Exploring linguistic complexity in readability analysis and L2 development

Detmar Meurers
Universität Tübingen

based on joint research with

Sowmya Vajjala and Xiaobin Chen

Oberseminar English Linguistics
Universität Frankfurt
January 9, 2017

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

Linguistic Complexity

- Aspects of linguistic complexity are used to characterize
  - the increasingly elaborate and varied language produced by learners in **Second Language Acquisition** research
  - which audience can read a text in **Readability** research
  - how hard it is for humans to process sentences (lexical frequency, Dependency Locality Theory, surprisal, . . .)
- Aspects of linguistic complexity not touched on here:
  - comparison of linguistic systems (are some languages more complex than others, recursion, . . .)
  - comparison of linguistic theories (are some analyses less complex than others, . . .)
  - language change (historical development from more to less complex, where does complexity come from?, . . .)

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

What is readability analysis?

We want to determine how difficult it is to read

- a given **text**
- for a given **purpose**, e.g.,
  - skimming for information
  - answering comprehension questions
- for a given individual **reader** with
  - their knowledge of the topic domain
  - individual differences in cognition, affect, personality

⇒ Which characteristics of the texts can we consider?

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

Traditional approaches to readability

- Long history of readability formulas (DuBay 2004)
  - Developed for specific purposes, e.g., characterizing demands of military training manuals (Caylor et al. 1973)
- Formulas are based on shallow, easy to count features:
  - typically avg. sentence length and avg. word length, e.g., Flesch-Kincaid formula (Kincaid et al. 1975)
  - counts of words on specific word lists (Dale & Chall 1948)
- Problems of traditional readability formulas:
  - based on rough generalizations:
    - long words are rare, long sentences are difficult
  - formulas are domain dependent
  - provide only a quantitative measure, not a characterization of the language aspects involved in readability

What can we observe about a given text?

I. Which language **forms** are used, how are they combined?
- **type** of forms in the **linguistic system**
  e.g.: complex NPs per sentence
- **use** of forms in terms of personal **language experience**, evidence via proxy of representative language records
  e.g.: word frequency, average AoA

II. What type & amount of **meaning** is encoded by the forms, and how is it organized into a coherent discourse?
  e.g.: concreteness, lexical density, referential cohesion

III. What are its demands on **human processing**?
  e.g. memory load for referents, surprisal

What can we observe about a given text?

How can we obtain relevant measures?

I. Which language **forms** are used, how are they combined?
⇒ linguistic observations (complex NPs, embedding, . . .)
→ measures of language proficiency established in SLA
⇒ norms for frequency (SUBTLEX), AoA (crowd sourcing)

II. What type & amount of **meaning** is encoded by the forms and how is it organized into a coherent discourse?
⇒ lexical semantic information in databases (MRC, WordNet), CohMetrix measures of coherence/cohesion

III. What are its demands on **human processing**?
⇒ measures from human sentence processing literature, e.g., surprisal (Boston et al. 2008)

SLA measures of proficiency development

- Second Language Acquisition research has developed a rich inventory of measures for monitoring development.
- Skehan (1989) characterized proficiency in terms of the three dimensions Complexity, Accuracy, und Fluency (CAF, Wolfe-Quintero et al. 1998; Housen & Kuiken 2009)
- **Complexity**:
  *The extent to which the language produced in performing a task is **elaborate** and **varied**.*
  (Ellis 2003, p. 340)

Connecting Readability and L2 Complexity

- How about making use of
  - SLA measures of the **complexity** of **learner language**
  - for determining the **readability** of **native texts**?
- Motivation:
  - profit from rich set of SLA measures operationalizing complexity at all levels of linguistic modeling
  - using the same features to characterize reading texts and language proficiency can make it possible to
    - tailor complexity of input to learner proficiency (i+1)
- Putting the idea to the test:
  - Vajjala & Meurers (2012, 2013, 2014a,b,c), Vajjala (2015)
  - Chen & Meurers (2016a,b)

Testing how well the idea works

A supervised machine learning setup as experimental sandbox

- Take a **corpus** of texts for which reading levels are known.
- Spell out hypotheses which properties matter as **features**.
- Train a machine learning **model**.
  - classification: discrete levels (e.g., beg., int., adv.)
  - regression: continuous levels (e.g., age)
  - ranking: relative level (which of two is easier)
- **Evaluate** model by predicting levels of unseen texts.

Corpus

- Needed: a corpus with gold-standard labels
- Previous work: graded reading material in WeeklyReader
- We compiled the WeeBit corpus (Vajjala & Meurers 2012):

Grade Level	*Age in Years*	*Number of Articles*	*Avg. Number of Sentences/Article*
from WeeklyReader			
Level 2	7–8	629	23.41
Level 3	8–9	801	23.28
Level 4	9–10	814	28.12
from BBCBitesize			
KS3	11–14	644	22.71
GCSE	14–16	3500	27.85

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Features from SLA research

Lu (2010, 2011, 2012) surveyed complexity features used in SLA research, which we select many of our features from.

Lexical Features

- Lexical Variation
  - Type-Token Ratio = *Typ* / *Tok*
    - influenced by text length
  - Measure of Textual Lexical Diversity (MTLD, McCarthy 2005)
    - average number of words needed to reach stable TTR point
- Lexical Density = Tok_{Lex} / Tok
  - Lex = open lexical classes (N, Adj, Adv, V)
- Overall we use: 19 lexical features (16 SLA, 3 others)

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Features from SLA research

Syntactic complexity features

- analyze different units: sentences, T-units, clauses
  - a) mean length per unit
    - e.g., mean length of sentences
  - b) number of occurrences per unit
    - e.g., number of clauses per sentence
  - c) ratios of different subtypes (subordination, coordination)
    - e.g., dependent clauses per clause, . . .
  - d) specific constructions
    - e.g., complex nominals per clause, . . .
- Overall we use: 25 syntactic features (14 SLA, 11 others)

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

More Features

- Other syntactic features
  - Average parse tree height
  - Average number of NPs, VPs, and PPs per sentence
  - Mean length of NP, VP, and PP
- Traditional features
  - Traditional Features: avg. word length, sentence length
  - Traditional Formulas: Flesch-Kincaid, Coleman-Liau score
  - Word lists: Academic Word List

Experimental Setup
(Vajjala & Meurers 2012)

- Corpus used for experiment: WeeBit
  - 500 training documents per level
  - 125 testing documents per level
- Features computed using standard NLP tools:
  - OpenNLP part-of-speech tagger
  - Berkeley Parser (Petrov & Klein 2007)
  - Tregex pattern matcher (Levy & Andrew 2006) using definitions from Lu (2010)
- Machine learning setup:
  - classification with five classes (levels 2, 3, 4, KS3, GCSE)
  - explored various algorithms in WEKA:
    - decision trees, support vector machines, logistic regression
    - reporting multi-layer perceptron results

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Results on WeeBit (5 classes)

	Number of Features	Performance Accuracy	RMSE
WeeklyReader: Previous Work			
Feng (2010)	122	74.0%	
Petersen & Ostendorf (2009)	25	63.2%	
P. & O. syntactic features only	4	50.9%	
WeeklyReader (Vajjala & Meurers 2012)			
Replication P. & O. syntactic feat.	4	50.7%	
Our Syntactic Features	25	64.3%	0.37
Our Lexical Features	19	84.1%	0.23
All our Features	46	**91.3%**	0.17
WeeBit (Vajjala & Meurers 2012)			
SLALEX	16	68.1%	0.29
SLASYN	14	71.2%	0.28
SLALEX + SLASYN	30	**82.3%**	0.23
All our Features	46	**93.3%**	0.15
BEST10FEATURES	10	89.7%	0.18

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Ten Best Features (Information Gain)

- Half of the best features are **SLA complexity measures**:
  - mean length of a sentence
  - dependent clause to clause ratio
  - complex NPs per clause
  - modifier variation (proportion of adjectives & adverbs)
  - adverb variation (proportion of adverbs)
- The other features in the Top 10:
  - avg. num. characters per word
  - avg. num. syllables per word
  - proportion of words on Academic Word List
  - num. co-ordinate phrases per sentence
  - Coleman-Liau score

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Extending the feature set

(Vajjala & Meurers 2014a)

- Add more features which are meaningful for sentence-level analysis and comparison.

⇒ Add features on lexical system and language use:

- Morphological properties of a word (from Celex)
  e.g., Is the word derived from a stem along with an affix? *abundant=abound+ant*
- Lexical semantic properties of a word (from WordNet)
  e.g., Avg. number of senses per word
- Psycholinguistic features of words
  e.g., word abstractness, average age-of-acquisition (AoA)

Realization of the extended feature set

(Vajjala & Meurers 2014a)

- Resources:
  - Celex Lexical Database (http://celex.mpi.nl)
  - Kuperman et al. (2012)'s AoA ratings
  - MRC Psycholinguistic database (http://ota.oucs.ox.ac.uk/headers/1054.xml)
  - Wordnet (http://wordnet.princeton.edu)
- Tools:
  - Features computed using:
    - Stanford Tagger (Toutanova, Klein, Manning & Singer 2003)
    - Berkeley Parser (Petrov & Klein 2007)
    - Tregex Pattern Matcher (Levy & Andrew 2006)
  - Machine Learning using WEKA
    - SMOReg algorithm (modeling readability as regression) trained on WeeBit corpus

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

Results on standard CCSS corpus

- Common Core State Standards reading initiative of the U.S. education system (CCSSO 2010)
- Reference corpus: 168 texts for grade levels 2–12
- Results (Spearman's rank correlation since scales differ):

System	Spearman
Nelson et al. (2012):	
REAP	0.54
ATOS	0.59
DRP	0.53
Lexile	0.50
Reading Maturity	**0.69**
ETS SourceRater	**0.75**
Vajjala & Meurers (2014a)	**0.69**

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

Do the results generalize?

(Vajjala & Meurers 2014c)

- Does the WeeBit **model** generalize to other datasets?

Test set	Spearman
CommonCore	0.69
TASA corpus	0.86

- Impact of genre differences in CommonCore data:

Genre in CommonCore	Spearman
Informative	0.76
Misc.	0.69
Literature	0.51
Speech	0.35

Is the **feature set** informative enough for spoken language?

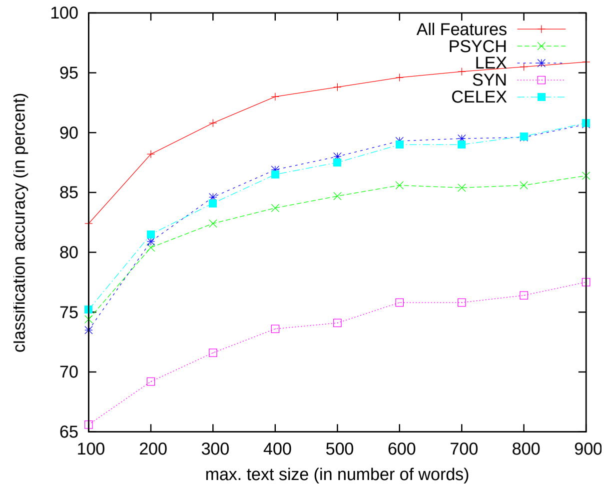
Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

Readability analysis of TV subtitles
(Vajjala & Meurers 2014b)

- We used our feature set to train a model that identifies age-specific TV programs.
- Data: subtitles of BBC TV channels (Van Heuven et al. 2014)
- Classification into three age groups:
  - less than 6, 6–12, adult

⇒ 96% classification accuracy (SMO, 10 fold CV)
  - single most predictive feature: average AoA of words, but accuracy is not reduced if this feature is removed
  - Classification is informed by a wide range of linguistic elaborateness, variedness, and cognitive characteristics.

Effect of Text Size on Classification Accuracy



- Training/testing with longer texts supports higher accuracy.
- But even with 100 words per text, one reaches >80%.
- Lex & Psych best in short texts, Syn more linear increase

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

From texts to sentences

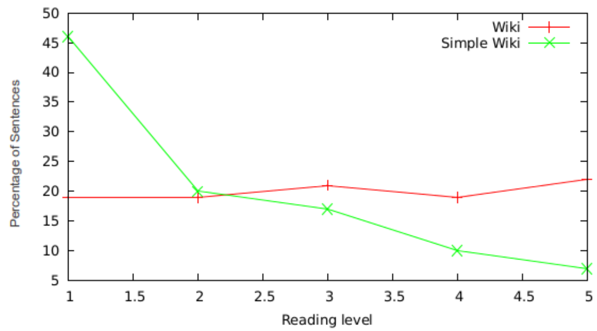
- Can we reliably analyze individual sentences?
- This would be useful
  - for text simplification
    - to identify targets for simplification
    - to evaluate aspects of simplification
  - to evaluate sentences in questionnaires
  - to rank candidates in generation systems

→ Test model trained on WeeBit texts on individual sentences

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Readability at the sentence level
(Vajjala & Meurers 2014a)

- Test on sentence-aligned Wiki–SimpleWiki (Zhu et al. 2010)
- Predictions of WeeBit text model:



- Simplification is relative: A simplified sentence is simpler than its unsimplified version, but can be harder than another one.
- Hard texts are not simply collections of hard sentences.

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Dealing with the multi-level nature of evidence
Beyond averages

- To classify texts, we rely on evidence at different levels:
  - words, sentences, texts
- What is the best way to combine the evidence?
  - Is computing averages really preserving what is relevant?
- Explored for word frequencies in Chen & Meurers (2016a) using SUBTLEX Zipf scale (Van Heuven et al. 2014)

Dealing with the multi-level nature of evidence
Results on WeeBit (Chen & Meurers 2016a)

- Accuracy of 10-fold CV classification on WeeBit (5 levels):
  - Average frequencies baseline:
    - 24.2% with average token frequency as feature
    - 32.1% with average type frequency as feature
  - Adding Standard Deviation:
    - 39.9% with average token frequency + SD as features
    - 43.3% with average type frequency + SD as features

→ Let's explore different levels of granularity.
  - most informative: characterize a text through the vector of frequencies of every token in the text, but:
    - unlikely to generalize, and
    - texts differ in length

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Dealing with the multi-level nature of evidence
Word frequencies in texts at different levels of granularity

- How about grouping tokens to obtain *n* averages per text?
  i) *n* **frequency bands of the language**
  ii) *n* **clusters of words in document** closest in frequency

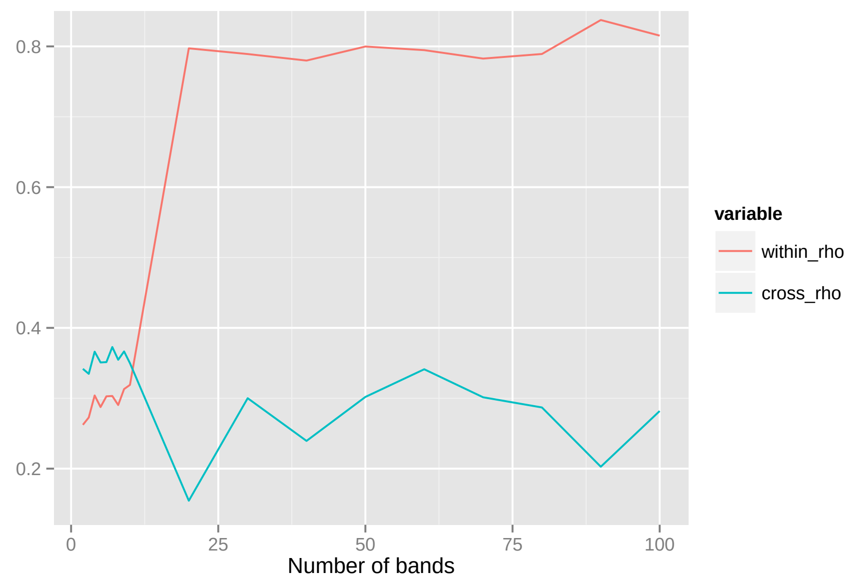
  A text is represented by one avg. frequency feature per group.

⇒ This works well for 10-fold CV in WeeBit corpus:
  i) 67.5% accuracy with 90 frequency bands (by types)
  ii) 54.6% accuracy with 100 clusters (by tokens)

- But does this generalize across corpora?
  → Compare WeeBit 10-fold CV with test on CommonCore, reporting Spearman's rank correlation coefficient (ρ)

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

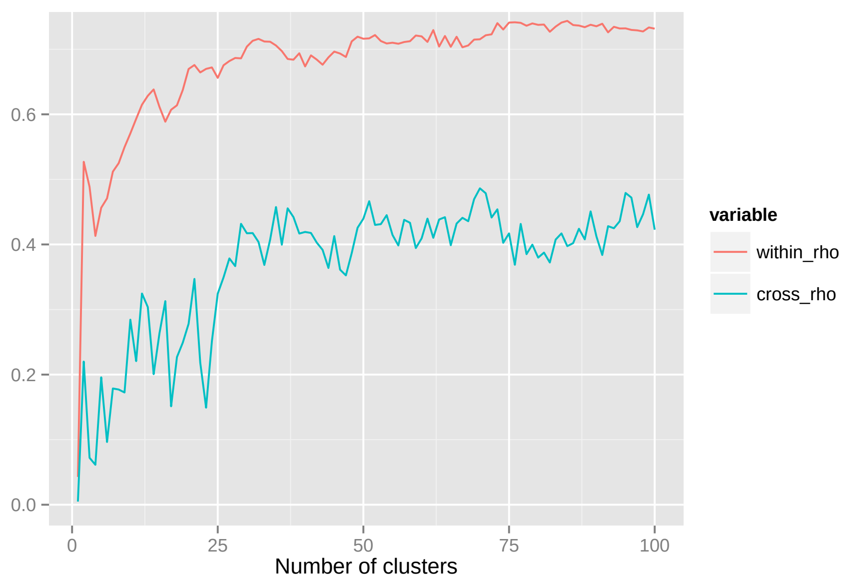
Grouping by frequency bands in the language
Spearman rank correlation within and across corpora



Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Hierarchical clustering of tokens in document

Spearman rank correlation within and across corpora



Dealing with the multi-level nature of evidence

Summary

- For aggregating word frequencies at the text level:
  - grouping by language frequency band better within-corpus
  - hierarchical clustering of words in text generalizes better
- Conclusion: We should put more thought into how to combine the multi-level nature of the readability evidence.
- Next idea to test:
  - How can the incremental process information provided by Surprisal (Boston et al. 2008) inform text difficulty?
  - → Hierarchical clustering of Surprisal profiles of sentences.

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Using readability to rank web search results

(Vajjala & Meurers 2013)

- Are state-of-the-art readability models actually useful for classifying texts as found on the web?
  - Can we re-rank search results based on reading levels?
- Implementation details:
  - feature set from Vajjala & Meurers (2012)
  - trained model on WeeBit corpus
  - modeling: regression, since we want output on a scale
- We applied the readability model to search results obtained through BING search API.
  - took 50 search queries from a public query log
  - computed reading levels for Top-100 results

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

Results: Reading levels of top search results

(Vajjala & Meurers 2013)

Result Rank:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Avg** Top100
Query:											
copyright copy law	1.8	4.6	1.4	2.7	4.6	6.2	2.7	**1.1**	3.9	5.6	4.6
halley comet	**1.7**	4.5	4.5	4.2	2.4	4.1	4.9	3.6	4.2	3.6	4.0
europe union politics	3.6	4.9	6.3	4.0	2.2	4.5	**1.5**	1.6	4.9	6.3	4.3
shakespeare	2.4	2.9	4.2	4.7	4.7	3.9	**1.5**	2.1	2.6	4.0	3.6
euclidean geometry	3.9	4.7	4.7	4.3	4.5	4.6	4.0	4.1	3.5	**2.6**	3.2
...											

- Avg. reading level of search results quite high (5 = GCSE)
- The model identifies a range of reading levels among most relevant results returned by search engine.
- Readability-based re-ranking of results potentially useful for language-aware search engines in Language Teaching (Ott & Meurers 2010; Chinkina & Meurers 2016)

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers
Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

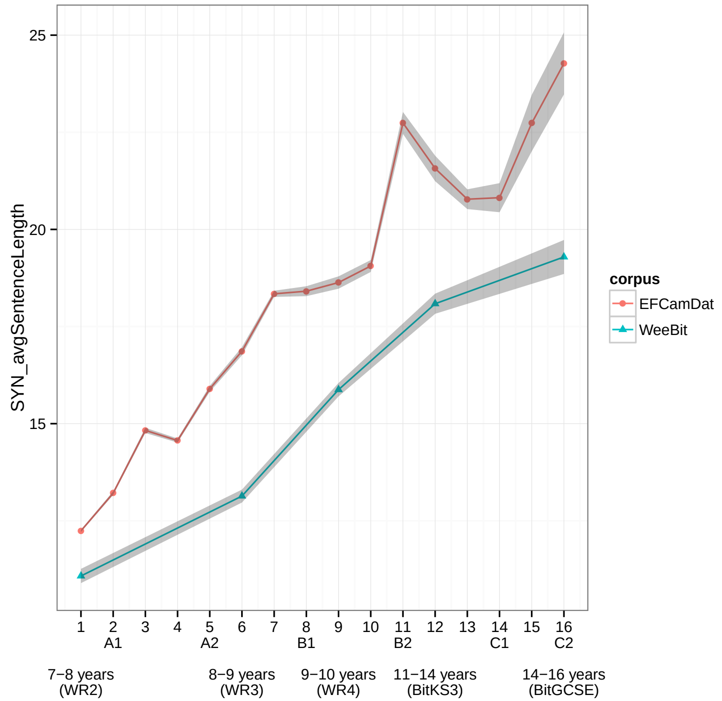
Linking readability and L2 development

- Can we use the complexity features as a looking glass on both readability and L2 development?
  - Idea: Provide texts just above level of student ($i + 1$)
- EFCamDat corpus (Geertzen et al. 2013), prerelease 2:
  - 1.2 million assignments (70 million words)
  - written by nearly 175 thousand learners
  - across a wide range of levels (CEFR A1–C2)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

LEAD Graduate School

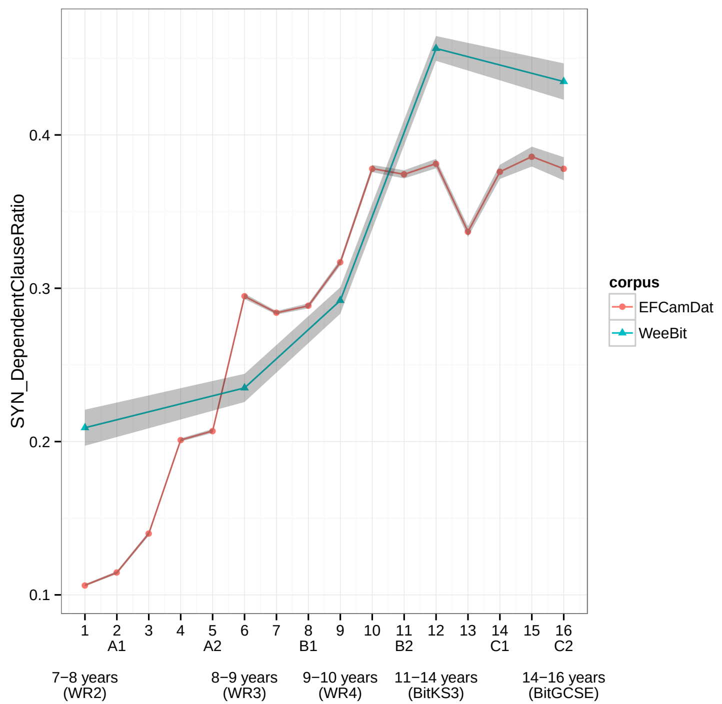
Mean length of a sentence



Exploring linguistic complexity in readability analysis & L2 development

Detmar Meurers

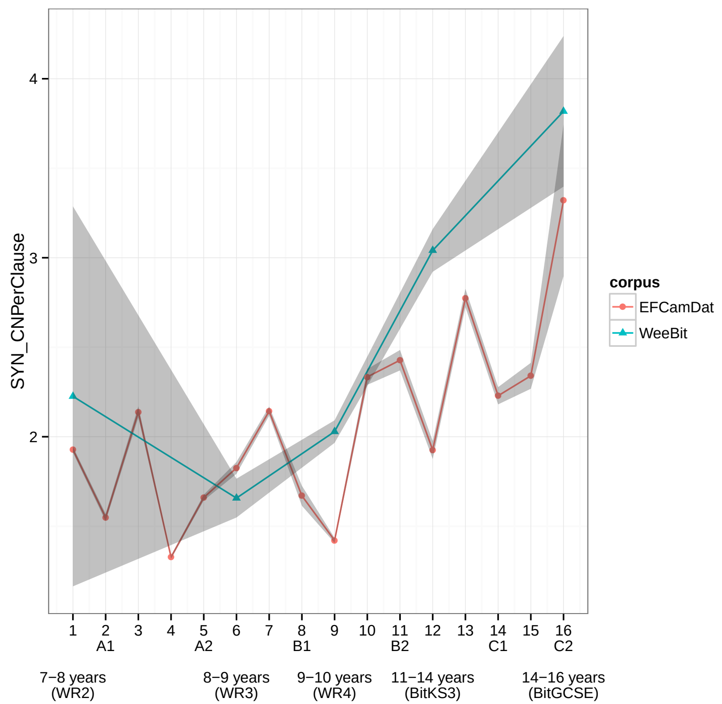
Dependent clause to clause ratio



Exploring linguistic complexity in readability analysis & L2 development

Detmar Meurers

Complex NPs per clause



Exploring linguistic complexity in readability analysis & L2 development

Detmar Meurers

Linking readability and L2 development

First conclusions

- Good face validity for using same measures for readability and L2 development.
- Various challenges that need to be addressed, e.g.:
  - impact of analyzing learner language: e.g., variable orthography should not be counted as lexical richness
  - systematic sentence segmentation difficult (both for learner language and certain text types, e.g., CVs)
- Questions to be addressed for selecting $i + 1$ text material:
  - How reliably can we determine reading ability based on measures of writing proficiency?
  - What does the "+1" amount to, for the different aspects of linguistic modeling (lexicon, syntax, ...)?

Summary

- Measures of development from SLA research turn out to be excellent predictors for readability classification.
- Our approach outperforms previously published readability assessment results on WeeklyReader data.
  - best non-commercial readability model on CCSS data
- Feature set generalizes well to other data sets and genres
  - including spoken language transcripts and web data
- Readability reflected in a wide range of linguistic properties
  - Linguistic and cognitive features complement each other.

Summary (cont.)

- Sentence-level analysis:
  - Documents at a given reading level contain sentences at a range of levels of complexity.
  - Readability analysis at the sentence level is feasible, but needs to take the relative nature of readability into account.
- Taking the multi-level nature of the evidence on readability (word, sent., text) into account can support significant gains.
- Approach is applicable to other languages:
  - German readability (Hancke, Meurers & Vajjala 2012) and proficiency (Hancke & Meurers 2013), readability for Bulgarian (Nikolova 2015) and Greek (Georgatou 2016)

Outlook on some current collaborations

- Analyzing linguistic complexity and accuracy in relation to task demands (Alexopoulou, Michel, Murakami & Meurers 2017)
- At interface with empirical educational science (LEAD):
  - Studying the cognitive correlates of readability using eye-tracking (Vajjala, Meurers, Eitel & Scheiter 2016)
  - Is the language in school books appropriate for grade level and school type? ReadingDemands with Berendes & Bryant
  - Linguistic and numerical factors contributing to the complexity of Word Problems (Daroczy et al. 2015)
  - Impact of linguistic complexity of questionnaires (Göllner et al. 2014)
  - CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis (Chen & Meurers 2016a)

http://www.ctapweb.com

References

Alexopoulou, T., M. Michel, A. Murakami & D. Meurers (2017). Analyzing learner language in task contexts: A study case of task-based performance in EFCAMDAT. *Language Learning* Special Issue on "Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and interpretation".

Boston, M. F., J. T. Hale, U. Patil, R. Kliegl & S. Vasishth (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research* 2(1), 1–12. URL http://www.jemr.org/online/2/1/1.

Caylor, J. S., T. G. Sticht, L. C. Fox & J. P. Ford. (1973). *Methodologies for determining reading requirements of military occupational specialties: Technical report No. 73-5*. Tech. rep., Human Resources Research Organization, Alexandria, VA.

CCSSO (2010). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects. Appendix B: Text Exemplars and Sample Performance Tasks*. Tech. rep., National Governors Association Center for Best Practices, Council of Chief State School Officers. http://www.corestandards.org/assets/Appendix_B.pdf.

Chen, X. & D. Meurers (2016a). Characterizing Text Difficulty with Word Frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego, CA.

Chen, X. & D. Meurers (2016b). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the COLING Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*.

Chinkina, M. & D. 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*. San Diego, CA, pp. 188–198. URL http://aclweb.org/anthology/W16-0521.pdf.

Dale, E. & J. S. Chall (1948). A Formula for Predicting Readability. *Educational research bulletin; organ of the College of Education* 27(1), 11–28.

Daroczy, G., M. Wolska, W. D. Meurers & H.-C. Nuerk (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology* 6(348). URL http://www.frontiersin.org/developmental_psychology/10.3389/fpsyg.2015.00348/abstract.

DuBay, W. H. (2004). *The Principles of Readability*. Costa Mesa, California: Impact Information. URL http://www.impact-information.com/impactinfo/readability02.pdf.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford, UK: Oxford University Press.

Feng, L. (2010). Automatic Readability Assessment. Ph.D. thesis, City University of New York (CUNY). URL http://lijun.symptotic.com/files/thesis.pdf?attredirects=0.

Geertzen, J., T. Alexopoulou & A. Korhonen (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum (SLRF)*. Cascadilla Press. URL http://purl.org/icall/efcamdat.

Georgatou, S. (2016). Approaching readability features in Greek school books. Master thesis in computational linguistics, Department of Linguistics, University of Tübingen.

Göllner, R., D. Meurers, W. Wagner, K. Berendes, B. Nagengast & U. Trautwein (2014). Linguistic complexity of questionnaires: Relevance for psychometric

quality of teaching evaluation from the student perspecrtive and prediction of learner success in large scale assessments. Approved project proposal for the BMBF call "Förderung von Forschungsvorhaben in Ankopplung an Large-Scale-Assessments".

Hancke, J. & D. Meurers (2013). Exploring CEFR classification for German based on rich linguistic modeling. In *Learner Corpus Research 2013, Book of Abstracts*. Bergen, Norway. URL http://purl.org/dm/papers/Hancke.Meurers-13.html.

Hancke, J., D. Meurers & S. Vajjala (2012). Readability Classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbay, India, pp. 1063–1080. URL http://aclweb.org/anthology-new/C/C12/C12-1065.pdf.

Housen, A. & F. Kuiken (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics* 30(4), 461–473. URL http://applij.oxfordjournals.org/content/30/4/461.full.pdf.

Kincaid, J. P., R. P. J. Fishburne, R. L. Rogers & B. S. Chissom (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.

Kuperman, V., H. Stadthagen-Gonzalez & M. Brysbaert (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4), 978–990. URL http://crr.ugent.be/archives/806.

Levy, R. & G. Andrew (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474–496.

Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly* 45(1), 36–62.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Languages Journal* pp. 190–208.

McCarthy, P. & S. Jarvis (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2), 381–392. URL https://serifos.sfs.uni-tuebingen.de/svn/resources/trunk/papers/McCarthy.Jarvis-10.pdf.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Ph.D. thesis, University of Memphis. URL https://umdrive.memphis.edu/pmmccrth/public/Papers/MTLD20dissertation.doc.

Meurers, D. (2012). Natural Language Processing and Language Learning. In C. A. Chapelle (ed.), *Encyclopedia of Applied Linguistics*, Oxford: Wiley, pp. 4193–4205. URL http://purl.org/dm/papers/meurers-12.html.

Meurers, D. (2015). Learner Corpora and Natural Language Processing. In S. Granger, G. Gilquin & F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, pp. 537–566.

Meurers, D. & M. Dickinson (2017). Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration with Computational Linguistics. *Language Learning, Special Issue on Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and Interpretation* URL http://purl.org/dm/papers/Meurers.Dickinson-17.html. To appear.

Nelson, J., C. Perfetti, D. Liben & M. Liben (2012). *Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance*. Tech. rep., The Council of Chief State School Officers. URL http://purl.org/net/Nelson.Perfetti.ea-12.pdf.

Nikolova, L. (2015). Readability Classification for Bulgarian. Master thesis in computational linguistics, Department of Linguistics, University of Tübingen.

Ott, N. & D. Meurers (2010). Information Retrieval for Education: Making Search Engines Language Aware. *Themes in Science and Technology Education. Special issue on computer-aided language analysis, teaching and learning: Approaches, perspectives and applications* 3(1–2), 9–30. URL http://purl.org/dm/papers/ott-meurers-10.html.

Petersen, S. E. & M. Ostendorf (2009). A machine learning approach to reading level assessment. *Computer Speech and Language* 23, 86–106.

Petrov, S. & D. Klein (2007). Improved Inference for Unlexicalized Parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York, pp. 404–411.

Skehan, P. (1989). *Individual Differences in Second Language Learning*. Edward Arnold.

Toutanova, K., D. Klein, C. Manning & Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*. Edmonton, Canada, pp. 252–259.

Vajjala, S. (2015). Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications. Ph.D. thesis, University of Tübingen.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In J. Tetreault, J. Burstein & C. Leacock (eds.), *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*. Montréal, Canada: Association for Computational Linguistics, pp. 163–173. URL http://aclweb.org/anthology/W12-2019.pdf.

Vajjala, S. & D. Meurers (2013). On The Applicability of Readability Models to Web Texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. pp. 59–68.

Vajjala, S. & D. Meurers (2014a). Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. ACL, Gothenburg, Sweden: Association for Computational Linguistics, pp. 288–297.

Vajjala, S. & D. Meurers (2014b). Exploring Measures of "Readability" for Spoken Language: Analyzing linguistic features of subtitles to identify age-specific TV programs. In *Proceedings of the Third Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Gothenburg, Sweden: ACL, pp. 21–29.

Vajjala, S. & D. Meurers (2014c). Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification* 165(2), 142–222.

Vajjala, S. & D. Meurers (new). Readability-based Sentence Ranking for Evaluating Text Simplification. So far unpublished.

Vajjala, S., D. Meurers, A. Eitel & K. Scheiter (2016). Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on*

Exploring linguistic complexity in readability analysis & L2 development

Detmar Meurers

Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

40 / 40

Computational Linguistics for Linguistic Complexity (CL4LC). Osaka, Japan: The COLING 2016 Organizing Committee, pp. 38–48. URL http://aclweb.org/anthology/W16-4105.pdf.

Van Heuven, W. J., P. Mandera, E. Keuleers & M. Brysbaert (2014). Subtlex-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology* pp. 1–15. URL http://dx.doi.org/10.1080/17470218.2013.850521.

Wolfe-Quintero, K., S. Inagaki & H.-Y. Kim (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Zhu, Z., D. Bernhard & I. Gurevych (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING), August 2010. Beijing, China*. pp. 1353–1361.

Exploring linguistic complexity in readability analysis & L2 development

Detmar Meurers

Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

40 / 40

Sentence-level analysis using ranking

Ranking experiments on Wiki-SimpleWiki (Vajjala & Meurers new)

- Employ a Ranking algorithm, as commonly used in information retrieval to rank search results.
- Setup: SVM^{Rank} on Wiki–Simple Wiki (10-fold CV)
- Result: 82.7% accuracy
  - accuracy = percentage of correctly ranked pairs
  - baseline: 72.3% accuracy for Flesch-Kincaid formula
- Do these results generalize to other data?
  → We compiled a new data set from OneStopEnglish.com

Exploring linguistic complexity in readability analysis & L2 development

Detmar Meurers

Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School

40 / 40

Sentence-level analysis using ranking
OneStopEnglish experiments (Vajjala & Meurers new)

- Texts from *The Guardian* manually rewritten at 3 levels.
- We extracted and aligned 3113 sentences at two levels (OSE2) and 837 across three levels (OSE3), e.g.:

Adv: *In Beijing, mourners and admirers made their way to lay flowers and light candles at the Apple Store.*

Int: *In Beijing, mourners and admirers came to lay flowers and light candles at the Apple Store.*

Ele: *In Beijing, people went to the Apple Store with flowers and candles.*

Sentence-level analysis using ranking
OSE and Cross-Corpus Results (Vajjala & Meurers new)

- Defined separate train and test sets for WIKI and OSE2
- Flesch-Kincaid **baselines**:

TEST	Accuracy
WIKI	69.0%
OSE2	69.6%

- **Same-Corpus Results** of RankSVM model:

TRAIN	TEST	Accuracy
WIKI	WIKI	81.8%
OSE2	OSE2	81.5%

- **Cross-Corpus Results** of RankSVM model:

TRAIN	TEST	Accuracy
WIKI	OSE2	74.6%
OSE2	WIKI	77.5%

$\Rightarrow$ Rich features set supports reliable sentence-level analysis.

Exploring linguistic complexity in readability analysis & L2 development
Detmar Meurers

Introduction
How can we obtain evidence?
Complexity features for readability
Features from SLA research
Experimental setup
Results on WeeBit
Extending the feature set
Results on CCSS
Generalizability
From texts to sentences
Wikipedia–SimpleWiki
Multi-level evidence
Results on WeeBit
Generalizability
Ranking web search
Linking readability & L2 development
Summary
Outlook

40 / 40