Introduction

- The use of NLP in ICALL has primarily centered on diagnosing learner errors and, more recently, testing and assessment.
- Idea: Explore how NLP technology can support other aspects of second language learning.
- Our specific focus: What can NLP contribute to awareness of language forms and rules, an important component of adult second language acquisition?
  - WERTi: Automatic generation of language awareness activities based on real-world texts.
  - IR4ALL: Retrieval of authentic texts at the appropriate level for language learners

Pedagogical grounding of our research

Awareness (Schmidt 1995):

- Noticing
  - “conscious registration of an event”
  - low level of awareness
  - implicit learning

E.g.: noticing that sometimes speakers of Spanish omit the subject pronoun

- Understanding
  - “recognition of a general principle, rule or pattern”
  - higher level of awareness
  - explicit learning
  - generalization can be internally generated or externally provided

E.g.: understanding that Spanish is a pro-drop language

⇒ Consequences:

- Learners have to be exposed to linguistic features to acquire them.
- Learners have to notice those features.
- Tools presenting such linguistic features in a contextualized way, allowing for student interaction, can be helpful.
Pedagogical grounding of our research

Linguistic information and how it is conveyed

- A wide range of linguistic features can be relevant for awareness, incl. morphological, syntactic, semantic, and pragmatic information (cf. Schmidt 1995, p. 30).

- Linguistic information can be conveyed to the learner
  - using explicit linguistic terminology/representations, e.g.:
    - parts of speech
    - verbal tense, mood and aspect
    - sentence classification
    - syntactic analyses (shown as trees or sentence diagrams)
  - using implicit presentation, e.g.:
    - coloring, underlining, moving, etc
    - pointing to correct or incorrect uses

⇒ Awareness activities can include both implicit and explicit presentation of linguistic features.

Modeling FLT practice

- A common pedagogical practice in FLT moves from target language presentation, to practice, on to production.

- Proposal: Create sequences of linguistic awareness activities following the initial stages of such a progression:
  I. Receptive presentation
  II. Productive presentation
  III. Controlled practice

- What makes this idea interesting?
  - NLP technology can identify certain relevant linguistic categories and forms in real-life texts.
  - The contents of these texts can be selected by the learners based on their interests.
  - The sentences turned into exercises can remain fully contextualized as part of the text selected by learner.
  - Automatic feedback for the activities is feasible since the original text is known.

The activity progression in WERTi

Using real world web-based texts (such as news articles) we provide a progression of activities:

**Step 1.** Receptive presentation
Ex. The system *colors* examples of targeted items.

**Step 2.** Productive presentation
Ex. The learner is asked to *find* and *mouse-click* all tokens of the targeted category. The system shows correct picks in green, incorrect ones in red.

**Step 3.** Controlled practice
Ex. The learner is asked to
  - *reorder* words/phrases given (scrambled) list
  - complete *fill-in-the-blank* (FIB) slots
  - created for tokens of targeted category
  - given some information, where needed (e.g., stems)

Examples and Target types

- **Examples:**
  - FIB Determiners
  - Colored Gerunds

- **Types of targets:**
  - Lexical targets:
    - prepositions
    - determiners
  - Lexical form targets with contextual triggers:
    - gerunds vs. *to*-infinitives
    - *if* conditionals
    - tense and aspect
  - Syntactic targets with discourse context triggers:
    - active vs. passive
What is involved in realizing such an approach?

- Two components can be distinguished:
  1. Obtaining and selecting appropriate texts:
     - Texts obtained through web search using terms provided by the language learner
       - restrict web to news sites (e.g., Reuters)
       - alternative: specific corpora
     - Texts could be filtered according to aspects relevant to language learning (text readability, frequency of relevant constructions, etc. → IR4LL discussion below)
  2. Identifying the targets in the selected texts and creating receptive and productive presentations, and controlled practice exercises using the texts.

- We illustrate the approach, focusing on the second component, by showcasing an activity progression targeting prepositions.

### Prototype realization

- Original prototype in Python, integrated into the Apache2 webserver using mod_python, including:
  - searching in the Reuters site providing news webpages
  - linguistic annotation using NLTK (Bird & Loper 2004), TreeTagger (Schmid 1994)
- Recently reimplemented as UIMA-based Java servlet on Tomcat server (Aleks Dimitrov, Ramon Ziai, Niels Ott).
- The annotated text is mapped into Color, Click, and FIB presentation code (HTML and JavaScript), and fully integrated in the original web page.
- Only a standard web browser is needed to use the system.
- We are working on integrating further target patterns and activities. Prototypes available at:
  - WERTi: http://purl.org/net/WERTi
  - WERTi2: http://delos.sfs.uni-tuebingen.de:8080/WERTi

### Realizing the proposal

#### Creating an activity sequence

- The system first annotates the web page text using efficient and robust NLP tools performing:
  - tokenization → tokens
  - lemmatization → word roots
  - part-of-speech tagging → lexical categories
  - morphological analysis → morphological properties
  - shallow parsing → phrasal categories

- The language items targeted by the activity are identified using regular expression matching of target and contextual items in the annotated text.

- The nature of the activity determines the complexity of the annotation and the regular expressions required:
  - Preposition activity: single instances of a lexical category
  - Tense and aspect: sequences of auxiliaries, inflected forms, and specific lexical items (contextual cues)

#### Prototype realization

- Recently reimplemented as UIMA-based Java servlet
  - IR4ICALL NLP pipeline
- The system first annotates the web page text using efficient and robust NLP tools performing:
  - tokenization → tokens
  - lemmatization → word roots
  - part-of-speech tagging → lexical categories
  - morphological analysis → morphological properties
  - shallow parsing → phrasal categories

#### Some challenges

- Annotation errors:
  - Statistical NLP tools are efficient and robust
  - Such tools make errors, e.g., 3–5% for POS tagging.
  - What impact do such errors have for the envisaged use?
  - It is known where errors are likely to arise (cf., e.g., Dickinson & Meurers 2003; Dickinson 2005), so one can avoid basing activities on likely error locations.

- The complexity of real life:
  - Real-life texts from the web often have
    - complex structure
    - mark-up and integrated multimedia
  - It is nontrivial to combine that web page structure with the activity based on the annotated text base.
Finding texts appropriate for language learners

- How can one find authentic texts as reading material or for activity generation (e.g., WERTi)?
- Such texts should
  - be in the language of interest
  - have the appropriate level of complexity for the learner
  - contain enough good instances of the language patterns and rules targeted by the activities.
- How about simply using the web and a standard web search engine (e.g., google)?
  - Pro: The Web is huge, and up-to-date information on virtually any topic is available.
  - Cons: Standard search engines are not aware of reading complexity and language patterns.
⇒ Create a dedicated search engine for language learning: IR4LL (Ott 2009)

Readability and how to measure it

- Readability or text difficulty: refers to the understandability or comprehensibility of a text (Klare 1963).
- The more reading proficient the reader, the less readable texts need to be in order to be understood by this reader.
- Traditional readability formulas try to measure the readability on a scale, e.g. the U.S. grade level scale.

U.S. grade level scale

Scale based on Gunning (1968, p. 40):

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>Named Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>College</td>
</tr>
<tr>
<td>16</td>
<td>graduate</td>
</tr>
<tr>
<td>15</td>
<td>senior</td>
</tr>
<tr>
<td>14</td>
<td>junior</td>
</tr>
<tr>
<td>13</td>
<td>sophomore</td>
</tr>
<tr>
<td>12</td>
<td>High School</td>
</tr>
<tr>
<td>11</td>
<td>senior</td>
</tr>
<tr>
<td>10</td>
<td>junior</td>
</tr>
<tr>
<td>9</td>
<td>sophomore</td>
</tr>
<tr>
<td>8</td>
<td>Eight grade</td>
</tr>
<tr>
<td>7</td>
<td>Seventh grade</td>
</tr>
<tr>
<td>6</td>
<td>Sixth grade</td>
</tr>
</tbody>
</table>
Traditional Readability Formulas

- Over two hundred traditional readability formulas have been developed (cf. Dubey 2004).
- They are generally developed for special purposes, such as determining the complexity of military training manuals (Caylor et al. 1973).
- A frequently used traditional readability measure is the Flesch-Kincaid formula (Kincaid et al. 1975)

![Example: Flesh-Kincaid](https://example.com/kincaid.png)

- Computes U.S. grade level needed to read a text.
- Derived empirically from set of hand-classified documents.

Flesch-Kincaid = \(-15.59 + 11.8 \cdot \text{AWL}_s + 0.39 \cdot \text{ASL}\)

Where

\[ \text{AWL}_s = \frac{\text{Number of Syllables}}{\text{Number of Words}} \]  
\[ \text{ASL} = \frac{\text{Number of Words}}{\text{Number of Sentences}} \]

- Idea:
  - The longer the word, the harder it is.
  - (and the less common it is, cf. Zipf 1936)
  - The longer the sentence, the harder it is to understand.

Another example: Dale & Chall (1948)

\[ \text{Dale-Chall} = 0.1579 \cdot \text{DS} + 0.0496 \cdot \text{ASL} + 3.6365 \]

Where

\[ \text{DS} = \text{Dale Score} \]  
\[ \text{ASL} = \frac{\text{Number of Words}}{\text{Number of Sentences}} \]

- Adds the idea of a specific list of “easy” words.
- List produced by “testing forth-graders on their knowledge in reading of a list of approximately 10,000 words”.
- The more words are outside the set of “easy” words, the more difficult the text is.

Traditional readability measures: Evaluation

- Pros:
  - Relatively simple to use.
  - ‘Simple’ NLP only: tokenizer, stemming, sentence splitter, sometimes syllable counter
- Cons:
  - Originally developed and validated using very small and often highly specific data sets (e.g., technical manuals).
  - No explicit validation of automatic analysis compared to original human analysis (e.g., syllable counting)
  - Measures such as sentence length are relative to domain.
  - Underlying assumptions (e.g., ‘long sentences are difficult’) are rather crude generalizations.
Lexical Frequency Profiles (LFPs)

- Introduced by Laufner & Nation (1995) for the purpose of measuring the vocabulary used by learners.
- Ott (2009) uses LFPs 'upside down': measuring vocabulary in texts for learners, not by learners.
- LFPs work with 3 word lists:
  - First 1000 words of the General Service List (West 1953).
  - General Service List: list of words sorted by frequency
  - Second 1000 words of the General Service List.
  - Academic Word List (Coxhead 2000).
  - Underlying assumption: lists are mutually exclusive.

Lexical Frequency Profile: Example

Results for a typical Wikipedia article:

<table>
<thead>
<tr>
<th>Word List</th>
<th>Tokens</th>
<th>Types</th>
<th>Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSL 1</td>
<td>2202</td>
<td>75.39%</td>
<td>542</td>
</tr>
<tr>
<td>GSL 2</td>
<td>121</td>
<td>4.14%</td>
<td>94</td>
</tr>
<tr>
<td>AWL</td>
<td>245</td>
<td>8.39%</td>
<td>136</td>
</tr>
<tr>
<td>Others</td>
<td>353</td>
<td>12.08%</td>
<td>227</td>
</tr>
<tr>
<td>Total</td>
<td>2921</td>
<td>100%</td>
<td>999</td>
</tr>
</tbody>
</table>

- Families: related by simple morphological processes
  - e.g., happy, happily, and happiness are in same family

Vocabulary-based measures

**Pros:**
- Vocabulary is an important issue for learners.
- ‘Simple’ NLP only: tokenizer, lemmatizer, perhaps tagger.
- Measure can be informed by controlled vocabulary lists of text books.
- Lists can also be extracted from corpora.

**Cons:**
- Vocabulary changes constantly, e.g., the General Service List was published in 1953 and correspondingly does not contain words such as Internet or e-mail?
- Vocabulary is domain-specific: Does the Academic Word List contain words of your field of research?

Syntactic Complexity

**Vocabulary useful indicator, but if sentences are complex, learners will still have trouble understanding them.**

**Sentence length as used in readability formulas simplistic.**

**How can syntactic complexity be measured?**

**Two simple units (Hunt 1965):**
- Clause: “a structure with a subject and a finite verb”
- T-unit: “a main clause plus any subordinate clauses”
Measuring syntactic complexity

Lu (2009) automates 14 measures of syntactic complexity which have been discussed as correlating with L2 proficiency:

<table>
<thead>
<tr>
<th>Type</th>
<th>Measure</th>
</tr>
</thead>
</table>
| Length of production| Mean length of clause  
Mean length of sentence  
Mean length of T-unit   |
| Sentence complexity  | Mean number of clauses per sentence                                      |
| Subordination        | Mean number of clauses per T-unit  
Mean number of complex T-units per T-unit  
Mean number of dependent clauses per clause  
Mean number of dependent clauses per T-unit |
| Coordination         | Mean number of coordinate phrases per clause  
Mean number of coordinate phrases per T-unit  
Mean number of T-units per sentence          |
| Particular structures| Mean number of complex nominals per clause  
Mean number of complex nominals per T-unit  
Mean number of verb phrases per T-unit       |

Textbook structures: Example

Linguistic structures taught in a textbook for English (Klett: Green Line 4, Weisshaar 2008):

<table>
<thead>
<tr>
<th>Unit</th>
<th>Structures taught</th>
</tr>
</thead>
</table>
| 1    | Present perfect progressive with *since* and *for*  
Past perfect progressive  
Attributive use of adjectives after nouns  
Adverbs of degree               |
| 2    | Perfect infinitive with modal verbs  
Passive infinitive with full verbs and modals                                       |
| 3    | Gerund as subject, object, and after verbs and adjectives with prepositions  
Object plus -ing form  
Present and past progressive passive  
Passive with verbs with prepositions                                         |
| 4    | Verb plus object plus infinitive  
Infinitive after question words and after superlatives  
Infinitives vs. Gerund                                                   |
| 5    | Non-defining relative clauses  
Participles as adjectives                                                      |

Textbook structures

- Textbooks introduce linguistic categories and forms in order of perceived complexity.
- For the purpose of teaching grammar, particular structures are especially relevant, e.g. ‘give me a text with a lot of gerunds’.
  - Ott & Ziai (2008) developed a constraint grammar-based approach for classifying -ing forms into gerunds, participles, and the progressive forms.

Information Retrieval

Manning et al. (2008, ch. 1):

“Information Retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).”
**Indexing does the trick in IR!**

**Simply put:**
- Usually one has documents that contain words (“terms”).
- Re-sort everything so that one has terms that are associated with documents → indexing.
- Result: the terms from the query can be mapped to terms in the index at low cost, giving you the corresponding documents quickly.

**Example: Boolean index**

<table>
<thead>
<tr>
<th>Doc1</th>
<th>Doc2</th>
<th>Doc3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Jackie</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Jon</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>likes</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>loves</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Vickie</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Index with weights: Example**

- **TF-IDF (Term Frequency · Inverse Document Frequency):** Weigh terms which occur in fewer documents more highly.

**Text models**

- In addition to the words themselves, any information about a text can be used as an index.
  - Here: readability measures
- All measures are stored in a table for each text, the so-called text model.
- The table contains the key (name) for each measure and a value.
### Example of a text model (extract)

#### Type | Key | Value
--- | --- | ---
General | Character Count | 14249
General | Sentence Count | 111
General | Token Count | 2542
General | Type-Token Ratio | 0.3703
LFP | Academic Word List Token Ratio | 0.0816
LFP | Academic Word List Type Ratio | 0.1389
LFP | General Service List 1k Token Ratio | 0.1389
LFP | General Service List 1k Type Ratio | 0.4191
LFP | General Service List 2k Token Ratio | 0.0557
LFP | General Service List 2k Type Ratio | 0.0841
LFP | Off-List Token Ratio | 1.3119
LFP | Off-List Type Ratio | 0.1325
Readability | Automatic Readability Index | 12.7182
Readability | Flesch Reading Ease | 57.6363
Readability | Gunning Fog Index | 19.4510
Readability | Original Dale-Chall Score | 8.8971

### Towards Evaluation

- An experiment with 190,872 unique documents downloaded from 7 online encyclopedias.
- Encyclopedias are likely to contain articles on one topic each, but with different text difficulty.
- Sample of 7,000 text models (1,000 models for each site).

### Towards Evaluation: Some results

#### Distribution of scores from two grade level-based measures:

- **RARI**
- **RColemanLiau**

![Distribution of scores from two grade level-based measures](image-url)

- This type of evaluation gives only a first impression.
- A gold standard (annotated corpus) should be created and used instead.
Summary

- Fostering language awareness is a well-motivated component of FLT.
- We discussed WERTi: web-based activity generator based on real-world texts selected by the learner.
  - a learner-driven approach, in which learners can
  - generate as many activities as they want
  - choose texts that match their interests
  - activities that remain fully contextualized as whole articles with the original web presentation intact
  - learner interaction with simple feedback based on the original text and linguistic analysis
- Develop search for real-world texts supporting a range of reading difficulty measures and specific linguistic categories → IR4LL.

References


