Introducing Multi-Word Units Profiler: Background, development, and outlook

Masaki Eguchi (江口 政貴)

Learner Corpus Research and Applied Data Science Lab

Linguistics Department, University of Oregon

11/27/2021 @ Methodology Special Interest Group, Kansai Chapter, Language Education and Technology (LET)

UNIVERSITY OF OREGON

Brief Self-Introduction

- PhD Candidate
- Linguistic dep, Univ. of Oregon
- SLA, NLP, SFL, L2 Vocabulary, Productive skills

- Learner Corpus Research + Applied Data Science Lab
  - Directed by Dr. Kris Kyle
  - We are working on easy-to-use webapp versions of:

TAALED TAALES TAASSC

Today's content

- Backgrounds and motivation
- Key features of the tools
- Quick demo video
- Methodological issues and tentative solutions in identifying MWUs
- Outlook (How do tools facilitate research and pedagogy?)

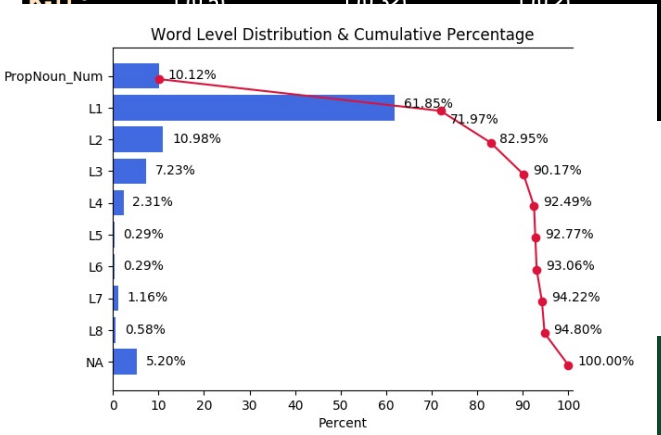
Computer assisted lexical analysis

Successful applications of lexical analysis for research and pedagogy.

- Range program (Prof. Nation)
- VocabProfiler (Prof. Cobb)
- AntWordProfiler (Prof. Anthony)
- New Word Level Checker (Prof. Mizumoto)

- Profiling user-input texts:
  - how many words learners need to know
  - "difficult" words for students
  - how many times target words occurs in the text.
  - See also Nation & Webb (2008)

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token (%)
K-1 :	118 (54.6)	141 (45.19)	389 (58.4)	58.4
K-2 :	36 (16.7)	40 (12.82)	53 (8.0)	66.4
K-3 :	36 (16.7)	37 (11.86)	49 (7.4)	73.8
K-4 :	6 (2.8)	6 (1.92)	8 (1.2)	75.0
K-5 :	4 (1.9)	4 (1.28)	5 (0.8)	75.8
K-6 :	6 (2.8)	7 (2.24)	11 (1.7)	77.5
K-7 :	2 (0.9)	2 (0.64)	2 (0.3)	77.8
K-8 :	3 (1.4)	3 (0.96)	3 (0.5)	78.3
K-9 :	2 (0.9)	2 (0.64)	2 (0.3)	78.6
K-10 :				
K-11 :	1 (0.5)	1 (0.32)	1 (0.2)	78.8
				79.0



Multiword units

- Multi-Word Units (here): "(semi-)fixed, recurrent phrases" (Siyanova-Chanturia & Martinez, 2014, p.549)
  - **Collocations**: *strong tea*, *meet + expectation*
  - **Multi-word verbs**: *put up with*
  - **Lexical bundles**: *in the middle of*
  - Binomials: *black and white*
  - Idioms: *spill the beans*

Current foci

Multiword Units (MWUs):

- are ubiquitous and prevalent in language use (e.g., Erman & Warren, 2000).
- are important means for socialization (e.g., Burdelski & Cook, 2012).
- can provide processing short-cuts (e.g., Siyanova & Schmitt, 2008)
- are found to contribute to L2 skills (Reading, Oral proficiency, writing, etc.)

(Kremmel et al., 2017; Uchihara et al, 2021; Kyle & Eguchi, 2021)

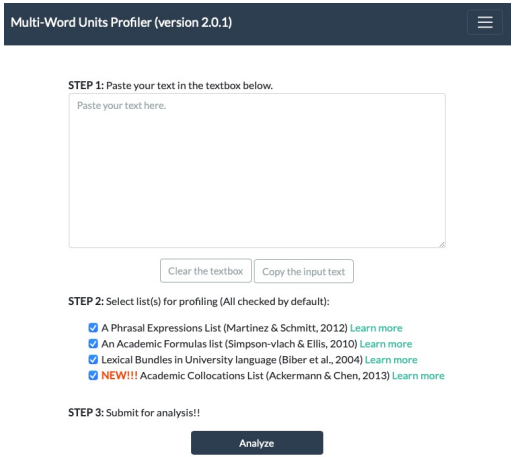
Learning MWU can be challenging

- MWU can be obstacles in L2 learning
  - Even highly proficient learners struggle (Durrant & Schmitt, 2009; Nesselfauf, 2003)

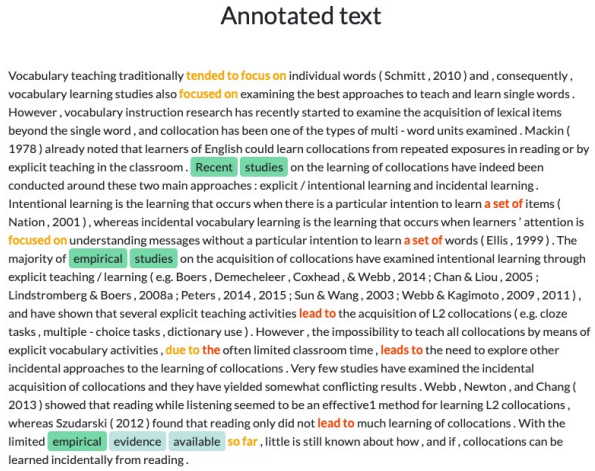
- Less favorable conditions for incidental learning (see Boers, 2021)
  - MWUs are less frequent than single words.
  - MWUs may be less perceptually salient
  - Learners may not recognize usefulness
  - Some MWUs are incongruent (for speakers of a certain L1)
- How can we aid learning of MWU?

Multi-Word Units Profiler

- A freely available web application for pedagogy (and research)
- Capability to identify and highlight research-based MWUs in user-input texts.



Input/settings



Annotation view

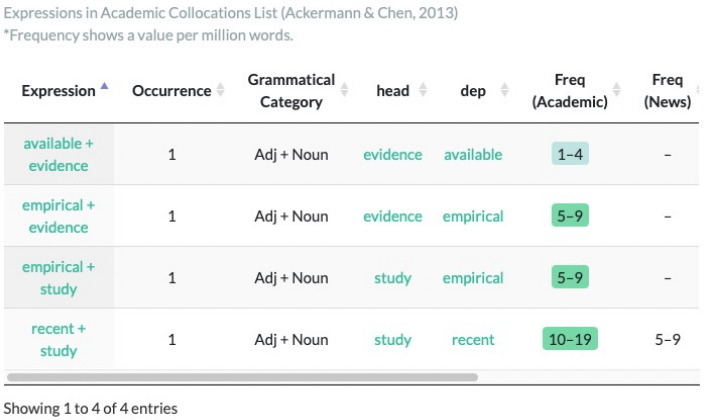


Table view

Quick demo

Multi-Word Units Profiler (version 2.0.1)

STEP 1: Paste your text in the textbox below.

Paste your text here.

Clear the textbox | Copy the input text

STEP 2: Select list(s) for profiling (All checked by default):

- ☑ A Phrasal Expressions List (Martinez & Schmitt, 2012) Learn more
- ☑ An Academic Formulas list (Simpson-vlach & Ellis, 2010) Learn more
- ☑ Lexical Bundles in University language (Biber et al., 2004) Learn more
- ☑ NEW!!! Academic Collocations List (Ackermann & Chen, 2013) Learn more

STEP 3: Submit for analysis!!

Analyze

Annotated text

List(s) of multi-word expressions

Summary: MWU profiler

- can identify, and highlight MWUs in user-input texts
- draws on several empirically based MWU lists
  - PHRASE list (Martinez & Schmitt, 2012)
  - Academic Formulas list (Simpson-Vlach & Ellis, 2010)
  - Academic Collocations List (ACL; Ackermann & Chen, 2013)
  - Etc.
- provides item-specific information in a table format
- provides hyper links to other corpus-based resources
  - Web Concordancer (by Prof. Tom Cobb)
  - Sketch Engine for Language Learning, or SKELL (by Sketch Engine)

Research Background

1) Intervention approaches to teach MWUs
2) Corpus-driven MWU lists

1) Intervention approach to MWU learning

- Research on the effects of systematic intervention on MWU learning is emerging.

- Teacher-led **chunk discovery activity** (Boers et al., 2006)
  - Exp outperformed in oral interviews: overall, fluency, ranges of expressions
    - (medium to large effect sizes; *d* = .91 – 1.28)
- **Textual enhancement** on retention of MWUs (Choi, 2017)
  - Enhancement led to longer reading time on target collocations (*d* = 1.127)
  - Enhanced group recalled 39% more collocations than control group.
  - Trade-off between retention of enhanced collocation vs. unenhanced text

2) Corpus-driven MWU lists

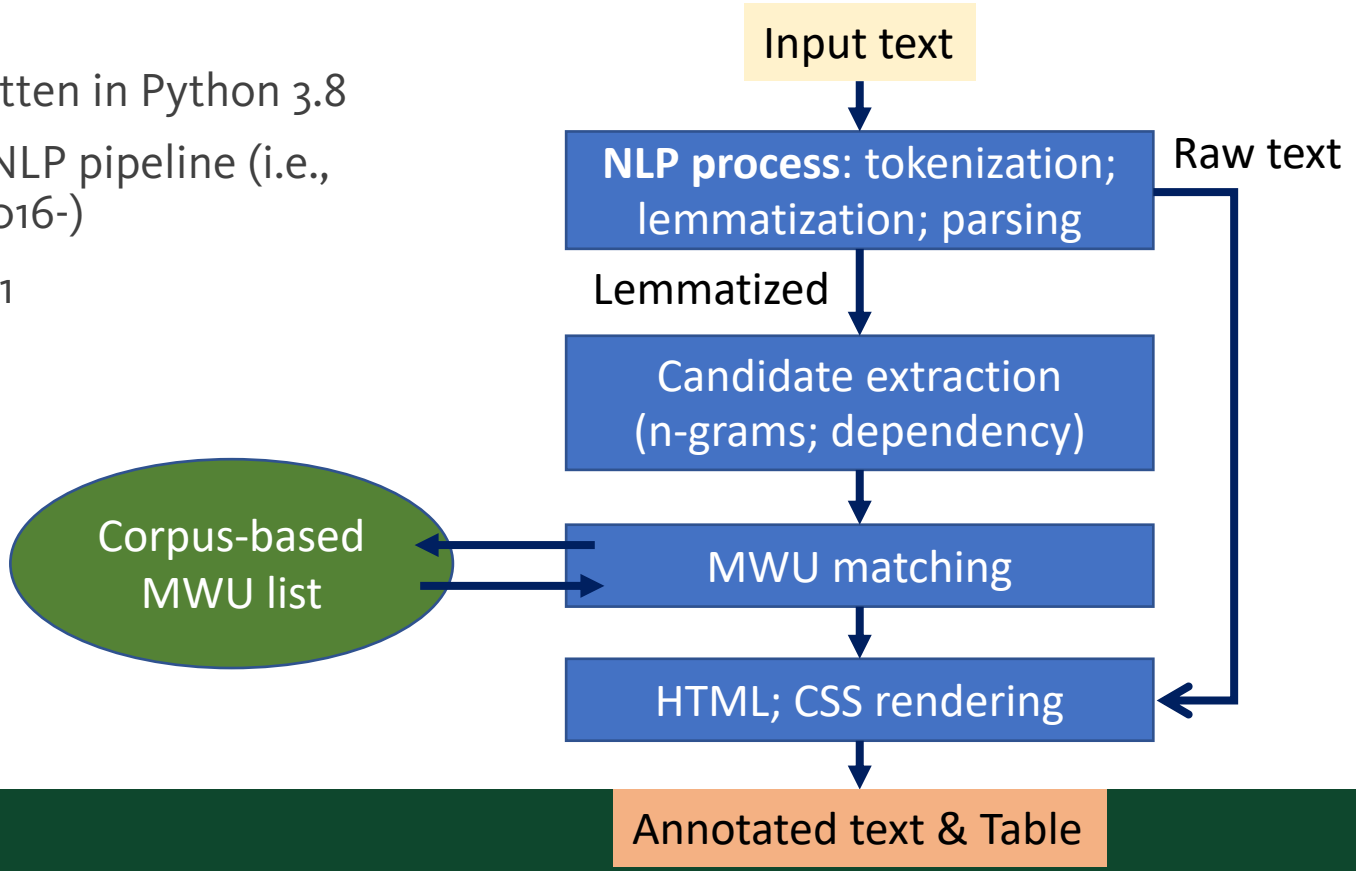
- What are the “important”, “pedagogically useful” MWUs?
  - It depends (e.g., English for Specific Purposes; genres, registers)
- Developments of MWU lists
  - PHRASal Expressions (PHRASE) list (Martinez & Schmitt, 2012)
  - Academic Formulas List (Simpson-Vlach & Ellis, 2010)
  - Academic Collocations List (ACL; Ackermann & Chen, 2013)
- Corpus-driven frequency list + expert judgements
  - such as, a number of, apart from, set out, account for
  - the ability to, in terms of a, a list of, a variety of
  - anecdotal + evidence, gather + information, seem + plausible, strongly + agree, highly + controversial

What does MWU profiler do behind the scene?

NLP pipeline for MWU analysis

The NLP pipeline for MWU profiler

- MWU profiler is written in Python 3.8
- Leverages modern NLP pipeline (i.e., spaCy, Explosion, 2016-)
- spaCy 2.3 in ver 2.0.1



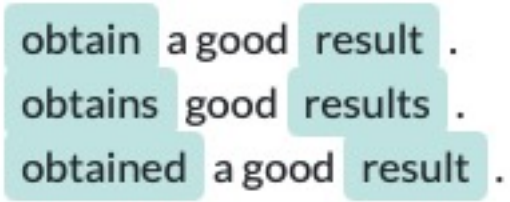
1) Tokenization and lemmatization

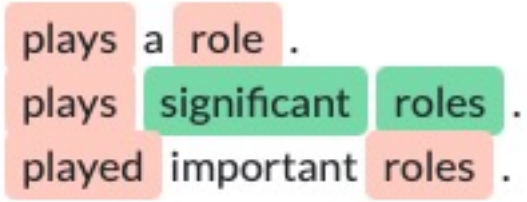
Tokenization: separating text into sentences and words.

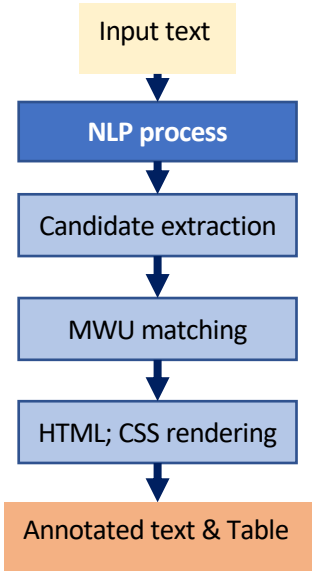
Lemmatization: converting inflectional form into the base form

- viewed, views, viewing > view viewer, viewers > viewer

- **Issue #1**: MWUs can occur in various inflected forms.
  - ✅ MWU profiler uses spaCy lemmatizer to overcome this:







2) Candidate extraction: N-gram

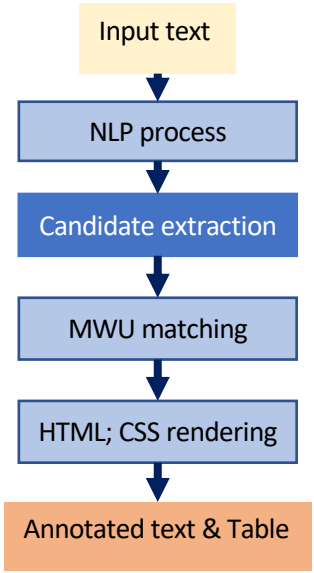
N-grams (adjacent sequence of *n*-words)

- 2-4 grams are extracted from the text and stored for item matching.

> English as we know it today came to be exported to other parts of the world through British colonisation, and is now the dominant language in Britain and Ireland, the United States and Canada, Australia, New Zealand and many smaller former colonies, as well as being widely spoken in India, parts of Africa, and elsewhere. Partially due to influence of the United States and its globalized

Bigrams ➔ English as/ as we/ we know …

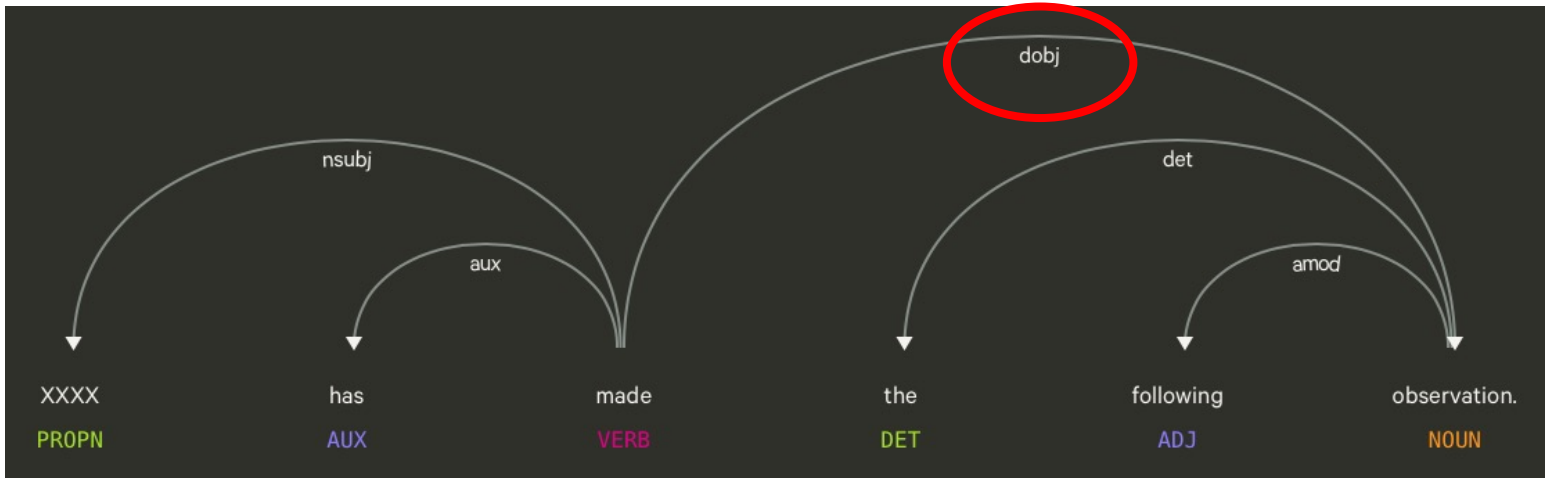
Trigrams ➔ English as we/ as we know/ we know …

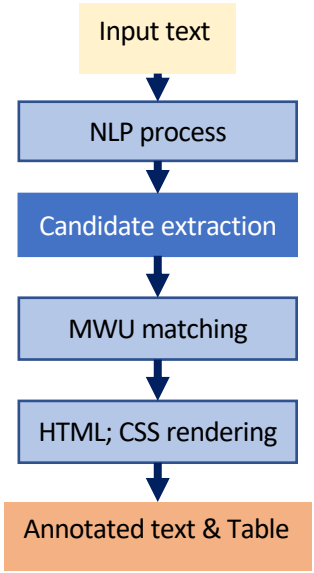


2) Candidate extraction: Dependency parsing

Process of identifying two words in a direct syntactic relation

- e.g., make + observation (dobj)



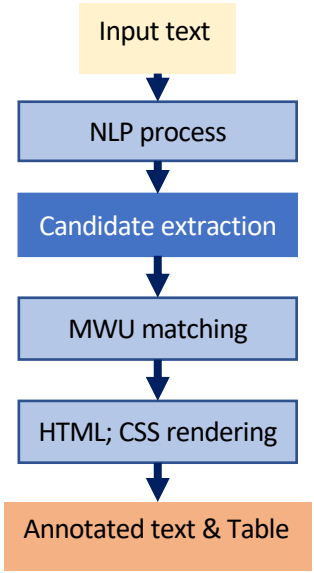


You can access demo from https://explosion.ai/demos/displacy

Dependency captures dislocated collocations

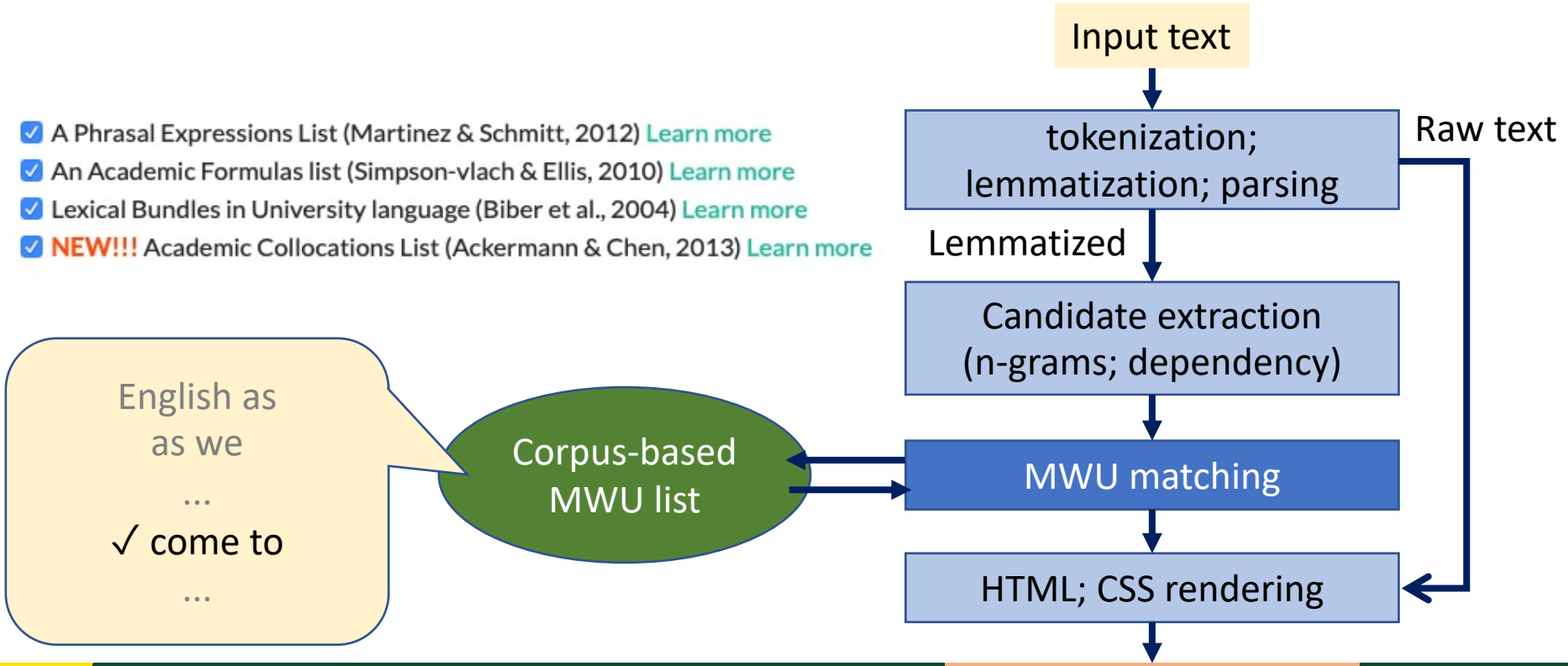
- **Issue #2**: Collocations can be dislocated (with intervening words).
- ✅ spaCy dependency parser is used to identify dislocated collocations.
- Does **NOT** have to specify arbitrary window sizes (e.g., +/- 4 words)

XXXX has made an observation .
XXXX has made the following observation .
XXXX has made the following pertinent observation .



3) MWU matching

- ☑ A Phrasal Expressions List (Martinez & Schmitt, 2012) Learn more
- ☑ An Academic Formulas list (Simpson-vlach & Ellis, 2010) Learn more
- ☑ Lexical Bundles in University language (Biber et al., 2004) Learn more
- ☑ NEW!!! Academic Collocations List (Ackermann & Chen, 2013) Learn more



Methodological issues & tentative solutions

Issues	Example	✅Solutions & ❑ development plans
Spelling variants	"widely recognised" vs "widely recognized"	✅ Spelling converted when items are matched
Grammatical inflection	MWUs play an important role. English plays an important role.	✅ spaCy lemmatizer is used
Dislocated collocations	"**obtain** the **results**" "**obtain** the best possible **results**"	✅ spaCy dependency parser is used
Syntactic variants	"**results** were **obtained**"	❑ passive construction (e.g., nsubjpass tag) should be identified in the future versions
Type specific MWUs	"in relation to" **but not** "in relations to"	❑ Specify which MWUs should be type specific e.g., using type specificity in a large-scale corpus

Outlook

Development plans and more...

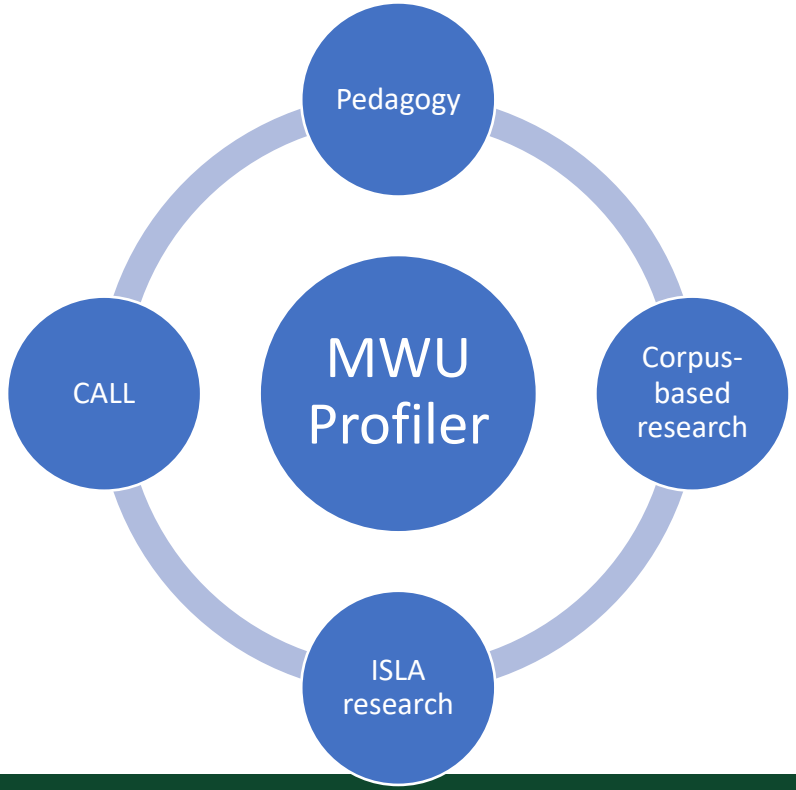
Development plans

- Fine-tuning NLP pipeline for better accuracy
  - See previous slides for issues to tackle
  - Simple n-gram analysis may identify false positives. (“at all times” > “at all times”)
  - Empirical evaluation of the current (and future) pipeline against expert manual annotations
- Adding more learner-centered features
  - Providing gloss (translation)? If so, how (e.g., machine translation…)?
- Adding, developing, and refining MWU lists
  - More purpose-specific MWU lists (e.g., target domains, genre and register, skill areas)
  - More easy-to-understand frequency information
- Empirical study on the effects of textual enhancement through MWU profiler
  - Replication of Choi (2017) or Boers et al (2006), etc.

Facilitating research–pedagogy interaction

MWU profiler can:

- be used in teaching/learning
- be used to create materials for ISLA research
- facilitate easier replication of research on textual enhancement
- provide a quick demo to implement a newly developed corpus-based MWU lists



How to cite MWU profiler:

- Eguchi, M. (2021). *Multi-Word Units Profiler* (Version 2.0.1) [Computer software]. https://multiwordunitsprofiler.pythonanywhere.com

Also, please see and try:

- Phrase profiler by Prof. Tom Cobb (https://www.lextutor.ca/vp/collocs/)

To learn more about dependency collocations

- Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using words, n-gram, and dependency bigram indices. In S. Granger (Ed.), *Perspectives on the Second Language Phrasicon: The View from Learner Corpora.* Multiligual Matters.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Anthony, L. (2021). *AntWordProfiler* (1.5.1) [Computer software]. https://www.laurenceanthony.net/software

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a Lexical Approach to the test. *Language Teaching Research*, *10*(3), 245–261. https://doi.org/10.1191/1362168806lr195oa

Burdelski, M., & Cook, H. M. (2012). Formulaic Language in Language Socialization. *Annual Review of Applied Linguistics*, *32*, 173–188. https://doi.org/10.1017/S0267190512000049

References

Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research*, *21*(3), 403–426. https://doi.org/10.1177/1362168816653271

Cobb, T. (2021). *Compleat Web VP* (2.5) [Computer software]. https://www.lextutor.ca/vp/comp/

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text - Interdisciplinary Journal for the Study of Discourse*, *20*(1). https://doi.org/10.1515/text.1.2000.20.1.29

Explosion AI. (2016). *SpaCy* (2.0) [Computer software]. https://spacy.io

References

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the Role of Phraseological Knowledge in Foreign Language Reading. *Applied Linguistics*, amv070. https://doi.org/10.1093/applin/amv070

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using words, n-gram, and dependency bigram indices. In S. Granger (Ed.), *Perspectives on the Second Language Phrasicon: The View from Learner Corpora.* Multiligual Matters.

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics*, *33*(3), 299–320. https://doi.org/10.1093/applin/ams010

Mizumoto, A. (2021). *New Word Level Checker*. https://nwlc.pythonanywhere.com/

References

Nation, I. S. P., & Heatley, A. (2002). *Range: A program for the analysis of vocabulary in texts*. http://www.vuw.ac.nz/lals/staff/paulnation/nation.aspx

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, *31*(4), 487–512. https://doi.org/10.1093/applin/amp058

Siyanova, A., & Schmitt, N. (2008). L2 Learner Production and Processing of Collocation: A Multi-study Perspective. *Canadian Modern Language Review*, *64*(3), 429–458. https://doi.org/10.3138/cmlr.64.3.429

References

Siyanova-Chanturia, A., & Martinez, R. (2014). The Idiom Principle Revisited. *Applied Linguistics*, amt054. https://doi.org/10.1093/applin/amt054

Uchihara, T., Eguchi, M., Clenton, J., Kyle, K., & Saito, K. (2021). To What Extent is Collocation Knowledge Associated with Oral Proficiency? A Corpus-Based Approach to Word Association. *Language and Speech*, 002383092110138. https://doi.org/10.1177/00238309211013865

Extra slides

MWU matching with research-based lists

PHRASE list
(Martinez & Schmitt, 2012)

Scope: frequent MWU list for receptive skills
Corpus: BNC (100M; Spoken+Written)
Method: Corpus frequency + Qual
Criteria: freq + meaningful + semantically less transparent
Examples: *such as, a number of, apart from, set out, account for*

Academic Formulas List
(Simpson-Vlach & Ellis, 2010)

Scope: useful MWUs for academic speech and writing
Corpus: 2.1M words (MICASE + Hyland's (2007) research paper corpus + BNC)
Method: Corpus frequency + MI + Qual
➔ freq + MI predicted teacher's 'usefulness' rating
Examples: *the ability to, in terms of a, a list of, a variety of*

If you have any recommendations for addition, please please let me know.

MWU draws on existing published MWU lists

Academic Collocations List (Ackermann & Chen, 2013)

Scope: Pedagogically relevant lexical collocations for EAP

Corpus: the Pearson International Corpus of Academic English (37M words; lecture, textbook, journal papers)

Criteria: Corpus freq (≥ 1 per M) /MI (≥ 3) / T-score (≥ 4)

Method: Frequency analysis → Manual exclusion → Expert judgement on inclusion (Likert scale) → 2468 High-scoring collocations

Examples: *anecdotal + evidence*, *gather + information*, *seem + plausible*, *strongly + agree*, *highly + controversial*

If you have any recommendations for addition, please please let me know.