

# Enhancing Authentic Web Pages for Language Learners

Detmar Meurers<sup>1</sup>, Ramon Ziai<sup>1</sup>,  
Luiz Amaral<sup>2</sup>, Adriane Boyd<sup>3</sup>, Aleksandar Dimitrov<sup>1</sup>, Vanessa Metcalf<sup>3</sup>, Niels Ott<sup>1</sup>

<sup>1</sup> Universität Tübingen

<sup>2</sup> University of Massachusetts Amherst

<sup>3</sup> The Ohio State University

## Abstract

Second language acquisition research since the 90s has emphasized the importance of supporting awareness of language categories and forms, and input enhancement techniques have been proposed to make target language features more salient for the learner.

We present an NLP architecture and web-based implementation providing automatic visual input enhancement for web pages. Learners freely choose the web pages they want to read and the system displays an enhanced version of the pages. The current system supports visual input enhancement for several language patterns known to be problematic for English language learners, as well as fill-in-the-blank and clickable versions of such pages supporting some learner interaction.

## 1 Introduction

A significant body of research into the effectiveness of meaning-focused communicative approaches to foreign language teaching has shown that input alone is not sufficient to acquire a foreign language, especially for older learners (cf., e.g., Lightbown and Spada, 1999). Recognizing the important role of consciousness in second-language learning (Schmidt, 1990), learners have been argued to benefit from (Long, 1991) or even require (Lightbown, 1998) a so-called *focus on form* to overcome incomplete or incorrect knowledge of specific forms or regularities. Focus on form is understood to be “an occasional shift of attention to linguistic code features” (Long and Robinson, 1998, p. 23).

In an effort to combine communicative and structuralist approaches to second language teaching, Rutherford and Sharwood Smith (1985) argued for the use of consciousness raising strategies drawing the learner’s attention to specific language properties. Sharwood Smith (1993, p. 176) coined the term *input enhancement* to refer to strategies highlighting the salience of language categories and forms.

Building on this foundational research in second language acquisition and foreign language teaching, in this paper we present an NLP architecture and a system for automatic visual input enhancement of web pages freely selected by language learners. We focus on learners of English as a Second Language (ESL), and the language patterns enhanced by the system include some of the well-established difficulties: determiners and prepositions, the distinction between gerunds and *to*-infinitives, *wh*-question formation, tense in conditionals, and phrasal verbs.

In our approach, learners can choose any web page they like, either by using an ordinary search-engine interface to search for one or by entering the URL of the page they want to enhance. In contrast to textbooks and other pre-prepared materials, allowing the learner to choose up-to-date web pages on any topic they are interested in and enhancing the page while keeping it intact (with its links, multimedia, and other components working) clearly has a positive effect on learner motivation. Input enhanced web pages also are attractive for people outside a traditional school setting, such as in the voluntary, self-motivated pursuit of knowledge often referred to as lifelong learning. The latter can be particularly relevant for adult immigrants, who are

already functionally living in the second language environment, but often stagnate in their second language acquisition and lack access or motivation to engage in language classes or other explicit language learning activities. Nevertheless, they do use the web to obtain information that is language-based and thus can be enhanced to also support language acquisition while satisfying information needs.

In terms of paper organization, in section 2 we first present the system architecture and in 2.1 the language phenomena handled, before considering the issues involved in evaluating the approach in 2.2. The context of our work and related approaches are discussed in section 3, and we conclude and discuss several avenues for future research in section 4.

## 2 The Approach

The WERTi system (*Working with English Real Texts interactively*) we developed follows a client-server paradigm where the server is responsible for fetching the web page and enriching it with annotations, and the client then receives the annotated web page and transforms it into an enhanced version. The client here is a standard web browser, so on the learner's side no additional software is needed.

The system currently supports three types of input enhancement: i) color highlighting of the pattern or selected parts thereof, ii) a version of the page supporting identification of the pattern through clicking and automatic color feedback, and iii) a version supporting practice, such as a fill-in-the-blank version of the page with automatic color feedback.

The overall architecture is shown in Figure 1. Essentially, the automated input enhancement process consists of the following steps:

1. Fetch the page.
2. Find the natural language text portions in it.
3. Identify the targeted language pattern.
4. Annotate the web page, marking up the language patterns identified in the previous step.
5. Transform the annotated web page into the output by visually enhancing the targeted pattern or by generating interaction possibilities.

Steps 1–4 take place on the server side, whereas step 5 happens in the learner's browser.<sup>1</sup> As NLP is only involved in step 3, we here focus on that step.

<sup>1</sup>As an alternative to the server-based fetching of web pages,

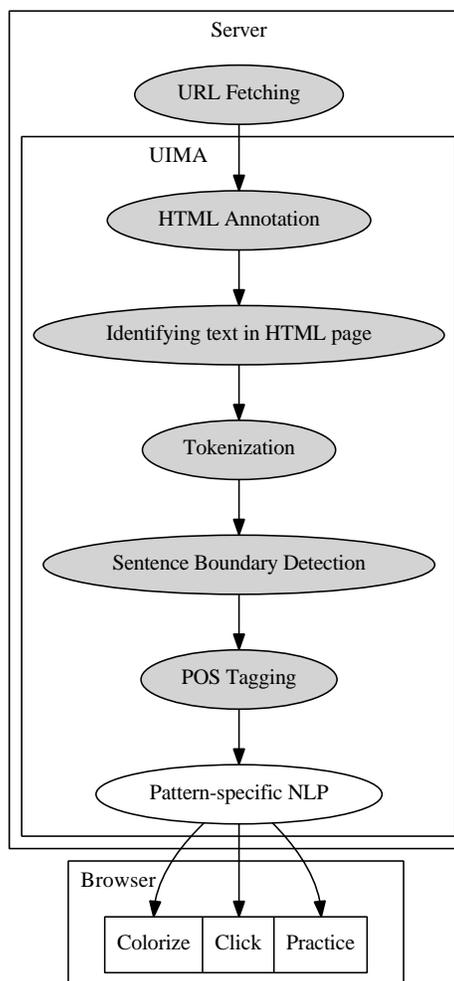


Figure 1: Overall WERTi architecture. Grey components are the same for all patterns and activities, cf. section 2.1.

While the first prototype of the WERTi system<sup>2</sup> presented at CALICO (Amaral, Metcalf and Meurers, 2006) and EUROCALL (Metcalf and Meurers, 2006) was implemented in Python, the current system is Java-based, with all NLP being integrated in the UIMA framework (Ferrucci and Lally, 2004). UIMA is an architecture for the management and analysis of unstructured information such as text, which is built on the idea of referential annotation and can be seen as an NLP analysis counterpart to current stand-off encoding standards for annotated corpora (cf., e.g., Ide et al. 2000). The input

we are developing a Firefox plugin, leaving only the NLP up to the server. This increases compatibility with web pages using dynamically generated contents and special session handling.

<sup>2</sup><http://purl.org/icall/werti-v1>

can be monotonically enriched while passing from one NLP component to the next, using a flexible data repository common to all components (Götz and Suhre, 2004). Such annotation-based processing is particularly useful in the WERTi context, where keeping the original text intact is essential for displaying it in enhanced form.

A second benefit of using the UIMA framework is that it supports a flexible combination of individual NLP components into larger processing pipelines. To obtain a flexible approach to input enhancement in WERTi, we need to be able to identify and analyze phenomena from different levels of linguistic analysis. For example, lexical classes can be identified by a POS tagger, whereas other patterns to be enhanced require at least shallow syntactic chunking. The more diverse the set of phenomena, the less feasible it is to handle all of them within a single processing strategy or formalism. Using the UIMA framework, we can re-use the same basic processing (e.g., tokenizing, POS tagging) for all phenomena and still be able to branch into pattern-specific NLP in a demand-driven way. Given that NLP components in UIMA include self-describing meta-information, the processing pipeline to be run can dynamically be obtained from the module configuration instead of being hard-wired into the core system. The resulting extensible, plugin-like architecture seems particularly well-suited for the task of visual input enhancement of a wide range of heterogeneous language properties.

Complementing the above arguments for the UIMA-based architecture of the current WERTi system, a detailed discussion of the advantages of an annotation-based, demand-driven NLP architecture for Intelligent Computer-Assisted Language Learning can be found in Amaral, Meurers, and Ziai (To Appear), where it is employed in an Intelligent Language Tutoring System.

## 2.1 Implemented Modules

The modules implemented in the current system handle a number of phenomena commonly judged as difficult for second language learners of English. In the following we briefly characterize each module, describing the nature of the language pattern, the required NLP, and the input enhancement results, which will be referred to as *activities*.

## Lexical classes

Lexical classes are the most basic kind of linguistic category we use for input enhancement. The inventory of lexical categories to be used and which ones to focus on should be informed by second language acquisition research and foreign language teaching needs. The current system focuses on functional elements such as prepositions and determiners given that they are considered to be particularly difficult for learners of English (cf. De Felice, 2008 and references therein).

We identify these functional elements using the LingPipe POS tagger (<http://alias-i.com/lingpipe>) employing the Brown tagset (Francis and Kucera, 1979). As we show in section 2.2, the tagger reliably identifies prepositions and determiners in native English texts such as those expected for input enhancement.

The input enhancement used for lexical classes is the default set of activities provided by WERTi. In the simplest case, *Color*, all automatically identified instances in the web page are highlighted by coloring them; no learner interaction is required. This is illustrated by Figure 2, which shows the result of enhancing prepositions in a web page from the British

**Car-free cities: an idea **with** legs**  
 Car-free neighbourhoods are no unrealistic utopia – they exist all **over** Europe

Comments (...)  
 Buzz up!  
 Digg it

Posted **by** Steve Melia Thursday 29 October 2009 08:00 GMT guardian.co.uk

**A** larger | smaller

**Life and style**  
 Cycling

**Environment**  
 Ethical and green living &mdash; Travel and transport

**Series**  
 Bike blog

**More from Green living blog on**

**Life and style**  
 Cycling

**Environment**  
 Ethical and green living &mdash; Travel and

'Not anti-car, just pro-choice' ... a cyclist **in** Vauban, Germany. Photograph: Sipa Press/Rex Features

A quarter **of** households **in** Britain – more **in** the larger cities, and a majority **in** some inner cities – live **without** a car. Imagine how quality **of** life would improve **for** cyclists and everyone else if traffic were removed **from** areas where people could practically choose to live **without** cars. Does this sound unrealistic, utopian? Did you know many European cities are already doing it?

**Vauban in Germany** is one **of** the largest car-free neighbourhoods **in** Europe, home **to** more than 5,000 people. If you live **in** the district, you are required to confirm once a year that you do not own a car – or, if you do own one, you must buy a space **in** a multi-storey car park **on** the edge **of** the district. One space was initially provided **for** every two households, but car ownership has fallen **over** time, and many **of** these spaces are

Figure 2: Screenshot of color activity for prepositions, cf. <http://purl.org/icall/werti-color-ex>

newspaper The Guardian.<sup>3</sup>

In this and the following screenshots, links already present in the original web page appear in light blue (e.g., *Vauban in Germany*). This raises an important issue for future research, namely how to determine the best visual input enhancement for a particular linguistic pattern given a specific web page with its existing visual design features (e.g., bold-facing in the text or particular colors used to indicate links), which includes the option of removing or altering some of those original visual design features.

A more interactive activity type is *Click*, where the learner during reading can attempt to identify instances of the targeted language form by clicking on it. Correctly identified instances are colored green by the system, incorrect guesses red.

Thirdly, input can be turned into *Practice* activities, where in its simplest form, WERTi turns web pages into fill-in-the-blank activities and provides immediate color coded feedback for the forms entered by the learner. The system currently accepts only the form used in the original text as correct. In principle, alternatives (e.g., other prepositions) can also be grammatical and appropriate. The question for which cases equivalence classes of target answers can automatically be determined is an interesting question for future research.<sup>4</sup>

### Gerunds vs. *to*-infinitives

Deciding when a verb is required to be realized as a *to*-infinitive and when as a gerund *-ing* form can be difficult for ESL learners. Current school grammars teach students to look for certain lexical clues that reliably indicate which form to choose. Examples of such clues are prepositions such as *after* and *of*, which can only be followed by a gerund.

In our NLP approach to this language pattern, we use Constraint Grammar rules (Karlsson et al., 1995) on top of POS tagging, which allow for straightforward formulation of local disambiguation rules such as: “If an *-ing* form immediately follows the preposition *by*, select the gerund reading.” Standard POS

<sup>3</sup>Given the nature of the input enhancement using colors, the highlighting in the figure is only visible in a color printout.

<sup>4</sup>The issue bears some resemblance to the task of identifying paraphrases (Androutsopoulos and Malakasiotis, 2009) or classes of learner answers which differ in form but are equivalent in terms of meaning (Bailey and Meurers, 2008).

tags for English contain a single tag for all *-ing* forms. In order to identify gerunds only, we introduce all possible readings for all *-ing* forms and wrote 101 CG rules to locally disambiguate them. The *to*-infinitives, on the other hand, are relatively easy to identify based on the surface form and require almost no disambiguation.

For the implementation of the Constraint Grammar rules, we used the freely available CG3 system.<sup>5</sup> While simple local disambiguation rules are sufficient for the pattern discussed here, through iterative application of rules, Constraint Grammar can identify a wide range of phenomena without the need to provide a full grammatical analysis.

The *Color* activity resulting from input enhancement is similar to that for lexical classes described above, but the system here enhances both verb forms and clue phrases. Figure 3 shows the system highlighting gerunds in orange, infinitives in purple, and clue phrases in blue.

"The government says it is expanding access to university, but they are actually blocking people's aspirations and betraying a generation."

The government was forced to cap student numbers after discovering a £200m black hole in the university financing budget at the end of last year. Labour was accused of abandoning its pledge to expand higher education, adding pressure to a growing debate about how to fund the growing number of young people who want to do a degree. The government is due to announce a review of student finance.

The massive increase in applicants has put a strain on the university system this year, with one university forced to convert single bedrooms in halls into doubles, and others putting students up in hotels.

Figure 3: Color activity for gerunds vs. *to*-infinitives, cf. <http://purl.org/ical1/werti-color-ex2>

For the *Click* activity, the web page is shown with colored gerund and *to*-infinitival forms and the learner can click on the corresponding clue phrases.

For the *Practice* activity, the learner is presented with a fill-in-the-black version of the web page, as in the screenshot in Figure 4. For each blank, the learner needs to enter the gerund or *to*-infinitival form of the base form shown in parentheses.

### Wh-questions

Question formation in English, with its particular word order, constitutes a well-known challenge for second language learners and has received significant attention in the second language acquisi-

<sup>5</sup><http://beta.vis1.sdu.dk/cg3.html>

"The government says it is expanding access to university, but they are actually blocking people's aspirations and betraying a generation."

The government was forced to cap student numbers after  (discover) a £200m black hole in the university financing budget at the end of last year. Labour was accused of  (abandon) its pledge to expand higher education, adding pressure to a growing debate about how to fund the growing number of young people who want  (do) a degree. The government is due to announce a review of student finance.

The massive increase in applicants has put a strain on the university system this year, with one university forced to convert single bedrooms in halls into doubles, and others putting students up in hotels.

Figure 4: Practice activity for gerunds vs. *to*-infinitives, cf. <http://purl.org/icall/werti-cloze-ex>

tion literature (cf., e.g., White et al., 1991; Spada and Lightbown, 1993). Example (1) illustrates the use of *do*-support and subject-aux inversion in *wh*-questions as two aspects challenging learners.

(1) *What do you think* it takes to be successful?

In order to identify the *wh*-question patterns, we employ a set of 126 hand-written Constraint Grammar rules. The respective *wh*-word acts as the lexical clue to the question as a whole, and the rules then identify the subject and verb phrase based on the POS and lexical information of the local context.

Aside from the *Color* activity highlighting the relevant parts of a *wh*-question, we adapted the other activity types to this more complex language pattern. The *Click* activity prompts learners to click on either the subject or the verb phrase of the question. The *Practice* activity presents the words of a *wh*-question in random order and requires the learner to rearrange them into the correct one.

### Conditionals

English has five types of conditionals that are used for discussing hypothetical situations and possible outcomes. The tenses used in the different conditional types vary with respect to the certainty of the outcome as expressed by the speaker/writer. For example, one class of conditionals expresses high certainty and uses present tense in the *if*-clause and future in the main clause, as in example (2).

(2) *If* the rain *continues*, we *will* return home.

The recognition of conditionals is approached using a combination of shallow and deep methods. We first look for lexical triggers of a conditional, such as

the word *if* at the beginning of a sentence. This first pass serves as a filter to the next, more expensive processing step, full parsing of the candidate sentences using Bikel's statistical parser (Bikel, 2002). The parse trees are then traversed to identify and mark the verb forms and the trigger word.

For the input enhancement, we color all relevant parts of a conditional, namely the trigger and the verb forms. The *Click* activity for conditionals requires the learner to click on exactly these parts. The *Practice* activity prompts users to classify the conditional instances into the different classes.

### Phrasal verbs

Another challenging pattern for English language learners are phrasal verbs consisting of a verb and either a preposition, an adverb or both. The meaning of a phrasal verb often differs considerably from that of the underlying verb, as in (3) compared to (4).

(3) He *switched* the glasses without her noticing.

(4) He *switched off* the light before he went to bed.

This distinction is difficult for ESL learners, who often confuse phrasal and non-phrasal uses.

Since this is a lexical phenomenon, we approached the identification of phrasal verbs via a database lookup in a large online collection of verbs known to occur in phrasal form.<sup>6</sup> In order to find out about noun phrases and modifying adverbs possibly occurring in between the verb and its particles, we run a chunker and use this information in specifying a filter for such intervening elements.

The visual input enhancement activities targeting phrasal verbs are the same as for lexical classes, with the difference that for the *Practice* activity, learners have to fill in only the particle, not the particle and the main verb, since otherwise the missing contents may be too difficult to reconstruct. Moreover, we want the activity to focus on distinguishing phrasal from non-phrasal uses, not verb meaning in general.

## 2.2 Evaluation issues

The success of a visual input enhancement approach such as the one presented in this paper depends on a number of factors, each of which can in principle

<sup>6</sup><http://www.usingenglish.com/reference/phrasal-verbs>

be evaluated. The fundamental but as far as we are aware unanswered question in second language acquisition research is for which language categories, forms, and patterns input enhancement can be effective. As Lee and Huang (2008) show, the study of visual input enhancement sorely needs more experimental studies. With the help of the WERTi system, which systematically produces visual input enhancement for a range of language properties, it becomes possible to conduct experiments in a real-life foreign language teaching setting to test learning outcomes<sup>7</sup> with and without visual input enhancement under a wide range of parameters. Relevant parameters include the linguistic nature of the language property to be enhanced as well as the nature of the input enhancement to be used, be it highlighting through colors or fonts, engagement in different types of activities such as clicking, entering fill-in-the-blank information, reordering language material, etc.

A factor closely related to our focus in this paper is the impact of the quality of the NLP analysis.<sup>8</sup> For a quantitative evaluation of the NLP, one significant problem is the mismatch between the phenomena focused on in second language learning and the available gold standards where these phenomena are actually annotated. For example, standard corpora such as the Penn Treebank contain almost no questions and thus do not constitute a useful gold standard for *wh*-question identification. Another problem is that some grammatical distinctions taught to language learners are disputed in the linguistic literature. For example, Huddleston and Pullum (2002, p. 1120) eliminate the distinction between gerunds and present participles, combining them into a class called “gerund-participle”. And in corpus annotation practice, gerunds are not identified as a class by the tagsets used to annotate large corpora, making it unclear what gold standard our gerund identification component should be evaluated against.

While the lack of available gold standards means that a quantitative evaluation of all WERTi modules is beyond the scope of this paper, the determiner and preposition classes focused on in the lexical classes module can be identified using the stan-

<sup>7</sup>Naturally, online measures of noticing, such as eye tracking or Event-Related Potentials (ERP) would also be relevant.

<sup>8</sup>The processing time for the NLP analysis as other relevant aspect is negligible for most of the activities presented here.

dard CLAWS-7 or Brown tagsets, for which gold-standard corpora are available. We thus decided to evaluate this WERTi module against the BNC Sampler Corpus (Burnard, 1999), which contains a variety of genres, making it particularly appropriate for evaluating a tool such as WERTi, which learners are expected to use with a wide range of web pages as input. The BNC Sampler corpus is annotated with the fine-grained CLAWS-7 tagset<sup>9</sup> where, e.g., prepositions are distinguished from subordinating conjunctions. By mapping the relevant POS tags from the CLAWS-7 tagset to the Brown tagset used by the LingPipe tagger as integrated in WERTi, it becomes possible to evaluate WERTi’s performance for the specific lexical classes focused on for input enhancement, prepositions and determiners. For prepositions, precision was 95.07% and recall 90.52% while for determiners, precision was 97.06% with a recall of 94.07%.

The performance of the POS tagger on this reference corpus thus seems to be sufficient as basis for visual input enhancement, but the crucial question naturally remains whether identification of the target patterns is reliable in the web pages that language learners happen to choose. For a more precise quantitative study, it will thus be important to try the system out with real-life users in order to identify a set of web pages which can constitute an adequate test set. Interestingly, which web pages the users choose depends on the search engine front-end we provide for them. As discussed under outlook in section 4, we are exploring the option to implicitly guide them towards web pages containing enough instances of the relevant language patterns in text at the appropriate reading difficulty.

### 3 Context and related work

Contextualizing our work, one can view the automatic visual input enhancement approach presented here as an enrichment of Data-Driven Learning (DDL). Where DDL has been characterized as an “attempt to cut out the middleman [the teacher] as far as possible and to give the learner direct access to the data” (Boulton 2009, p. 82, citing Tim Johns), in visual input enhancement the learner stays in con-

<sup>9</sup><http://www.natcorp.ox.ac.uk/docs/c7spec.html>

trol, but the NLP uses ‘teacher knowledge’ about relevant and difficult language properties to make those more prominent and noticeable for the learner.

In the context of Intelligent Computer-Assisted Language Learning (ICALL), NLP has received most attention in connection with Intelligent Language Tutoring Systems, where NLP is used to analyze learner data and provide individual feedback on that basis (cf. Heift and Schulze, 2007). Demands on such NLP are high given that it needs to be able to handle learner language and provide high-quality feedback for any sentence entered by the learner.

In contrast, visual input enhancement makes use of NLP analysis of authentic, native-speaker text and thus applies the tools to the native language they were originally designed and optimized for. Such NLP use, which we will refer to as Authentic Text ICALL (ATICALL), also does not need to be able to correctly identify and manipulate all instances of a language pattern for which input enhancement is intended. Success can be incremental in the sense that any visual input enhancement can be beneficial, so that one can focus on enhancing those instances which can be reliably identified in a text. In other words, for ATICALL, precision of the NLP tools is more important than recall. It is not necessary to identify and enhance all instances of a given pattern as long as the instances we do identify are in fact correct, i.e., true positives. As the point of our system is to enhance the reading experience by raising language awareness, pattern occurrences we do not identify are not harmful to the overall goal.<sup>10</sup>

We next turn to a discussion of some interesting approaches in two closely related fields, exercise generation and reading support tools.

### 3.1 Exercise Generation

Exercise generation is widely studied in CALL research and some of the work relates directly to the input enhancement approach presented in this paper. For instance, Antoniadis et al. (2004) describe the plans of the MIRTO project to support “gap-filling” and “lexical spotting” exercises in combination with a corpus database. However, MIRTO seems to fo-

<sup>10</sup>While identifying all instances of a pattern indeed is not crucial in this context, representativeness remains relevant to some degree. Where only a skewed subset of a pattern is highlighted, learners may not properly conceptualize the pattern.

cus on a general architecture supporting instructor-determined activity design. Visual input enhancement or language awareness are not mentioned. The VISL project (Bick, 2005) offers games and visual presentations in order to foster knowledge of syntactic forms and rules, and its KillerFiller tool can create slot-filler exercises from texts. However, KillerFiller uses corpora and databases as the text base and it presents sentences in isolation in a testing setup. In contrast to such exercise generation systems, we aim at enhancing the reader’s second language input using the described web-based mash-up approach.

### 3.2 Reading Support Tools

Another branch of related approaches consists of tools supporting the reading of texts in a foreign language. For example, the Glosser-RuG project (Nerbonne et al., 1998) supports reading of French texts for Dutch learners with an online, context-dependent dictionary, as well as morphological analysis and examples of word use in corpora. A similar system, focusing on multi-word lexemes, was developed in the COMPASS project (Breidt and Feldweg, 1997). More recently, the ALPHEIOS project<sup>11</sup> has produced a system that can look up words in a lexicon and provide aligned translations. While such lexicon-based tools are certainly useful to learners, they rely on the learner asking for help instead of enhancing specific structures from the start and thus clearly differ from our approach.

Finally, the REAP project<sup>12</sup> supports learners in searching for texts that are well-suited for providing vocabulary and reading practice (Heilman et al., 2008). While it differs in focus from the visual input enhancement paradigm underlying our approach, it shares with it the emphasis on providing the learner with authentic text in support of language learning.

## 4 Conclusion and Outlook

In this paper we presented an NLP architecture and a concrete system for the enhancement of authentic web pages in order to support language awareness in ESL learners. The NLP architecture is flexible enough to integrate any processing approach that lends itself to the treatment of the language phe-

<sup>11</sup><http://alpheios.net>

<sup>12</sup><http://reap.cs.cmu.edu>

nomenon in question, without confining the developer to a particular formalism. The WERTi system illustrates this with five language patterns typically considered difficult for ESL learners: lexical classes, gerunds vs. *to*-infinitives, *wh*-questions, conditionals and phrasal verbs.

Looking ahead, we already mentioned the fundamental open question where input enhancement can be effective in section 2.2. A system such as WERTi, systematically producing visual input enhancement, can help explore this question under a wide range of parameters in a real-life language teaching setting. A more specific future research issue is the automatic computation of equivalence classes of target forms sketched in section 2.1. Not yet mentioned but readily apparent is the goal to integrate more language patterns known to be difficult for language learners into WERTi (e.g., active/passive, tense and aspect distinctions, relative clauses), and to explore the approach for other languages, such as German.

A final important avenue for future research concerns the starting point of the system, the step where learners search for a web page they are interested in and select it for presentation with input enhancement. Enhancing of patterns presupposes that the pages contain instances of the pattern. The less frequent the pattern, the less likely we are to find enough instances of it in web pages returned by the standard web search engines typically used by learners to find pages of interest to them. The issue is related to research on providing learners with texts at the right level of reading difficulty (Petersen, 2007; Miltsakaki and Troutt, 2008), but the focus for us is on ensuring that texts which include instances of the specific language pattern targeted by a given input enhancement are ranked high in the search results. Ott (2009) presents a search engine prototype which, in addition to the content-focused document-term information and traditional readability measures, supports indexing based on a more general notion of a text model into which the patterns relevant to input enhancement can be integrated – an idea we are exploring further (Ott and Meurers, Submitted).

## Acknowledgments

We benefited from the feedback we received at CALICO 06, EUROCALL 06, and the ICALL

course<sup>13</sup> at ESSLLI 09, where we discussed our work on the Python-based WERTi prototype. We would like to thank Chris Hill and Kathy Corl for their enthusiasm and encouragement. We are grateful to Magdalena Leshtanska, Emma Li, Iliana Simova, Maria Tchalakova and Tatiana Vodolazova for their good ideas and WERTi module contributions in the context of a seminar at the University of Tübingen in Summer 2008. Last but not least, the paper benefited from two helpful workshop reviews.

## References

- Luiz Amaral, Vanessa Metcalf, and Detmar Meurers. 2006. Language awareness through re-use of NLP technology. Presentation at the CALICO Workshop on *NLP in CALL – Computational and Linguistic Challenges*, May 17, 2006. University of Hawaii. <http://purl.org/dm/handouts/calico06-amaral-metcalf-meurers.pdf>.
- Luiz Amaral, Detmar Meurers, and Ramon Ziai. To Appear. Analyzing learner language: Towards a flexible NLP architecture for intelligent language tutors. *Computer-Assisted Language Learning*. <http://purl.org/dm/papers/amaral-meurers-ziai-10.html>.
- Ion Androutsopoulos and Prodrimos Malakasiotis. 2009. A survey of paraphrasing and textual entailment methods. Technical report, NLP Group, Informatics Dept., Athens University of Economics and Business, Greece. <http://arxiv.org/abs/0912.3747>.
- Georges Antoniadis, Sandra Echinard, Olivier Kraif, Thomas Lebarbé, Mathieu Loiseau, and Claude Ponton. 2004. NLP-based scripting for CALL activities. In *Proceedings of the COLING Workshop on eLearning for CL and CL for eLearning*, Geneva.
- Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In (Tetreault et al., 2008), pages 107–115.
- Eckhard Bick. 2005. Grammar for fun: IT-based grammar learning with VISL. In P. Juel, editor, *CALL for the Nordic Languages*, pages 49–64. Samfundslitteratur, Copenhagen.
- Daniel M. Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of the Second Int. Conference on Human Language Technology Research*, San Francisco.
- Alex Boulton. 2009. Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*, 35(1):81–106.

<sup>13</sup><http://purl.org/dm/09/esslli/>

- Elisabeth Breidt and Helmut Feldweg. 1997. Accessing foreign languages with COMPASS. *Machine Translation*, 12(1–2):153–174.
- L. Burnard, 1999. *Users Reference Guide for the BNC Sampler*. Available on the BNC Sampler CD.
- Rachele De Felice. 2008. *Automatic Error Detection in Non-native English*. Ph.D. thesis, St Catherine's College, University of Oxford.
- Catherine Doughty and J. Williams, editors. 1998. *Focus on form in classroom second language acquisition*. Cambridge University Press, Cambridge.
- 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.
- W. Nelson Francis and Henry Kucera, 1979. *Brown corpus manual*. Dept. of Linguistics, Brown University.
- Thilo Götz and Oliver Suhre. 2004. Design and implementation of the UIMA Common Analysis System. *IBM Systems Journal*, 43(3):476–489.
- Trude Heift and Mathias Schulze. 2007. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.
- Michael Heilman, Le Zhao, Juan Pino, and Maxine Eskenazi. 2008. Retrieval of reading materials for vocabulary and reading practice. In (Tetreault et al., 2008), pages 80–88.
- Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.
- Nancy Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the 2nd Int. Conference on Language Resources and Evaluation*.
- Tim Johns. 1994. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin, editor, *Perspectives on Pedagogical Grammar*, pages 293–313. CUP, Cambridge.
- Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, editors. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin, New York.
- Sang-Ki Lee and Hung-Tzu Huang. 2008. Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition*, 30:307–331.
- Patsy M. Lightbown and Nina Spada. 1999. *How languages are learned*. Oxford University Press, Oxford.
- Patsy M. Lightbown. 1998. The importance of timing in focus on form. In (Doughty and Williams, 1998), pages 177–196.
- Michael H. Long and Peter Robinson. 1998. Focus on form: Theory, research, and practice. In (Doughty and Williams, 1998), pages 15–41.
- M. H. Long. 1991. Focus on form: A design feature in language teaching methodology. In K. De Bot, C. Kramsch, and R. Ginsberg, editors, *Foreign language research in cross-cultural perspective*, pages 39–52. John Benjamins, Amsterdam.
- Vanessa Metcalf and Detmar Meurers. 2006. Generating web-based English preposition exercises from real-world texts. Presentation at EUROCALL, Sept. 7, 2006. Granada, Spain. <http://purl.org/dm/handouts/eurocall106-metcalf-meurers.pdf>.
- Eleni Miltsakaki and Audrey Troutt. 2008. Real time web text classification and analysis of reading difficulty. In (Tetreault et al., 2008), pages 89–97.
- John Nerbonne, Duco Dokter, and Petra Smit. 1998. Morphological processing and computer-assisted language learning. *Computer Assisted Language Learning*, 11(5):543–559.
- Niels Ott and Detmar Meurers. Submitted. Information retrieval for education: Making search engines language aware. <http://purl.org/dm/papers/ott-meurers-10.html>.
- Niels Ott. 2009. Information retrieval for language learning: An exploration of text difficulty measures. Master's thesis, International Studies in Computational Linguistics, University of Tübingen.
- Sarah E. Petersen. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. Ph.D. thesis, University of Washington.
- William E. Rutherford and Michael Sharwood Smith. 1985. Consciousness-raising and universal grammar. *Applied Linguistics*, 6(2):274–282.
- Richard W. Schmidt. 1990. The role of consciousness in second language learning. *Applied Linguistics*, 11:206–226.
- Michael Sharwood Smith. 1993. Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition*, 15:165–179.
- Nina Spada and Patsy M. Lightbown. 1993. Instruction and the development of questions in L2 classrooms. *Studies in Second Language Acquisition*, 15:205–224.
- Joel Tetreault, Jill Burstein, and Rachele De Felice, editors. 2008. *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. ACL, Columbus, Ohio, June.
- Lydia White, Nina Spada, Patsy M. Lightbown, and Leila Ranta. 1991. Input enhancement and L2 question formation. *Applied Linguistics*, 12(4):416–432.