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.

The role of awareness

Research on awareness shows:

- There is no learning without noticing.
- Awareness without input is not sufficient.
- “Learning takes place within the learner’s mind and cannot be completely engineered by teachers or syllabus designers.”

One can only provide opportunities for developing learner awareness.

⇒ Consequences:

- 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.
  - IR4LL: Retrieval of authentic texts at the appropriate level for language learners
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.

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)

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:
  1. Receptive presentation
  2. Productive presentation
  3. 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.

Examples for an activity progression

1. Pronouns

**Step 1.** Receptive presentation
Ex. System colors different pronoun types.

(1) **Someone told me** that **he** accidentally hit **himself** in the face with **his** car keys.

**Step 2.** Productive presentation
Ex. Click on examples of a particular type of pronoun.

**Step 3.** Controlled practice
Ex. Fill in all pronouns in a text.
Ex. Find and correct incorrect pronoun choices in text.
E.g.: That's **him** car. → That's **his** car.
Examples for an activity progression
2. Passive

Step 1. Receptive presentation
Ex. System colors passive verb forms.
(2) Her purse was taken while she wasn’t looking.

Step 2. Productive presentation
Ex. Click on passive sentences

Step 3. Controlled practice
Ex. Given the main verb stem, fill in the passive verb string (i.e., the correct form of be and the past participle form of the main verb).
Ex. Given an active sentence, transform the sentence to a passive using a combination of click and drag, and FIB.

Examples for an activity progression
3. Adverb placement

Step 1. Receptive presentation
Ex. System colors verbs and verb-modifying adverbs.
(3) The house had already been damaged.

Step 2. Productive presentation
Ex. Click on adverbs in a particular position:
- at the beginning of a sentence
- between a main verb and a prepositional phrase
- before an auxiliary verb

Step 3. Controlled practice
Ex. Given constituent chunks and an adverb, with instructions on where this adverb should go, put the sentence together.

Examples for an activity progression
4. Tense and Aspect

Step 1. Receptive presentation
Ex. System colors examples of different aspectual meanings together with relevant contextual cues.
(4) a. We are going to New York tomorrow.
b. We usually go to the grocery store on Fridays.

Note: While the effect is semantic, the cues are lexical.

Step 2. Productive presentation
Ex. Click on sentences expressing a particular kind of meaning with the targeted verb forms, e.g., expressing future plans using present tense.

Step 3. Controlled practice
Ex. Given a main verb stem, provide the appropriate verb string using cues from context.

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.
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

- 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:
  - original prototype: http://purl.org/net/WERTi
  - current prototype: http://delos.sfs.uni-tuebingen.de:8080/WERTi
Introduction
Pedagogical grounding
The WERTi system
Modeling FLT practice
Progression in WERTi
Example 1: Pronouns
Example 2: Passive
Example 3: Adverb placement
Example 4: Tense and Aspect
Ranking proposal
Creating an example progression
Prototype
Some challenges
Related approaches
IR4LL (Ott 2009)
Measuring Text Difficulty
Readability Formulas
Lexical Frequency Profiles
Syntactic Complexity
Textbook structures
A Search Engine Prototype
Information Retrieval
Adding Text Difficulty
Related Approaches
Towards Evaluation
Summary

Prepositions

Select an activity based on the results of your search:

- Moove slow and don’t hug cows, hikers told (Wednesday, August 30, 2006 1:01 AM ET)
  - Evelyne Zaugg of the Swiss Hiking Federation said that while there were no precise statistics on incidents involving cows, walkers are reporting more run-ins...

- Cows 'moo' with an accent, farmers believe (Thursday, August 24, 2006 2:54 AM ET)
  - LONDON (Reuters) - Cows have regional accents, a group of British farmers claims, and phonetics experts say the idea is not as far fetched as it sounds...

- US drivers subsidize European pump prices - report (Thursday, August 31, 2006 9:13 AM ET)
  - Average 33 cent. American consumers have become the "cash cows" for the International oil industry, the study said. Unlike US...

- Accidental death of bear fuels passions in France (Wednesday, August 30, 2006 5:53 AM ET)
  - A bear's accidental death has sparked protests in the French Alps. An Austrian girl was killed by the creature outside a hotel in the area...

Prototype

Some challenges
Related approaches
IR4LL (Ott 2009)
Measuring Text Difficulty
Readability Formulas
Lexical Frequency Profiles
Syntactic Complexity
Textbook structures
A Search Engine Prototype
Information Retrieval
Adding Text Difficulty
Related Approaches
Towards Evaluation
Summary

Credits
Universität Tübingen
17 / 54

Introduction
Pedagogical grounding
The WERTi system
Modeling FLT practice
Progression in WERTi
Example 1: Pronouns
Example 2: Passive
Example 3: Adverb placement
Example 4: Tense and Aspect
Ranking proposal
Creating an example progression
Prototype
Some challenges
Related approaches
IR4LL (Ott 2009)
Measuring Text Difficulty
Readability Formulas
Lexical Frequency Profiles
Syntactic Complexity
Textbook structures
A Search Engine Prototype
Information Retrieval
Adding Text Difficulty
Related Approaches
Towards Evaluation
Summary

Credits
Universität Tübingen
18 / 54

Introduction
Pedagogical grounding
The WERTi system
Modeling FLT practice
Progression in WERTi
Example 1: Pronouns
Example 2: Passive
Example 3: Adverb placement
Example 4: Tense and Aspect
Ranking proposal
Creating an example progression
Prototype
Some challenges
Related approaches
IR4LL (Ott 2009)
Measuring Text Difficulty
Readability Formulas
Lexical Frequency Profiles
Syntactic Complexity
Textbook structures
A Search Engine Prototype
Information Retrieval
Adding Text Difficulty
Related Approaches
Towards Evaluation
Summary

Credits
Universität Tübingen
19 / 54

Introduction
Pedagogical grounding
The WERTi system
Modeling FLT practice
Progression in WERTi
Example 1: Pronouns
Example 2: Passive
Example 3: Adverb placement
Example 4: Tense and Aspect
Ranking proposal
Creating an example progression
Prototype
Some challenges
Related approaches
IR4LL (Ott 2009)
Measuring Text Difficulty
Readability Formulas
Lexical Frequency Profiles
Syntactic Complexity
Textbook structures
A Search Engine Prototype
Information Retrieval
Adding Text Difficulty
Related Approaches
Towards Evaluation
Summary

Credits
Universität Tübingen
20 / 54
**Realizing the proposal**

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?
- The complexity of real life:
  - Real-life texts from the web often have
    - complex structure
    - mark-up and integrated multimedia
  - It is nontrivial to preserve that structure and mark-up during linguistic annotation of the text base.
  - Receptive and productive presentation can be added modularly to an existing document (mark-up/javascript); inserting forms for practice more challenging.

---

**Related approaches**

The MIRTO project (Antoniadis et al. 2004)

- **Similarities**
  - Emphasizes pedagogical practice and integration
  - Automatic exercise generation:
    - Plans to support “gap-filling” and “lexical spotting” exercises in combination with a corpus database.
- **Differences**
  - Aims at creating a general toolbox architecture supporting instructor-determined activity design.
  - General toolbox = no explicit mention of language awareness or specific pedagogical progressions or aims

**Related approaches**

VISL: Visual Interactive Syntax Learning (Bick 2001, 2005a)

- **Similarities**
  - Emphasis on language awareness:
    - VISL offers games and visual presentations to foster knowledge of syntactic forms and functions.
  - Automatic exercise generation:
    - The "exercise building tool" KillerFiller automatically creates slot-filler exercises from texts.
- **Differences**
  - KillerFiller intended as evaluative tool, not for teaching.
  - Annotated corpora and databases used as text base.
  - Sentences presented in isolation, not in context.
  - Slots determined by general category (e.g., prepositions, verbs), not more specific or other linguistic features.
Related approaches
Generating cloze tests

Automatic generation of multiple choice “cloze tests” (FIB) for language testing and vocabulary drill
(cf., e.g. Coniam 1997; Irvine & Kylönen 2002; Deane & Sheehan 2003; Huang et al. 2005; Liu et al. 2005)

- Sumita et al. (2005): automatic generation of FIB questions for testing English proficiency
  - selection of seed sentence mentioned as relevant issue
  - uses web to test whether potential distractor items are indeed incorrect
  - addresses testing, not pedagogical exercise progression
  - sentences not selected by learner or contextualized

IR4LL (Ott 2009)
Measuring Text Difficulty
Readability Formulas
Lexical Frequency Profiles
Syntactic Complexity
Textbook structures
A Search Engine Prototype
Information Retrieval
Adding Text Difficulty
Related Approaches
Towards Evaluation

IR4LL Proposal

- Create a search engine that is aware of variations in text difficulty.
- Challenges and research questions:
  - How to measure text difficulty?
  - Is there enough variety in text difficulty on the web?
  - Are there enough ‘easy’ web pages?

Some challenges
Related approaches

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

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>Named Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>College graduate</td>
</tr>
<tr>
<td>16</td>
<td>senior</td>
</tr>
<tr>
<td>15</td>
<td>junior</td>
</tr>
<tr>
<td>14</td>
<td>sophomore</td>
</tr>
<tr>
<td>13</td>
<td>freshman</td>
</tr>
<tr>
<td>12</td>
<td>High School senior</td>
</tr>
<tr>
<td>11</td>
<td>junior</td>
</tr>
<tr>
<td>10</td>
<td>sophomore</td>
</tr>
<tr>
<td>9</td>
<td>freshman</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>

**Example: Flesh-Kincaid**

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

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

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)

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

Where

\[ DS = \text{Dale Score} \]

The percentage of words outside the Dale list of 3000 words.

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

Average sentence length.

- 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.

Lexical Frequency Profiles (LFPs)

- Introduced by Laufer & 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.

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).
  - Whether the automated analysis using computers agrees with the original human analysis has generally not been validated.
  - Measures such as sentence length are domain-dependent.
  - Underlying assumptions (e.g., ‘long sentences are difficult’) are rather crude generalizations.

Lexical Frequency Profile: Example

<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>542</td>
<td>384</td>
</tr>
<tr>
<td>GSL 2</td>
<td>121</td>
<td>94</td>
<td>78</td>
</tr>
<tr>
<td>AWL</td>
<td>245</td>
<td>136</td>
<td>109</td>
</tr>
<tr>
<td>Others</td>
<td>353</td>
<td>227</td>
<td>n.a.</td>
</tr>
<tr>
<td>Total</td>
<td>2921</td>
<td>999</td>
<td>100%</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?

### 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  |

### 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”

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

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

**Doc1:**
Jon loves Vickie.

**Doc2:**
Vickie likes Jackie.

**Doc3:**
Jackie loves Ian.
Ian loves Jackie.

<table>
<thead>
<tr>
<th>Doc1</th>
<th>Doc2</th>
<th>Doc3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jon</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>1</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>
Text models

- All measures are stored in a table for each text.
- The table contains a key (name) for each measure and a value.
- This is flexible since this text model can be extended easily in future versions.
- For IR, an index is generated which contains the terms as well as the information encoded in the text model.

### Example of a text model (extract)

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

### NLP pipelines in the indexer

- HTML Preprocessing
  - HTML2plaintextMapper
  - ParagraphSpanAnnotator
  - GenericRelevanceAnnotator
  - html2plaintextMapper
- Generic NLP
  - LanguageChecker
  - SplitWordSentenceAnnotator
  - OpenNlpTagger
  - morphLemmatizer
- Readability
  - oldDaleChall
  - SimpleReadabilityMeasures
  - LexicalFrequencyProfiler
  - RelevantText2Model

---

Index with weights: Example

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

Doc1:
Jon loves Vickie.

Doc2:
Vickie likes Jackie.

Doc3:
Jackie loves Ian.
Ian loves Jackie.
Read-X (Miltsakaki & Troutt 2008)

▶ Similar to REAP, Read-X uses Yahoo!.
▶ Queries are submitted to Yahoo via its API. Then they are all downloaded and post-filtered for readability.
▶ Since Read-X is a program running on the learner's PC (not a web application, using IR system or web crawling).
▶ The system is internally based on the Coleman-Liau index (Coleman & Liau 1975) and the RIX formula (Anderson 1983).
▶ Read-X also uses text models.
▶ No classification according to grammatical structures.

REAP Search

The REAP system by Heilman et al. (2008) aims at a similar task from a different perspective.

▶ In their system, a digital library of readings is created by querying AltaVista ('query-based crawling').
▶ Texts are controlled by a human instructor before they are presented to learners.
▶ The system aims at reading practice and vocabulary learning. Therefore it uses a special reading interface.
▶ Instead of text models and query models, the documents are classified using machine learning.
    ▶ This is less flexible because one cannot merge multiple classifications at query time.
▶ Due to the focus on vocabulary, there is no possibility to query for specific linguistic forms that could be practiced in WERTi.
Towards Evaluation: Some results

Distribution of scores from two grade level-based measures:

- R_ARI
- R_ColemanLiau

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


