclude with a discussion of limitations and future work.

2 Related work

Learning a foreign language requires being exposed to, practicing and producing the language in question (Gass and Mackey, 2013). A range of pedagogical techniques are designed to engage learners in functionally using language, and a balance between fluency and accuracy as well as between receptive and productive skills is sought (Brown, 2007). Integrating ITSs in a school context has the potential advantage of enabling teachers to focus on the communicative aspects of language in the classroom, while the system supports individualized learning of grammar, vocabulary, listening and reading skills – aspects where individual differences also play an important role (Dörnyei and Skehan, 2003).

The definition of fine-grained foreign language curricula including a formal specification of language structures based on communicative goals is an endeavor argued for by modern approaches to language instruction (Estaire and Zanón, 1994; Bachman and Palmer, 1996). However, the link between the communicative goals and the linguistic syllabus is rarely made explicit in practice.

In an effort to spell out aspects of the CEFR linguistic competence scales (Council of Europe, 2020, p. 130), the English Grammar Profile (EGP) Project[1] and Pearson's Global Scale of English[2] have compiled databases linking can-do statements to vocabulary and grammar structures. They include detailed information on the linguistic structures as well as the mastery levels at which such structures are produced (not just taught).

The EGP organizes its inventory based on 19 super-categories[3] (from *adjectives* to *verbs* over *adverbs*, *clauses*, etc.), with up to ten sub-categories each (e.g., for the super-category *present*, the sub-categories are *simple* and *continuous*). For each sub-category, a number of level-specific can-do statements is provided. Figure 1 illustrates the first three items for the super-category *present (tenses)* for the CEFR level A1, including both form and functional use characterizations.

The EGP is designed to help analyze and evaluate learner productions. To analyze teaching materials, verify curriculum coverage and generate activity models supporting adaptive selection and sequencing in an ITS, we need to go a step further and analyze the language in the input given that it "[i]s an essential component for learning in that it provides the crucial evidence from which learners can form linguistic hypotheses" (Gass and Mackey, 2015). When considering practice, we need to analyze the learner activities to determine which language students are expected to produce.

The few language tutoring systems that so far have been developed and used in real-life contexts (Heift, 2010; Nagata, 2009; Amaral and Meurers, 2011; Choi, 2016; Ziai et al., 2018) are based on manual activity specifications and do not provide a fine-grained characterization of the language means they cover. While some research tackles the task of automatically annotating texts with lin-

[1] https://englishprofile.org/english-grammar-profile

[2] https://english.com/gse/teacher-toolkit/user/grammar

[3] https://englishprofile.org/english-grammar-profile/grammatical-categories

guistic properties to support language-aware document retrieval (Chinkina and Meurers, 2016) or input enrichment and enhancement (Meurers et al., 2010), the work so far fell short of generating tutoring system activities that are pedagogically linked to a linguistic syllabus or curriculum.

The approach we are presenting in this paper goes a step further in automatically deriving fine-grained metalinguistic characterizations of the language used in or elicited by some given learning material, including both the linguistic phenomena targeted by the materials as well as those incidentally occurring in it.

3 Implementation context and resources

The research presented here is being carried out in the context of the development of Didi (http://didi.schule), an adaptive ITS for English as a Foreign Language based on the FeedBook system (Rudzewitz et al., 2017; Meurers et al., 2018). It integrates the feedback mechanisms from FeedBook and offers immediate, specific feedback on grammar (Rudzewitz et al., 2018), spelling (Ziai et al., 2019), and meaning (Ziai et al., 2018). Instead of offering exercises from an existing workbook, the Didi system provides independent exercises on more diverse levels of difficulty.

The minimal components of an ITS, such as Didi or FeedBook, are illustrated in Figure 2.

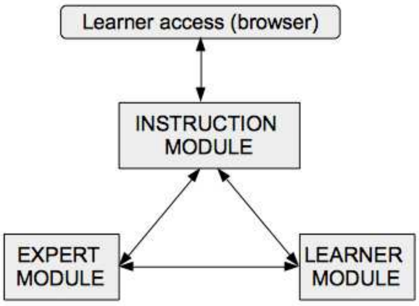


Figure 2: ITS architecture (adapted from Amaral, 2007, p. 85)

The method we present aims at using the linguistic structures identified in a given set of learning activities to link (i) language as an object of study (language as a system), which belongs to the expert module, (ii) language as organized and presented in instruction materials (language as a pedagogical goal), which is part of the instruction module, and (iii) language as knowledge that has been or is being acquired, which is part of the learner module (language as a competence). The three perspectives on language need to be anchored in a common characterization of language properties supporting the goals of the three modules of an ITS.

Our approach makes it possible to automatically populate the knowledge domain as part of the expert module on the basis of the language properties of the activities in the instruction module. The knowledge domain results as an aggregate of all the linguistic constructions found in the activities produced by material authors and organized as learning sequences. As we will see in section 5, it also allows us to monitor and make explicit learner competencies by enriching the learner model and ultimately perform adaptive sequencing.

As a starting point, we describe three resources that facilitate the automatization of this process: (i) a hierarchical structure of language phenomena relevant for English as a Foreign Language, (ii) a general linguistic annotation module, and (iii) a rule-based module for the annotation of language structures.

3.1 Knowledge hierarchy

The English as a Foreign Language (domain) knowledge of our ITS is organized as a hierarchy that consists of three levels of characterization exemplified in Figure 3. The first level includes categories such as word formation (morphology), sentence structure (syntax) and language use, levels of linguistic description common in Second Language Acquisition and Foreign Language Instruction. Each of these categories is in turn divided into smaller categories extracted and/or extended from the official curriculum for secondary schools, grades 7 to 9 (Kultusministerium, 2016, p. 50), which is the second level of characterization. This second level of characterization is exemplified in Figure 3 with superlative forms of adjectives, child nodes of the category word formation: regular forms (*reg. forms*), irregular forms (*irreg. forms*) and periphrastic forms (*most* + ADJ).

This third level of characterization is extracted or extended from the EGP, and it maps to level 2 categories so that each level 3 element relates to one and only one level 2 element. In Figure 3 this is exemplified with finer-grained labels for language means that are child nodes of the level 2 element *Superlative regular forms*: plain regular forms (*cheap - cheapest*), regular forms of ad-

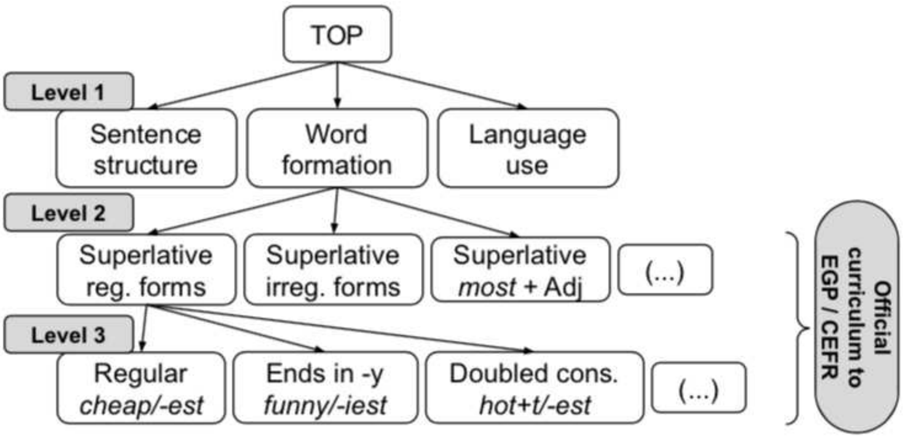


Figure 3: Hierarchical knowledge structure for English as a Foreign Language.

jectives ending in *-y* (*funny - funniest*) and regular forms ending in *-e* (*nice - nicest*). Moreover, the language means in level 3 constitute the specifications for the automatic analysis with the rule-based annotation tool.

3.2 Linguistic annotation

As we will describe in further detail in the following section, the input to the annotation module is the set of activities included in Didi.

The NLP analysis is realized in the Unstructured Information Management Architecture (UIMA, Ferrucci and Lally, 2004). As the first step, each language learning exercise provided as input is turned into a UIMA Common Analysis Structure (CAS) object. These CAS objects are linguistically annotated using the standard NLP tools specified in Table 1. Then the annotated CAS documents are exported as XMI files, the input format for the module responsible for the annotation of language means.

3.3 Annotation of language means

The module for the annotation of language means is implemented as a set of rule-based grammars in UIMA Ruta (Kluegl et al., 2016), a formalism and annotator development environment within UIMA that supports the robust and modular integration of this functionality in the processing pipeline. UIMA Ruta enables grammar writers to access annotations in the CAS that were provided by the NLP analysis modules and offers a set of operators and property check functions to map, review and remove annotations at the word, phrase, clause, sentence and document level.

Figure 4 exemplifies a UIMA Ruta rule that checks for the presence of a simple present tense form and, when found, records that there is a present simple verb form in terms of word formation and, in terms of sentence structure, that we are dealing with an affirmative sentence in the present. It thereby translates the NLP analysis output into two labels of level 3 in our knowledge hierarchy of English as a Foreign Language, namely "PresentSimpleForms" and "SyntAffirmativeSentencePresentSimple".

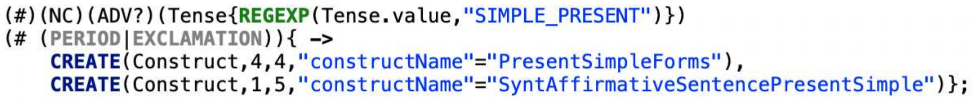
Currently the annotation module contains more than 200 rules. The module includes annotators for tenses (including present, past and future verb forms), comparatives (including comparatives and superlatives), passive voice, conditional sentences types 1 and 2, and relative clauses.

NLP task	**tool**
segmentation	ClearNLP (Choi and Palmer, 2012)
part-of-speech (POS) tagging	ClearNLP
dependency parsing	ClearNLP
lemmatization	Morpha (Minnen et al., 2001)
morphological analysis	Sfst (Schmid, 2005)

Table 1: NLP tools adding linguistic annotations as input to UIMA Ruta

4 Generation of activity models

Activity models, which belong to the instruction module of an ITS, are particularly important for



```
(#)(NC)(ADV?)(Tense{REGEXP(Tense.value,"SIMPLE_PRESENT")})
(# (PERIOD|EXCLAMATION)){ ->
    CREATE(Construct,4,4,"constructName"="PresentSimpleForms"),
    CREATE(Construct,1,5,"constructName"="SyntAffirmativeSentencePresentSimple")};
```

Figure 4: UIMA Ruta rule to annotate present verb forms and affirmative sentences.

supporting adaptive selection and sequencing of activities for a given user since they make explicit what the activity demands and offers.

Figure 5 shows the information included in our activity model. The first block shows the specifications provided manually during activity creation. The second block, shown in italics, lists the type of information added automatically using the NLP-based and other activity-specification-based annotation modules.

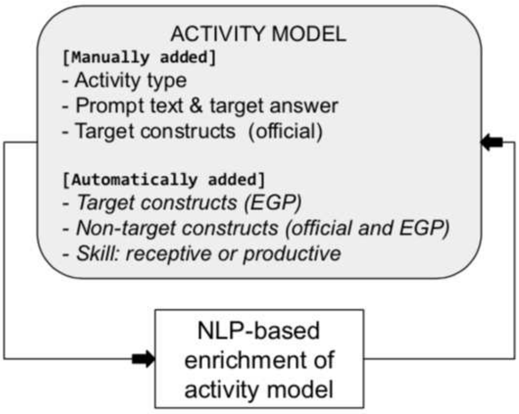

ACTIVITY MODEL
[Manually added]
- Activity type
- Prompt text & target answer
- Target constructs (official)
[Automatically added]
- Target constructs (EGP)
- Non-target constructs (official and EGP)
- Skill: receptive or productive
NLP-based enrichment of activity model

Figure 5: Linguistic enrichment of activity model

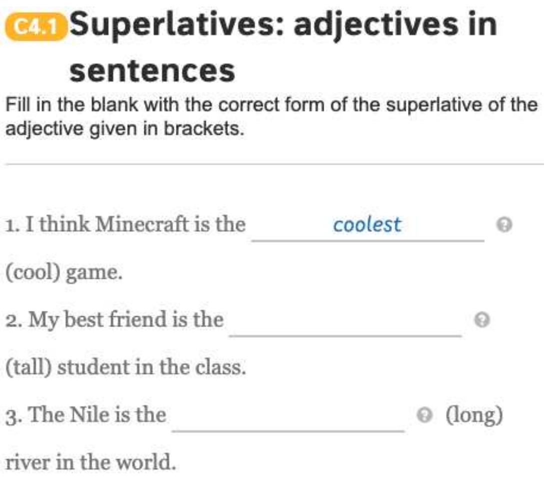
The manually determined properties include the activity's format (fill-in-the-blanks, multiple choice, etc.), the actual items, each including a prompt (textual or not), an expected answer (which may be typed in or selected), and optionally distractors. Among them there is also the learning goal, which maps to the level 2 language phenomena introduced in section 3.1 for which the activity has been designed. These phenomena become then language target (the target is to teach or learn them), as opposed to other language means, which are just accompanying the target of the activities – thus, non-target.

The automatically generated properties include language targets of level 3, and non-targets of level 2 and level 3. Non-target means are language elements present in the activities with which a specific language structure is to be practiced, but they do not belong together. For instance, to learn the use of comparatives, one needs to be able to produce sentences with them; therefore, a learner has to be able to use some sentence structure (e.g., basic SVO) and at least one tense form (e.g., the present simple).

The activity model also encodes the distinction between receptive and productive skills, which is computed on the basis of the activity's format, not its linguistic characteristics.

4.1 Input to NLP module

To illustrate the process, let us take a look at two sample activities with slightly different properties. Figure 6 shows part of a fill-in-the-blanks activity.



C4.1 Superlatives: adjectives in sentences

Fill in the blank with the correct form of the superlative of the adjective given in brackets.

1. I think Minecraft is the coolest (cool) game.

2. My best friend is the ______ (tall) student in the class.

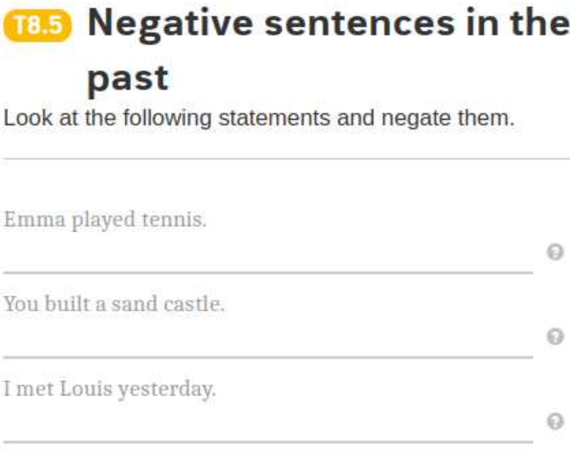
3. The Nile is the ______ (long) river in the world.

Figure 6: Activity C4.1 targeting superlatives

In this activity, students are given an adjective base form that has to be turned into its superlative form. According to specifications this is a fill-in-the-blanks activity, with items whose prompt consists of a sentence and whose answer length is a word. In addition, the activity is labeled with the language means in the curriculum *Superlative regular forms*, *Superlative irregular forms* and *Superlative most + Adj*, all of which are language means of level 2 in the hierarchy (see section 3.1).

Figure 7 shows a short answer activity. The activity includes a sentence as a prompt and requires complete sentence as a response. This activity gets the level 2 label *Sentences using the simple past*, a language structure appearing in the curriculum.

For this tutoring system, activity specification requires not only writing the instructions and



T8.5 Negative sentences in the past

Look at the following statements and negate them.

Emma played tennis.

You built a sand castle.

I met Louis yesterday.

Figure 7: Activity T8.5 on past simple negation

prompts in them but also entering a list of correct answers. For the activity in Figure 6, the expected answers are the superlative forms of the corresponding adjectives, but for the activity in Figure 7 the expected answers are the negated version of the sentences, such as “Emma did not play tennis” or “You did not build a sand castle”, for the first two items, respectively.

It is such activity specifications that are sent to the NLP-module performing the annotation of finer-grained language means (level 3).

4.2 Automatically generated properties

The first step of the automatic annotation process is the identification of language means based on the activity specification, using the NLP resources we introduced in section 3.

The second step performed in the annotation process distinguishes between so-called receptive and productive skills given that any linguistic phenomenon can be practiced in the context of understanding or producing language. What elements of an item are considered receptive or productive depends on the activity type. For fill-in-the-blanks activities, such as the one in Figure 6, the text in the expected answers for each blank constitutes the productive part (e.g, “coolest” in the first gap). In contrast, the language found in the text surrounding the blanks is handled as receptive since learners use them to complete the answer (e.g., “I think Minecraft is the … (cool) game.”). For short answer tasks, the receptive part are the prompts, and the productive parts are the answers to be elicited from the learners. For example, in Figure 9, the prompt “Emma played tennis” from the activity shown in Figure 7 is analyzed as language practiced in receptive mode (SyntAffirmativeSentenceSimplePast), while “Emma didn’t play tennis.” is language practiced in the productive mode (SyntNegativeSentenceSimplePast).

4.3 Information visualization

On this basis, we can systematically visualize the language means found in a given activity. Didi includes a visualization module that uses spider web charts to present this information.

Figure 8 illustrates the output for the fill-in-the-blanks activity targeting superlative forms we saw in Figure 6. We see that language means at the word and sentence level are classified as receptive or productive. For instance, a target goal at the word level, *SuperlativeFormRegularHigherDegree*, is classified as receptive once for “coolest”, which is given as a sample answer, and as productive multiple times for gaps such as “tallest” and “longest”, appearing in items 2 and 3 of the activity, respectively. At the sentence level, the target language means *AffirmativeSuperlativeSentence* and *InterrogativeSuperlativeSentence* are classified as productive, corresponding to the sentence containing the expected answer. Non-target means, such as *SyntAffirmativeSentencePresentSimple*, the rule for which was exemplified in Figure 4, are also represented in the spider web chart and can also be classified as receptive or productive.

Similarly, Figure 9 shows the spider web chart for the activity we saw in Figure 7, the one on the negative sentences in the past. In this activity, the affirmative sentences given as the prompts are classified as receptive, for instance, as *SyntAffirmativeSentenceSimplePast* at the sentence level. The expected answers contain the language mean *SyntNegativeSentenceSimplePast*, also at the sentence level and are classified as productive. In this activity, the annotation does not include non-target language means outside of the learning unit on tenses.

5 Applications of the approach

The approach described to enrich activity models is useful both from a pedagogical perspective for the design and selection of activities in relation to the curriculum and from the perspective of designing adaptive tutoring systems, where it supports the implementation of activity sequencing.

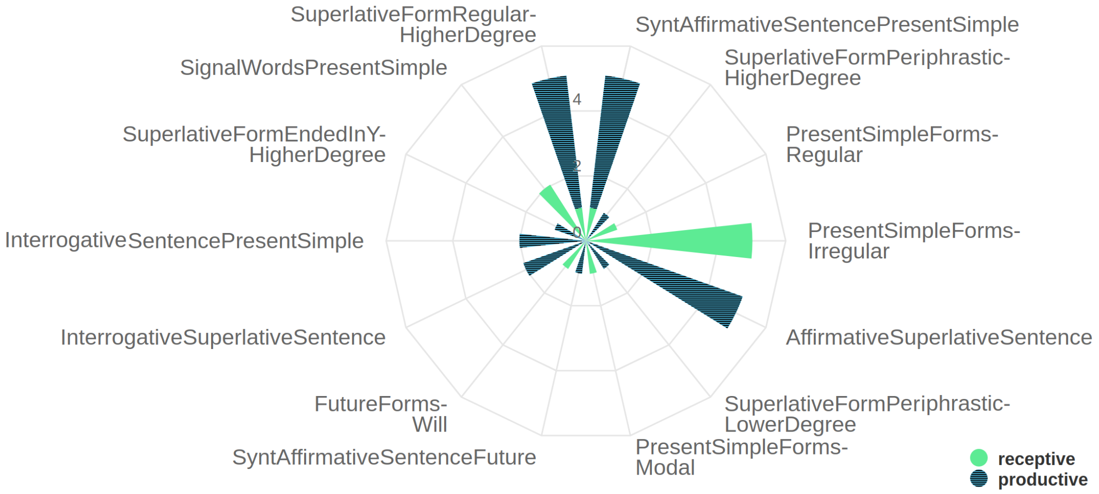


Figure 8: Visualization of annotated language means for activity C4.1

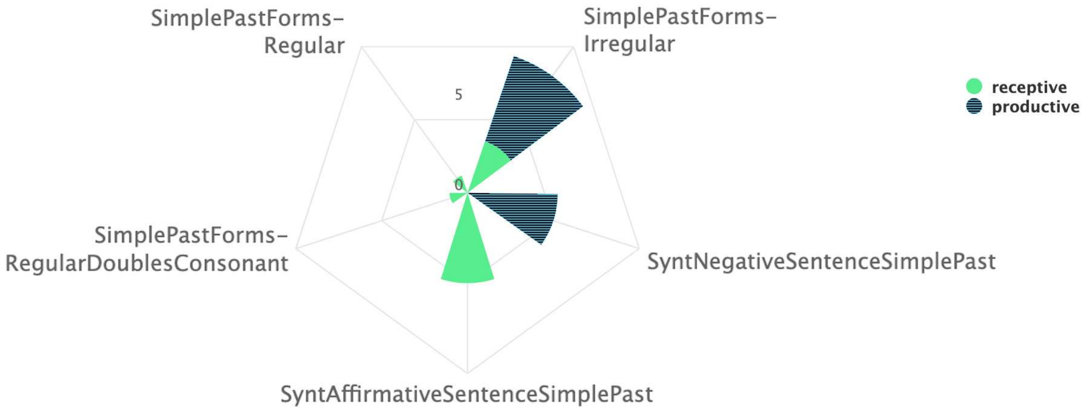


Figure 9: Visualization of annotated language means for activity T8.5

5.1 Evaluating curriculum coverage

A first approach to evaluating curriculum coverage can be carried out at the most abstract level of description, as in Table 2.

The table shows the number of automatically identified target and non-target language means at the receptive and productive level for the current set of activities implemented in four learning units. Pedagogically speaking, the table reflects that there are in total 845 opportunities either to produce (482) or to understand (363) one of the target language means of the tenses topic. We can also see that the numbers for the other three topics (comparatives, conditional sentences type 2 and elative clauses) are smaller. This tells us about the number of activities written for each of the topics which, as shown in the last column, is quite imbalanced – productive target language means in tenses amount to 56% of the opportunities to produce a piece of language in the current version of the materials.

The table also indicates that while tenses, comparatives and conditional sentences present a relatively balanced number of opportunities to practice target productive and receptive skills, relative clauses has a very low proportion (7%) of opportunities to practice target receptive skills. In this case a manual inspection of the activities in relative clauses confirms that the sequence of activities includes much more production activities than receptive ones.

If we take a look at the numbers under non-target language means, we see these are much higher and proportionally bigger for comparatives, conditional sentences type 2 and relative clauses. For instance, for comparatives the total number of non-target language means adds up to 417 (121

Table 2: Language means automatically identified in the activities of four learning units

	TARGET		NON-TARGET		
LEARNING UNIT	PROD.	REC.	PROD.	REC.	ACTIVITIES
tenses	482	363	133	152	49
comparatives	84	107	121	296	27
cond. sent. type 2	209	237	404	672	30
relative clauses	95	7	263	301	20
TOTAL	870	714	921	1,421	126

+ 296) while the total number of target language means adds up to 191 (84 + 107). A plausible explanation for this is the fact that although the learning of comparatives often focuses on word formation (building its forms) or some essential syntactic patterns (... ADJ *than* ..., ... *as* ADJ *as* ...), it is usually learned in the context of comparing different options (e.g., travel preferences, product prices and quality, etc.); since making comparisons requires the use of sentences that include different tenses and structures, a variety of non-target structures is expected here. Similar interpretations can be made for conditional sentences type 2 and relative clauses, two topics for which the use of sentences with all their underlying properties is required.

Finer-grained analyses of curriculum coverage are possible by quantifying the language means included in the materials as shown in Tables 3 and 4.

Table 3: Tenses: distribution of language means by level 1 categories

	TARGET		NON-TARGET	
CATEGORY	PROD.	REC.	PROD.	REC.
Word formation	248	233	95	106
Sentence structure	224	107	38	39
Language use	10	23	0	7
TOTAL	482	363	133	152

Table 3 offers a level 1 characterization of the opportunities to learn word formation, sentence structure and language use at the receptive and productive level on the unit on tenses. We can see that target language means are relatively proportionate between word formation and sentence structure, but not language use. At the same time, non-target means are much more frequent in word formation than in the other two categories.

Table 4 offers an even finer-grained representation of the distribution of language means – in this case for the category *word formation* in tenses. The table shows both target and non-target language means in productive and receptive skills. The horizontal line that divides the table in language means that are genuinely part of the grammar topic tenses and those that are not part of it.

Looking at the table, we can confirm that the unit on tenses has: (i) much more practice opportunities on the formation of irregular verbs (228 as target and 37 as non-target), than on any other verb form. However, we also see that some of the language means that are genuinely part of the grammar topic tenses are also used as non-target. This can be explained by activities in which a verb form is used to give a context in which then another verb form can be used. For instance, when practicing the past continuous forms, one will often see the pattern "*while* VP-PAST PARTICIPLE FORM ..., VP-PAST SIMPLE".

Now whether the presence and distribution of the language means as found in the learning activities in these units actually leads to mastery or not and whether they are compliant with a specific curricula is not within the scope of this paper. The goal of the paper is to show that this kind of evaluation is possible thanks to the information made explicit by the automated annotation strategy.

5.2 Automatic derivation of learner models

The rich activity models enable the ITS to generate learner models that track the progress of individual learners across activities. This serves two purposes: first, to inform learners about their observed competence in an inspectable, open learner model (Bull and Kay, 2006) and second, to inform the adaptive sequencing algorithm in Didi about the current level of proficiency of learners to suggest a suitable next exercise.

Whenever a learner works on an activity, the learner model for that learner records both the language means the learner was exposed to (i.e. the ones appearing in the activity) and the subset of language means that the learner was able to produce correctly. The learner model stores an update making explicit the exposure and accuracy for each level 3 language mean involved – together with a time stamp to enable temporal tracking. Taken together, the learner model records the dif-

Table 4: Detailed characterization of the language means found in *Tenses* at the morphology level

	Target		Non-target	
Language means: level 2 & 3	prod.	rec.	prod.	rec.
Future - *will*	4	1	0	1
Pres. cont. forms	18	1	0	0
Pres. simple forms: - irreg	9	8	0	7
- modal	5	10	0	2
- reg	18	5	0	1
Past cont. forms	16	0	0	1
Simple past forms: - doubCons	6	18	3	5
- in -y	7	10	3	5
- in -e	17	24	5	12
- irreg	111	116	17	20
- modal	2	5	0	1
- reg	35	35	13	18
Comparative intens.: - doubCons	0	0	1	1
- reg	0	0	0	2
Comparative equal.	0	0	0	2
Imperative forms	0	0	0	9
Passive - simple past	0	0	3	3
Past perfect forms	0	0	19	6
Pres. perfect forms	0	0	30	3
Reflexive pronoun	0	0	0	3
Rel. pron. - subject	0	0	1	4
Total	248	233	95	106

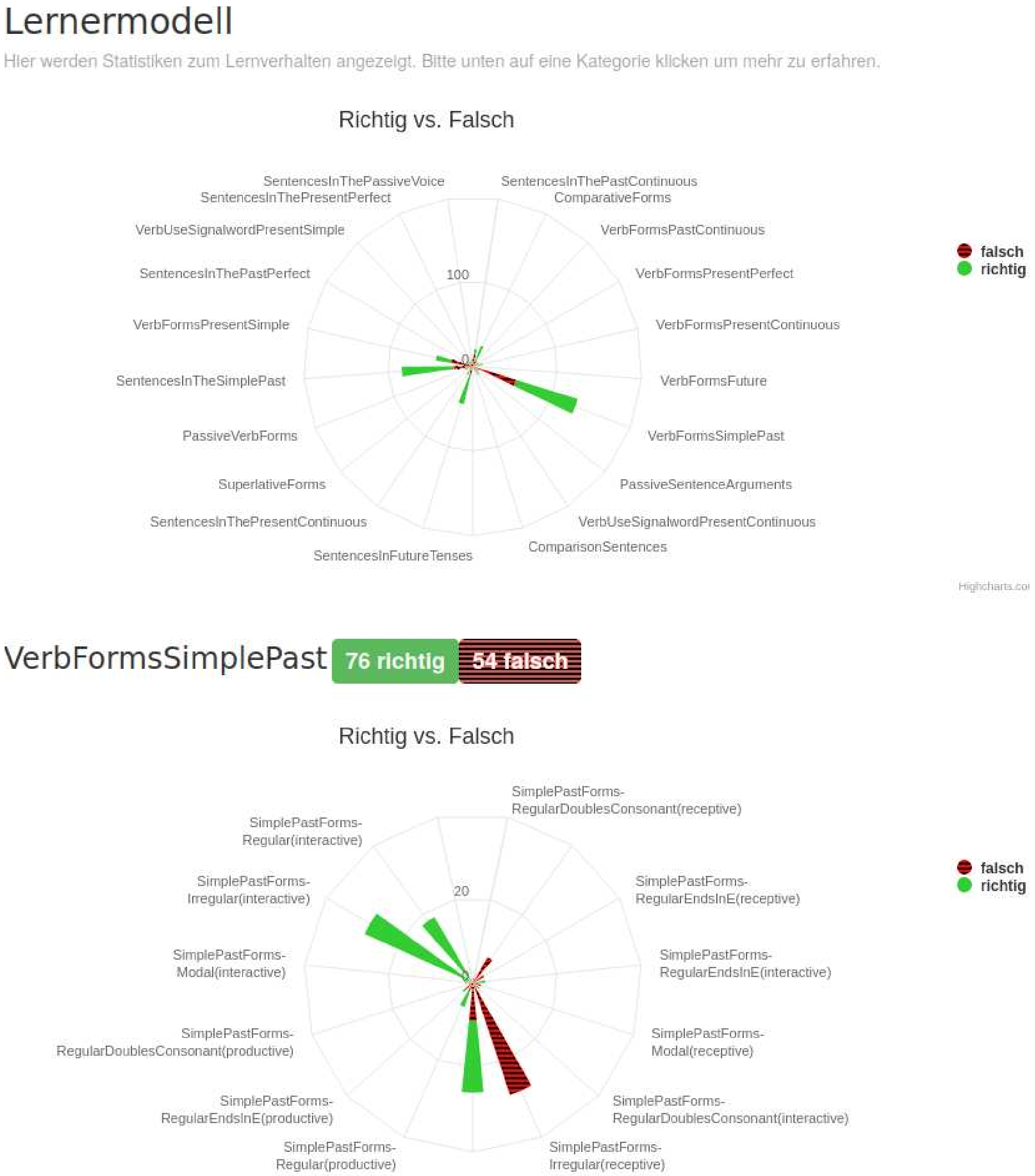
ference between what language means an exercise exposed a learner to and which of them the learner was able to produce. The learner model is updated independent of whether the language means have been marked as target or non-target.

The open learner model in Didi presents the collected information structured in two levels (Figure 10). On the top, the system presents the collected evidence aggregated for language means at level 2, providing the learner with an overview on the performance for all the pedagogically relevant categories. If learners want to get more detailed insights, the system displays a more detailed view of language means of level 3 below the aggregate view (bottom graph). For each of the level 2 labels, this detailed view lists a learner's performance for each of the language means of level 3 belonging to this level 2 label according to the domain model (cf. Figure 3), distinguishing receptive from productive skills. The learner model also presents language structures in the dimension *interactive* – learning opportunities in which language means need to be selected as opposed to typed in, e.g., in a multiple-choice task.

For example, in Figure 10, the category *VerbFormsSimplePast* appears in the top chart in the south east with a long green (correct usage) and shorter red (incorrect usage) bar. The bottom chart lists all the associated child language means, e.g., *SimplePastFormsRegular* or *SimplePastFormsModal*. This allows the learner to see which specific language means (s)he is struggling with the most – in this example it is *SimplePastFormsIrregular*.

5.3 Combining activity and learner models for adaptive sequencing

The learner and activity model together are the basis for the adaptive sequencing algorithm. This algorithm operates at the level of subsections within learning units, for which target structures are specified, and suggests a next suitable exercise to individual learners. In the first step, the system identifies the target language means that still need to be learned by filtering out all those structures for which the learner has obtained *mastery*. Mastery is assessed by comparing both the exposure to and accuracy achieved in the language means by externally defined thresholds in a configurable lookback window. Exposure is measured as the number of times an exercise provided an opportunity to practice a specific construction, and accuracy indicates how many times a specific learner was able to produce it correctly. The lookback window makes it possible to base decisions only on the recent performance, so that trying out different forms in earlier acquisition stages is not penalized. In the second step, the system queries exercises that contain the language means to be practiced by the learner. At this stage, Didi ranks the queried exercises using a linguistic affinity score by computing the closeness between the language



Lernermodell

Hier werden Statistiken zum Lernverhalten angezeigt. Bitte unten auf eine Kategorie klicken um mehr zu erfahren.

Richtig vs. Falsch
SentencesInThePassiveVoice
SentencesInThePresentPerfect
SentencesInThePastContinuous
ComparativeForms
VerbUseSignalwordPresentSimple
VerbFormsPastContinuous
SentencesInThePastPerfect
100
VerbFormsPresentPerfect
VerbFormsPresentSimple
VerbFormsPresentContinuous
0
SentencesInTheSimplePast
VerbFormsFuture
PassiveVerbForms
VerbFormsSimplePast
SuperlativeForms
PassiveSentenceArguments
SentencesInThePresentContinuous
VerbUseSignalwordPresentContinuous
SentencesInFutureTenses
ComparisonSentences
● falsch
● richtig
Highcharts.com

VerbFormsSimplePast 76 richtig 54 falsch

Richtig vs. Falsch
SimplePastForms-RegularDoublesConsonant(receptive)
SimplePastForms-Regular(interactive)
SimplePastForms-Irregular(interactive)
20
SimplePastForms-RegularEndsInE(receptive)
SimplePastForms-Modal(interactive)
0
SimplePastForms-RegularEndsInE(interactive)
SimplePastForms-RegularDoublesConsonant(productive)
SimplePastForms-Modal(receptive)
SimplePastForms-RegularEndsInE(productive)
SimplePastForms-RegularDoublesConsonant(interactive)
SimplePastForms-Regular(productive)
SimplePastForms-Irregular(receptive)
● falsch
● richtig

Figure 10: Open learner model visualizing performance on language means at level 2 (top) and level 3 (bottom)

means an activity offers and the current learning goals. The third step is to rank candidate exercises using pedagogically driven categories stored in the activity model (cf. Figure 5). For example, the adaptivity algorithm suggests closed activity types before open activities, activities with shorter gaps before activities with longer gaps, or activities where inflected forms are provided before tasks without any given lexical material. The automatically derived activity models allow for an activity selection process that not only takes into account what was actually learned as a target, but also everything learned as non-target.

6 Concluding remarks and future work

The work presented here shows how combining manual and automatic annotation of learning activities facilitates the enrichment of activity models. Making the linguistic properties explicit in this way supports a link between the language to be learned (expert model), the strategies to present this language as a learning goal (instruction model), and the language competence as recorded (learner model). The linguistically enriched activity models do not only take into account the language produced (or seen) when this was a learning goal, but also the language produced (or seen) as co-material – as a consequence of embedding the actual learning goals in more complex linguistic structures or larger language units.

While the approach described in this paper is fully implemented, it represents ongoing work and comes with certain limitations. First of all, there is currently no gold standard against which the accuracy of the NLP annotation module can be evaluated. However, the quantification of language means identified in the four learning units seems to indicate good face validity.

An additional limitation of the approach is that it only annotates language phenomena that appear in the input materials, which are used as a basis for specification. Comparing our aggregated annotations with resources such as the EGP informs us about the areas of language for which no activities exist – or appear only as non-target, which we have not systematically addressed yet.

Finally, the adaptivity algorithm described here is still under development and has not been tested in practice yet. Piloting and evaluating it in an authentic school context to assess the external validity is planned for the next project phase. We will conduct a randomized controlled field trial study for testing the effectiveness of adaptive sequencing of activities compared to static sequences defined in advance by teachers. Students across a range of different types of secondary schools will randomly be assigned to either the intervention group (adaptive sequences) or control group (static sequences). By employing a pre-post test design, we will be able to associate learning gains with experimental conditions and to test for which types of schools and learning goals adaptivity makes a difference.

Our most immediate goal at this point is to further develop both the knowledge hierarchy and the annotation rules. The sequencing algorithm requires a rich linguistic characterization and explicit interrelationships between specific language means. For instance, conditional type 2 sentences cannot be practiced if past simple forms and conditional forms have not been learned. Additionally, information from the expert model determining priorities between specific linguistic structures at a given point in the instruction plan can be used if more than one linguistic structure competes to be the “next” one.

In the mid-term, the creation of a gold-standard to evaluate the quality of the annotation process is also a task that we cannot escape. Since we have access to activities from other e-learning platforms, we can use those to perform a semi-automatic evaluation of the module. An evaluation from the perspective of the end-user in terms of the system’s efficacy will be possible as soon as the system starts to be piloted in schools. For that purpose we will simulate learning paths that will then end up proposing “next tasks” that a teacher will then judge as pedagogically meaningful or not.

Acknowledgments

This research has been largely funded by the Robert-Bosch-Stiftung GmbH as part of the program “Förderung von Wissenschaft-Praxis-Kooperationen für Unterrichtskonzepte mit digitalen Medien”.

References

Luiz Amaral. 2007. *Designing Intelligent Language Tutoring Systems: integrating Natural Language Processing technology into foreign language teaching*. Ph.D. thesis, The Ohio State University.

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

Lyle F. Bachman and Adrian S. Palmer. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press.

H. Douglas Brown. 2007. *Principles Of Language Learning and Teaching*, 5th edition. Pearson Education.

Susan Bull and Judy Kay. 2006. *Student models that invite the learner in: The SMILI open learner modelling framework*. Citeseer.

Maria Chinkina and Detmar Meurers. 2016. Linguistically-aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 188–198, San Diego, CA. ACL.

Inn-Chull Choi. 2016. Efficacy of an ICALL tutoring system and process-oriented corrective feedback. *Computer Assisted Language Learning*, 29(2):334–364.

Jinho D Choi and Martha Palmer. 2012. Fast and robust part-of-speech tagging using dynamic model selection. In *Proceedings of the 50th Annual Meeting of the ACL*, pages 363–367.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion Volume*. Cambridge University Press, Cambridge.

Zoltán Dörnyei and Peter Skehan. 2003. *Individual Differences in Second Language Learning*, chapter 18. John Wiley & Sons, Ltd.

Sheila Estaire and Javier Zanón. 1994. *Planning classwork: A task-based approach*. Educational Language Teaching. MacMillan-Heinemann, Oxford.

David Ferrucci and Adam Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.

S.M. Gass and A. Mackey. 2013. *The Routledge Handbook of Second Language Acquisition*. Routledge Handbooks in Applied Linguistics. Taylor & Francis.

Susan M. Gass and Alison Mackey. 2015. Input, interaction and output in second language acquisition. In Bill VanPatten and Jessica Williams, editors, *Theories in Second Language Acquisition: An Introduction (2nd edition)*. Routledge, New York and London.

Trude Heift. 2010. Prompting in CALL: A longitudinal study of learner uptake. *Modern Language Journal*, 94(2):198–216.

Peter Kluegl, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. 2016. Uima ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22(1):1–40.

Kultusministerium. 2016. Englisch als erste Fremdsprache [English as a first foreign language]. Bildungsplan des Gymnasiums 2016 [State curriculum for academic track schools 2016]. Ministerium für Kultus, Jugend und Sport, Baden Württemberg.

Shawn Loewen and Masatoshi Sato. 2017. *The Routledge handbook of instructed second language acquisition*. Routledge New York.

Wenting Ma, Olusola O. Adesope, John C. Nesbit, and Qing Liu. 2014. Intelligent tutoring systemsand learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4):901–918.

Melinda Martin-Beltrán, Natalia L. Guzman, and Pei-Jie Jenny Chen. 2017. 'let's think about it together:' how teachers differentiate discourse to mediate collaboration among linguistically diverse students. *Language Awareness*, 26(1):41–58.

Detmar Meurers, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2019. Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, 39:161–188.

Detmar Meurers, Kordula De Kuthy, Verena Möller, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2018. Digitale Differenzierung benötigt Informationen zu Sprache, Aufgabe und Lerner. Zur Generierung von individuellem Feedback in einem interaktiven Arbeitsheft. *FLuL – Fremdsprachen Lehren und Lernen*, 47(2):64–82.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 10–18, Los Angeles. ACL.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–233.

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

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, and Detmar Meurers. 2017. Developing a web-based workbook for English supporting the interaction of students and teachers. In *Proceedings of the Joint*

6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition, pages 36–46.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. 2018. Generating feedback for English foreign language exercises. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 127–136. ACL.

Helmut Schmid. 2005. A programming language for finite state transducers. In *Proceedings of the 5th International Workshop on Finite State Methods in Natural Language Processing*, pages 308–309.

Carol Ann Tomlinson. 2015. Teaching for excellence in academically diverse classrooms. *Society*, 52(3):203–209.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.

Ramon Ziai, Florian Nuxoll, Kordula De Kuthy, Björn Rudzewitz, and Detmar Meurers. 2019. The impact of spelling correction and task context on short answer assessment for intelligent tutoring systems. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 93–99, Turku, Finland. ACL.

Ramon Ziai, Björn Rudzewitz, Kordula De Kuthy, Florian Nuxoll, and Detmar Meurers. 2018. Feedback strategies for form and meaning in a real-life language tutoring system. In *Proceedings of the 7th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)*, pages 91–98. ACL.