

---

---

# Способы измерения лексической сложности текста

Яркова В.В.,  
стажер-исследователь научно-учебной лаборатории  
учебных корпусов НИУ ВШЭ — Пермь

---

---

# Определение лексической сложности

**lexical density**



the ratio of the number of lexical (open-class) words to the total number of words in a text

**lexical diversity**



the range of a learner's vocabulary

the type-token ratio

**lexical sophistication**



the ratio of the number of advanced words to the total number of words

# Способы измерения лексической сложности

- **Обработка вручную**
- **Использование цифровых инструментов (!)**

Coh-Metrix (Graesser et al, 2004)

**Lexical Complexity Analyzer (Ai & Lu, 2010)**

Text Inspector (Baх, 2012)

TAALED and TAALES (Kyle & Crossley, 2015)

Pitt Lexical Toolkit (Naismith, forthcoming)

Lex Complexity Tool (Bottini, forthcoming)

**Какие могут быть  
проблемы?**

Отсутствие доступа

Платный контент

Отсутствие возможности  
измерения **всех**  
**параметров** ЛС

# Lexical Complexity Analyzer

- Бесплатная и стабильная программа
- Измеряет **все 3 параметра** лексической сложности
- Можно как скачать, так и работать **онлайн** (!)



Не поддерживает ОС Windows

Нужны навыки программирования

Отсутствие автоматической обработки текстов  
(лемматизация и разметка)

# Материал для анализа

Корпус докладов русскоговорящих студентов  
с научной конференции iTELL



Материал за 2019-2020 гг.  
52 доклада (11/41)  
83508 слов

# Материал для сравнения

MICASE vs BASE



Презентации студентов (11)

Междисциплинарность

143369 слов

# Полученные результаты

Lexical density (LD)
<b>Lexical Sophistication</b>
Lexical sophistication-I (LS1)
Lexical sophistication-II (LS2)
Verb sophistication-I (VS1)
Verb sophistication-II (VS2)
Corrected VS1 (CVS1)
<b>Lexical Variation</b>
<b>NDW</b>
Number of different words (NDW)
NDW (first 50 words) (NDWZ-50)
NDW (expected random 50) (NDW-ER50)
NDW (expected sequence 50) (NDW-ES50)
<b>TTR</b>
Type/Token ratio (TTR)
Mean Segmental TTR (50) (MSTTR-50)
Corrected TTR (CTTR)
Root TTR (RTTR)
Bilogarithmic TTR (logTTR)
Uber Index (Uber)
<b>Verb diversity</b>
Verb variation-I (VV1)
Squared VV1 (SVV1)
Corrected VV1 (CVV1)
<b>Lexical word diversity</b>
Lexical word variation (LV)
Verb variation-II (VV2)
Noun variation (NV)
Adjective variation (AdjV)
Adverb variation (AdvV)
Modifier variation (ModV)

- **Lexical Density Values (Mean)**

iTELL 0.487

MICASE 0.483

- **Lexical Diversity Values, NDW (Mean)**

	NDW	NDW	NDWERZ	NDWESZ
iTELL	422.673	36.942	37.471	35.13
MICASE	1027.565	38.130	40.352	36.591

# Полученные результаты

- Lexical density (LD)
- Lexical Sophistication**
  - Lexical sophistication-I (LS1)
  - Lexical sophistication-II (LS2)
  - Verb sophistication-I (VS1)
  - Verb sophistication-II (VS2)
  - Corrected VS1 (CVS1)
- Lexical Variation**
  - NDW**
    - Number of different words (NDW)
    - NDW (first 50 words) (NDWZ-50)
    - NDW (expected random 50) (NDW-ER50)
    - NDW (expected sequence 50) (NDW-ES50)
  - TTR**
    - Type/Token ratio (TTR)
    - Mean Segmental TTR (50) (MSTTR-50)
    - Corrected TTR (CTTR)
    - Root TTR (RTTR)
    - Bilogarithmic TTR (logTTR)
    - Uber Index (Uber)
  - Verb diversity**
    - Verb variation-I (VV1)
    - Squared VV1 (SVV1)
    - Corrected VV1 (CVV1)
  - Lexical word diversity**
    - Lexical word variation (LV)
    - Verb variation-II (VV2)
    - Noun variation (NV)
    - Adjective variation (AdjV)
    - Adverb variation (AdvV)
    - Modifier variation (ModV)

- **Lexical Diversity Values, TTR (Mean)**

	TTR	MSTTR	CTTR	RTTR	LOGTTR	UBER
iTELL	0.245	0.706	7.109	10.05	0.81	17.13
MICASE	0.146	0.726	8.54	12.078	0.781	17.588

- **Lexical Diversity, VD (Mean)**

	VV1	SVV1	CVV1
iTELL	41.408	4.516	0.408
MICASE	50.529	4.966	0.263



# Полученные результаты

- Lexical density (LD)
- Lexical Sophistication
  - Lexical sophistication-I (LS1)
  - Lexical sophistication-II (LS2)
  - Verb sophistication-I (VS1)
  - Verb sophistication-II (VS2)
  - Corrected VS1 (CVS1)
- Lexical Variation
  - NDW
    - Number of different words (NDW)
    - NDW (first 50 words) (NDWZ-50)
    - NDW (expected random 50) (NDW-ER50)
    - NDW (expected sequence 50) (NDW-ES50)
  - TTR
    - Type/Token ratio (TTR)
    - Mean Segmental TTR (50) (MSTTR-50)
    - Corrected TTR (CTTR)
    - Root TTR (RTTR)
    - Bilogarithmic TTR (logTTR)
    - Uber Index (Uber)
  - Verb diversity
    - Verb variation-I (VV1)
    - Squared VV1 (SVV1)
    - Corrected VV1 (CVV1)
  - Lexical word diversity
    - Lexical word variation (LV)
    - Verb variation-II (VV2)
    - Noun variation (NV)
    - Adjective variation (AdjV)
    - Adverb variation (AdvV)
    - Modifier variation (ModV)

- **Lexical Diversity Values, LWD (Mean)**

	LV	VV2	NV	ADJV	ADV	MODV
iTELL	0.525	0.095	0.368	0.082	0.039	0.12
MICASE	0.252	0.057	0.296	0.052	0.028	0.08

- **Lexical Sophistication Values (Mean)**

	LS1	LS2	VS1	VS2	CVS1
iTELL	0.312	0.304	0.109	2.004	0.952
MICASE	0.286	0.39	0.06	3.11	1.2

# Перспективы исследования

- Изучение лексической сложности в динамике
- Совершенствование педагогических методик (как для неязыковых, так и языковых направлений подготовки)
- Изучение особенностей устной академической речи, создание инструмента для анализа

**Спасибо за внимание!**