Text readability and simplification
Connecting linguistics, processing, and educational applications

- **Readability assessment**: Are web pages accessible to people with diverse language proficiencies?
- **Text Simplification**: Can we simplify texts to lower reading levels without losing meaning?

I explore these questions in a language learning context
- to identify suitable texts for language learners
- to produce simplified texts or target users, if necessary.

1 **Research Question 1: Readability Assessment**
   - task: assessing the reading level of a given document.
   - primary focus: on utilizing insights from Second Language Acquisition (SLA) and cognitive psychology.
   - This approach yielded good results. (Vajjala & Meurers, 2012)
     - classification accuracy: 93%
     - for a classification in to five reading levels.
   - The approach also worked with German, giving a 89.7% accuracy. (Hancke, Vajjala & Meurers 2012)

2 **Can it be Applied in Real World?**
   - I used my readability models to classify web documents in to various reading levels.
   - The model successfully identified diverse reading levels across several web texts, that broadly fell into **easy** and **difficult** categories

   ![Graph showing readability levels](image)

   - Is this useful to rank search results?
     - There are actually texts of diverse reading levels even within the top-10 results.
     - But, the average reading levels of the texts are still higher than the high-school reading level.

   ![Graph showing average readability levels of search results](image)

   - Sample size: 50 search queries

   - So, to conclude,
     - Readability assessment works on the web documents.
     - But, text simplification is necessary as most of the news texts fall in to higher reading levels.

3 **Research Question 2: Text Simplification**
   - Is readability assessment useful to identify target sentences for text simplification?
   - It turned out that this is not the case.
     - What worked at 93% accuracy at the document level only worked at 70-75% accuracy on sentence level.
   - Identifying readability at a sentence level is difficult
     - because we have very little data to analyze?
   - Current work, focused on
     1. identifying a sentence as simple or complex
     2. identifying the class of simplification: expansion, compression, word replacement, split
     3. identifying actual simplification points within the sentence

4 **Way Ahead: Role of LEAD**
   - final step: generating the actual simplified sentences.
   - primary focus so far: on linguistic modeling
   - LEAD connection:
     1. Insights on reading and comprehension from related LEAD areas may aid in building better systems.
     2. Educational context is crucially needed to evaluate simplification for language learners in real life.

5 **References**
   - Readability classification for German using Lexical, Syntactic and Morphological features. Julia Hancke, Sowmya Vajjala and Detmar Meurers, COLING 2012.