Corpus linguistics

Raksangob Wijitsopon and Richard Watson Todd

TAAL AL Day 2023

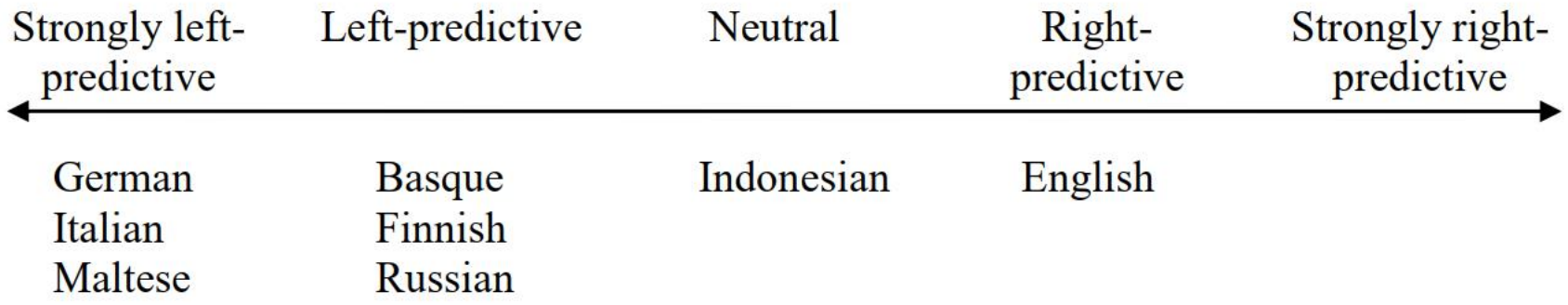
Overview of uses of corpus linguistics

- Theoretical
- Descriptive
- Interpretive
- Critical
- Pedagogic
- Combined with QUAL

- Directions of collocations in 8 languages
- Distinguishing features in the sermons of cult leaders
- Literature
- LGBT news
- Word lists
- Data-driven learning
- The language of textbooks
- Learner Interlanguage
- Social media at different levels

Theoretical corpus linguistics

- Right-predictive collocations e.g. *Pyrrhic victory* where the first word predicts the second
- Left-predictive collocations e.g. *deadly nightshade* where the second word predicts the first
- Use ΔP
- Analyse all collocations in 8 corpora of 8 languages



Watson Todd, R. (2019). Exploring the direction of collocations in eight languages. *Canadian Journal of Linguistics/Revue Canadienne De Linguistique, 64*(1).

Descriptive corpus linguistics

- Types of religious groups

Mainstream ⟷ Sect ⟷ Cult ⟷ Destructive cult

- Analyse sermons of religious group leaders for
  - Keywords
  - Key parts of speech
  - Key semantic tags
- Features associated with sermons of cults and destructive cults
  - Non-religious topics (e.g. politics)
  - Language of othering (separating *we* and *they*)
  - Intensifying language
  - Elaborative language

Palayon, R. T., Watson Todd, R. and Vungthong, S. (2022) From the temple of life to the temple of death: keyness analyses of the transitions of a cult. Corpora 17(3).

Interpretive corpus linguistics

- Relationship between language use in a text & interpretations/ effects of it on readers
  - “linguistic features whose densities in a text are appreciably different from those found in its contextually related norm” (Enkvist 1973: 21)
- Different densities => Finding keywords: A case study of *The Great Gatsby*
  - GG vs. AmFic & FitzFic (Contextually related norms)
- Colour words making up 4 of 22 keywords in GG (Text)
- Colour symbolism (Interpretation)
- Finding co-occurrence patterns of key colour words
  - ‘white’ + the heroine, luxury & filth (textual patterns)
  - The irony of ‘white’ (interpretation)
- CL findings provides or adds empirical textual evidence to existing interpretative remarks

Wijitsopon, R. (2022) Corpus stylistics and colour symbolism in *The Great Gatsby* and its Thai translations. *Language and Literature* 31(3): 267–295.

Critical corpus linguistics

- Dialectical relationship between language in use and socio-cultural/ political/ economic structures
- A case study: LGBT representation in Thailand's and international newspapers
- Using both self-compiled and existing newspaper corpora (COCA & SiBol)
- Collocation Analysis of LESBIAN, GAY, BISEXUAL, TRANSGENDER & LGBT
- Similar collocates
  - LGBT movement for equal rights in various social domains
  - Association of gays with HIV
  - Association of LGBT with crime.
    - => Reflection and construction of LGBT movement as a global phenomenon
    - => Naturalization of LGBT and HIV infection or criminal behavior => social stigmatization or marginalization (Singhakowinta 2014)
- Different collocates:
- beauty contests => socio-cultural factors in news discourse => beauty is power (cf. Poompruek et al., 2014)
  - This ideology dialectically reflected and promoted discursively through the positive semantic prosody of the collocates related to beauty contests of TRANSGENDER in BP.
- Underrepresentation of disclosure/ openness of sexual orientation

Chuaikun, D. & Wijitsopon, R. (2021) A corpus-based study of LGBT-related news discourse in Thailand's and international English-language newspapers. Applied Linguistics Review. https://doi.org/10.1515/applirev-2021-0036

Pedagogical corpus linguistics 1: Word lists

- What words is it most useful for a teacher to focus on in the classroom?
- Words students cannot deal with autonomously
- Polysemous words where the meaning required is not the usual meaning (opaque words)
- Corpus of textbooks used by engineering undergraduates
- Rate keywords for opaqueness in corpus
- E.g. *constant* (fixed number), *given* (nominated adj.)
- List of 40 opaque words to teach

Watson Todd, R. (2017) An opaque engineering word list: Which words should a teacher focus on? English for Specific Purposes 45.

Pedagogical corpus linguistics 2: Data-driven learning

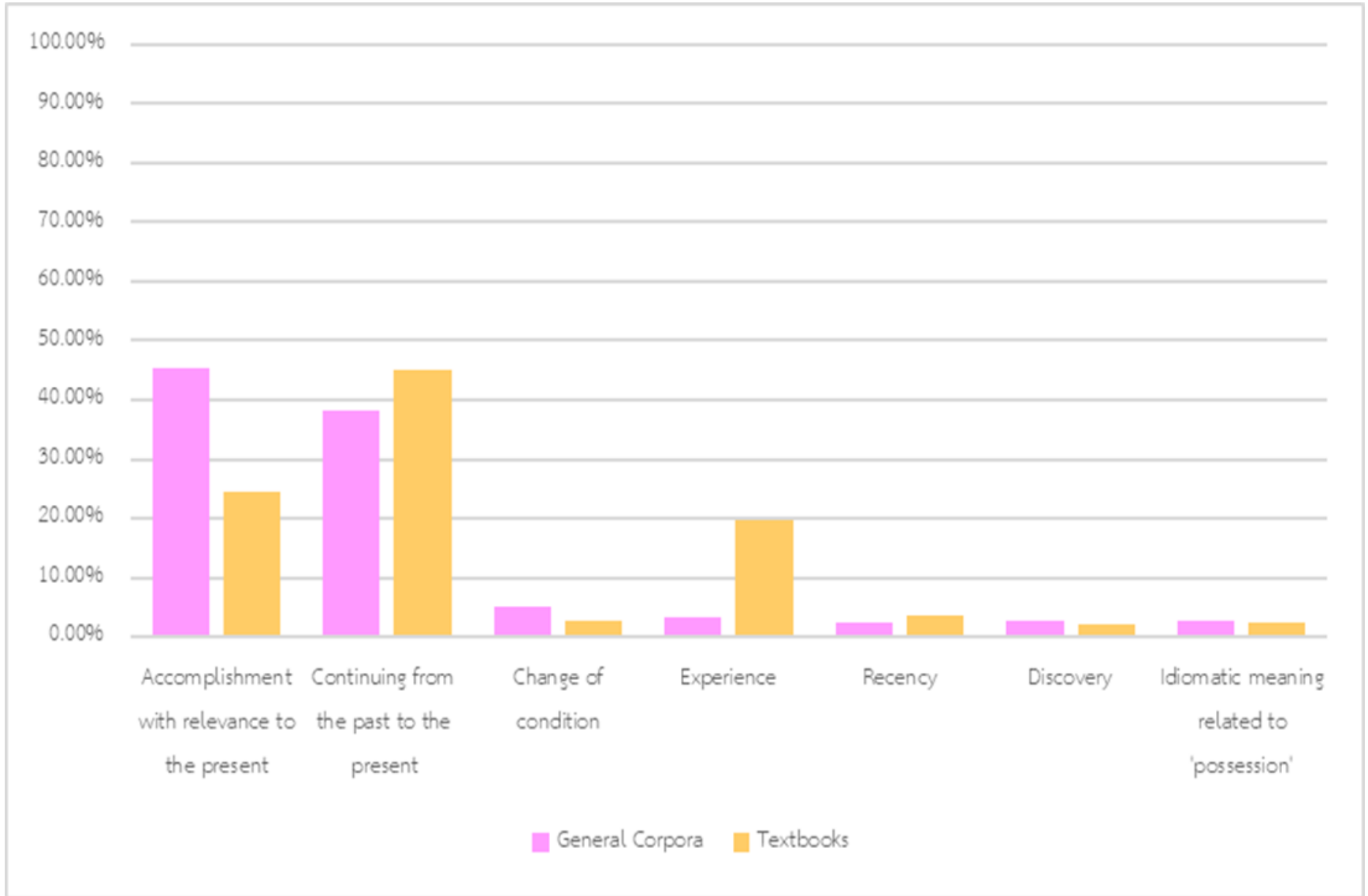
- The teacher identifies words needing correction
- Students generate concordance lines for their words
- Students induce patterns of use
- Students apply these patterns to self-correct their writing
- Students could self-correct in about 80% of cases
- Adjectives are the easiest word class for inductions and self-corrections
- There is a significant negative correlation between number of possible meanings of a word and ability to self-correct

Watson Todd, R. 2001 Induction from self-selected concordances and self-correction. System 29(1).

Pedagogical corpus linguistics 3: Textbooks

- Present perfect meanings in real use vs. taught in textbooks
- HAVE + past participle verbs of different kinds => Meanings of PP
- BrE2006 + AmE2006 + a self-compiled corpus of upper-intermediate textbooks
- Correlations and disparities in major uses of PP
  - Real use: Accomplished > Continuing frm past to prsnt > Changes
  - Textbooks: Continuing frm past to prsnt > Accomplished > Experience
- 2 core meanings taught but highlighted differently from real use.
- 1 major use in real life under-presented in textbooks
- 1 minor use in real life over-presented in textbooks

Charoenkul, C. & Wijitsopon, R. (2020) The English Present Perfect in Authentic Use and Textbooks: A Corpus-driven Study. *PASAA* Vol. 60: 275-308.


100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Accomplishment with relevance to the present
Continuing from the past to the present
Change of condition
Experience
Recency
Discovery
Idiomatic meaning related to 'possession'
General Corpora
Textbooks

Pedagogical corpus linguistics 4: Learner interlanguage

- Thai A & B+ learner English vs. Native speaker learner’
- Key clusters => Fomulaicity => Fluency
- Thai learner corpus vs. Native speaker learner corpus vs. COCA
- Ideational clusters:
  - Spoken > Written (Large quantities dominant)
- Interpersonal clusters:
  - Spoken > Written (Boosters/ categorical)
- Textual clusters (Largest group of key clusters)
  - Written > Spoken (transitional words)
- Suggested improvements:
  - Awareness of registers
  - Hedging
  - Support strategies other than referring to a large number

Wijitsopon, R. (2019) Key multi-word expressions in Thai learner English argumentative essays. Asian EFL Journal: 115-141.

Combining corpus linguistics and qualitative analyses

- Investigating a discussion forum on a controversial topic
- 2010 political unrest
- Categorise postings as red, yellow or neutral
- To understand the whole forum, conduct a keyword analysis of red postings versus yellow postings
- Both sides talk about the other side more than themselves
- Yellows used more words with strong affective connotations
- To understand how the discourse progresses, conduct qualitative analyses of threads

Jimarkon, P. and Watson Todd, R. (2013) Red or yellow, peace or war: Agonism and antagonism in online discussion during the 2010 political unrest in Thailand. In De Rycker, A. and Mohd Don, Z. (eds.) Discourse and Crisis: Critical Perspectives (pp. 301-322). Amsterdam: John Benjamins.

Methodology in corpus linguistics

- Keyword analysis as an example
- Comparing a target corpus with a benchmark corpus to identify ‘important’ words in the target corpus
    - Choose the target corpus
    - Choose the benchmark corpus
    - Formatting the target and benchmark corpora
    - Setting minimum frequency and dispersion criteria
    - Choosing the keyness statistic
    - Setting the cutoff threshold
    - Creating the keyword list
    - Checking for anomalies
    - Interpreting the keyword list

Methodology in corpus linguistics

- Collocation analysis
- Extracting lexical items that co-occur significantly with a target word
    - Signficantly: co-occurrences appear with greater than random probability in its (textual) context
        - Choose a target word/ a set of target words
        - Choose the co-occurrence span: 3/ 4/ 5 words to the left and right of the target word
        - Setting minimum frequency
        - Choosing the statistics, e.g. Mutual Information (MI), Log-Ratio, Log-Likelihood, Z-Score, etc.
        - Setting off the cutoff threshold, e.g. MI score ≥ 3
        - Interpreting the collocational relationship
- Concordance analysis

Collocational group	BP	SiBol and COCA
LGBT term	*bisexual, lesbian, intersex, transgender LGBTI, LGBT*	*lesbian(s), bisexual(s), transgendered, transgender*
Family and relationship	*couple, couples, marriage*	*lover, marriage(s), couples, weddings, relationships, adoption(s), unmarried*
Religion	*monks*	*bishop(s), clergy, ordination*
Law and politics	*law, rights*	*activists, rights, equality, advocate(s), Stonewall, movement, liberation, legislation, unions, feminist(s)*
Masculinity	*men, man*	*men, male*
Group of people	*community, people*	*community*
Health & body	*blood*	*HIV-positive*
Entertainment	*magazine*	*Parade(s), pornography, bar(s), club(s), nightclub, magazine, scene, characters, fantasia*
City, country & nationality	*Thailand, Bangkok, Spanish, American, Thai*	–
Heterosexuality	–	*Heterosexual(s), straights*
Disclosure/ openness	–	*openly, closeted*

phrase frequently co-occurs with percentile figures, such as *29* and *100%*. This creates a scientific or academic style in the news reports, thereby projecting an impression of factuality and reliability of information in news about gay men with HIV. In some cases, though without numbers, expressions about a high number and an increase, e.g. *infection rates are still rising* and *HIV rate on rise among gay men*, are used to

Society has been condemned for not accepting "	gay	's blood". The video "Thai Red Cross
"Thai Red Cross Society does not accept	gay	's blood", posted to YouTube by a
labour market, with 77% of transgenders, 49% of	gay men	and 62.5% of lesbians stating a belief
uccessful, infection rates are still rising among	gay men	and teenagers, Thomas Davin, a representative
last year were aged 25 or younger, while	gay men	and transsexuals accounted for 40% of the
activities. In Chiang Mai, nine couples of	gay men	and women and transgender people video clips of
quickly". He said this has resulted in	gay men	being particularly at risk of contracting
last year released by the center, 29% of	gay men	in Bangkok are HIV-infected. The
100% of the infected blood was donated by	gay men	, she said. Dr Soisaang said the
heir sexual preference. Some 24% of lesbians and	gay men	were told not to show or
HIV infection rate on rise among	gay men	World AIDS Day yesterday was greeted suit against

Figure 5: Concordance line samples of *gay* in association with HIV infection in BP.

Corpus linguistics research in Thailand

- Synonyms
- Thai Learner interlanguage
  - Thai vs. Native speaker (learners)
  - Thai learners of different proficiency levels
- Discourse
  - An approach to discourse studies: A corpus of texts as a starting point
  - Discourse features: A linguistic feature in texts as a starting point
- Translation
  - English – Thai lexical equivalents

Trends and gaps

- Moving beyond words: multimodal corpora
  - Corpora of annotated video
  - Corpora of sign language
  - Corpora of visual texts
- MCAT: corpus of 1,500 online adverts analysed on 12 categories (e.g. dominant colour, objects, alignment)
- Adverts are most commonly verbal content left – picture right
- Design differences between 1. adverts for everyday products, and 2. adverts for services and expensive products

Watson Todd, R., Vungthong, S., Trakulkasemsuk, W., Pojanapunya, P. and Towns, S. G. (2023) Love is blue: Designing and using a multimodal corpus analysis tool. *Applied Corpus Linguistics* 3(1).

Trends and gaps

- Integrating CL with other fields/ methods
  - Law & CL conference
  - CL & health care
  - CL in digital humanities (history, literature, film studies, etc.)
  - CL & psycholinguistic methods
- Contrastive linguistic/ discourse analysis
  - Development of corpora of languages other than English + processing software
  - But Thai : (
- Diachronic analysis
  - Digitisation of historical texts

Struggles and challenges

- Analysing Thai, especially Thai social media
  - Word segmentation
  - Spelling variants in social media
  - Paucity of benchmark resources
- Compiling Thai L2 English learner corpora
  - Research ethics review for research involving human
- Availability of open-access data

The future

The view from the top of the mountain was truly awesome.
That concert last night was awesome - the band really rocked it.
I just tried this new restaurant and the food was awesome.
My friend gave me an awesome birthday present - I can't wait to use it.
I had an awesome time on my vacation - I didn't want to come back.

- AI vs. corpora
  - Language AI based on massive corpora
  - Pre-analysed through machine learning
- Generating concordance lines using ChatGPT
- Identifying common collocates using ChatGPT

Based on a corpus of written texts, some words that commonly appear within three words to the left of the word "pretty" are:

- "very"
- "so"
- "quite"
- "rather"
- "damn"

The future

- Innovations in corpus linguistics
- Programmer corpus linguist/ corpus linguist programmer