Methodological issues and challenges in the use of phrase-frames to investigate phraseology

Dr. Aysel Sahin Kızıl, İzmir Bakırçay University

Dr. Lee McCallum , Edinburgh Napier University

Dr. Benet Vincent , Coventry University

Outline

- Aims of our project
- The corpus
- Phraseology and p-frames background
- Methodological issues
  - P-frame extraction
  - Manual filtering procedures
- Conclusions

Aims of our project (CHSRA)

identify recurrent phrase-frames in introductions of RAs in Health Sciences

Compile a pedagogically useful list of phrases based on these

Align the extracted phrases with moves / steps in the introductions

Explore the pedagogical use of the list with a group of early career researchers in an ESP context.

We are here

Corpus Construction

Each sub-corpus	# Texts	# Tokens	# Types
Audiology	105	100,062	9,271
Healthcare	105	86,786	9,172
Nursing	105	62,395	7,836
Physiotherapy	105	52,411	7,547
Totals	**420**	**301,654**	**21,049**

Sampling Frame

- ✅ Consulted 24 subject specialists to choose journals.
- ✅ 20 key journals emerged.
- ✅ 7 volumes from each journal.
- ✅ Articles published 2015 – 2021
- ✅ Written in identifiable IMRAD structure

Note – initial focus is on **Introductions**

Background: phraseology and phraseological methods

Phrase: ‘tendency of words to occur in preferred sequences’ (Hunston 2002: 138); the ‘normal primary carrier of meaning’ (Sinclair 2008)

Phraseological phenomena

Patterns (e.g. Hunston & Francis 2000)

Constructions (e.g. Goldberg 1995)

Units of meaning (Sinclair 2004)

Methodologies

n-grams (lexical bundles)

phrase-frames / collocational frameworks

word sketches

Very important to distinguish between them

Why p-frames? Origins

collocational framework (Renouf & Sinclair 1991), e.g. *a/an * of*

n-gram / lexical bundle (e.g. Biber et al. 1999), e.g. *I don't know what, in the case of*

p-frame (Fletcher 2002-2007; Stubbs 2007)/ discontinuous frame (Eeg-Oloffson & Altenberg 1994), e.g. *the * of the, in the * of, on the * of*

'by investigating such frameworks it is possible to discover collocations that may be overlooked or missed entirely in a study of continuous word combinations'

Or, looked at another way…

it is necessary to occurs > 40 times pmw in academic prose (Biber et al. 1999)

But what about…

Or

it is	*important* *vital* *essential*	*to*

it	*seems* *was* *becomes*	*necessary to*

p-frame: allows for a free 'slot' in the string, i.e.

*it is * to*

*it * necessary to*

Essentially, they are n-grams with one (or more) variable slot(s) & are straightforward to retrieve using e.g. AntConc (Anthony 2023)

Types of p-frame study

Exploratory – more interested in nature/distribution of p-frames (in specific genres, disciplines)

- e.g. Eeg-Oloffson & Altenberg (1994), Stubbs (2007), Gray & Biber (2013), Grabowski (2015)

Pedagogical – interested in making some claim of pedagogical utility

- e.g. Marco (2000), Casal & Kessler 2020, Nekrasova-Beker (2019), Lu et al. (2018), Lu et al. (2021)

e.g. making a list of useful phrases

Types of p-frame study

Exploratory – more interested in nature/distribution of p-frames (in specific genres, disciplines)

- e.g. Eeg-Oloffson & Altenberg 1994, Stubbs, 2007, Biber & Gray 2013, Grabowski 2015

Pedagogical – interested in making some claim of pedagogical utility

- **e.g. Marco (2000), Casal & Kessler 2020, Nekrasova-Beker (2019), Lu et al. (2018), Lu et al. (2021)**

this talk is relevant to this sort of study

Example of list derived using p-frames

Move 1. Establishing a research territory

Step 1a. Claiming centrality or value of research area

Frame	***Filler(s)***
*an important * in the*	question, role
*as one of the **	most, least
*at the heart of **	petroleum, the, U.S.
*is an important * of*	indicator, aspect, component

Lu, Yoon & Kisselev (2021: 74)

Issue seems to derive from lack of distinction between the **unit of analysis** used and the **linguistic phenomenon** being investigated

Our questions

- What would a list based on phraseological principles look like?

i.e.

- What are some of the methodological issues involved in terms of
  - p-frame extraction/retrieval: thresholds etc.?
  - manual filtering to find ‘useful’ phrases?
- Can we resolve these in a principled manner?

Initial questions: extraction thresholds

- length of p-frame
- minimum frequency
- range/dispersion

p-frame length: previous studies

Liu, Jang & Du (2023)

He et al (2021)

Tan & Römer (2022)

Mbodj (2021)

Fuster-Marquez & Pennock-Speck (2015)

Walcott (2021)

Nekrasova-Becker (2019)

Geluso (2022)

Jukneviciene & Grabowski (2018)

Nuttal (2021)

Golparvar & Barabadi (2020)

Casal & Kessler (2020)

Ren (2022)

Cunningham (2017)

Lu et al (2018)

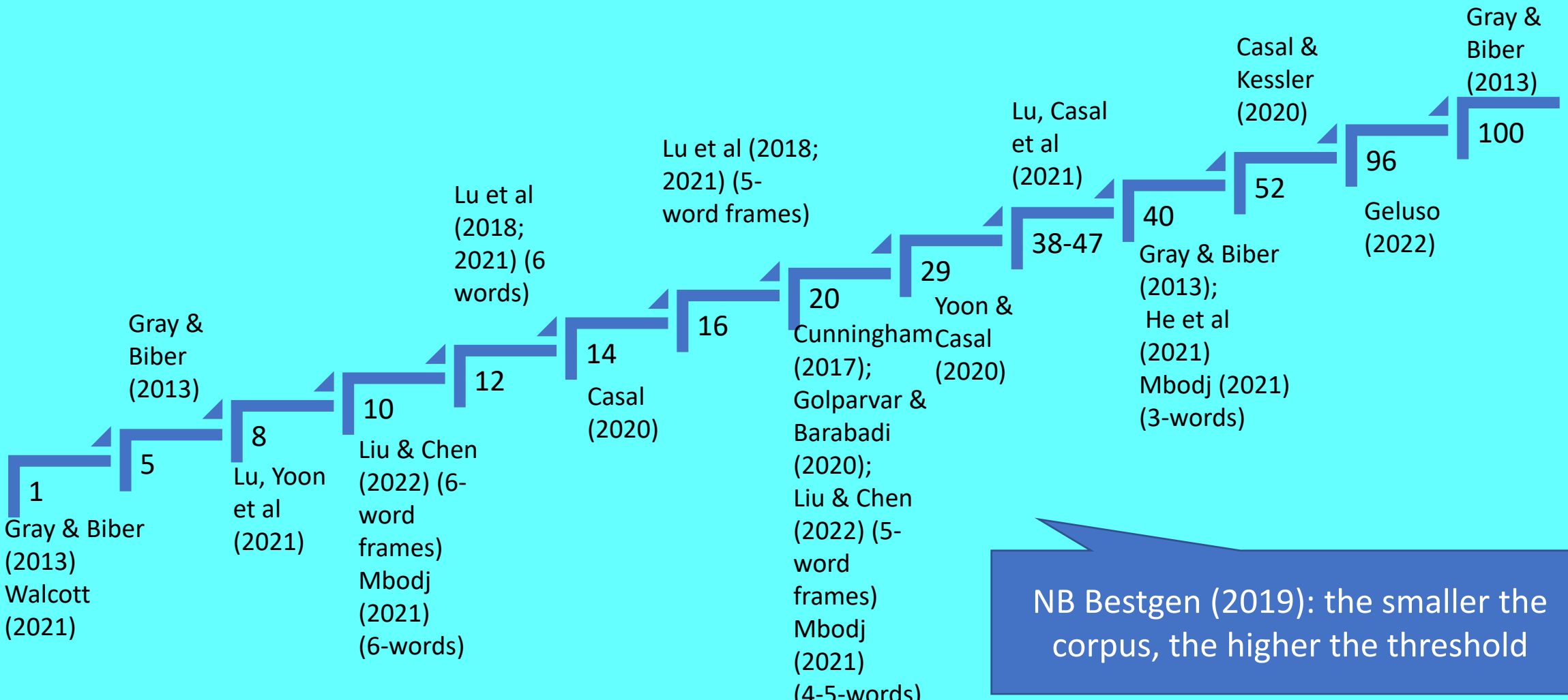
Yoon & Casal (2020)

Lu, Yoon et al (2021)

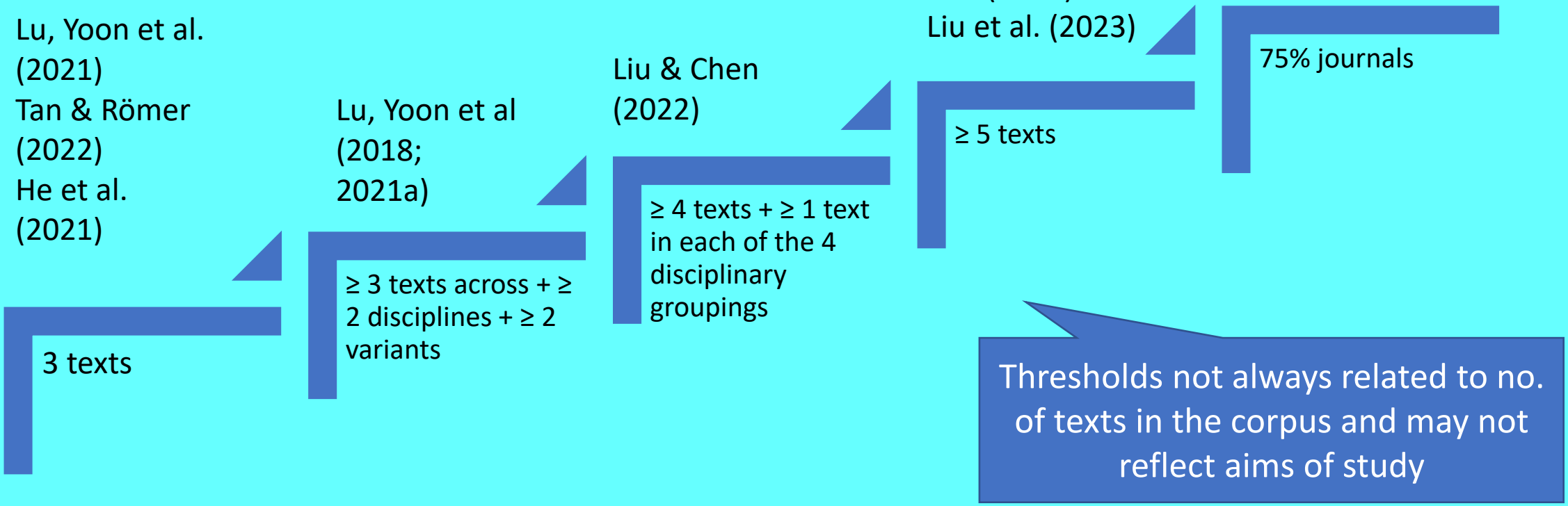
Lu, Casal et al (2021)

Liu & Chen (2022)

Frequency thresholds (pmw)



Range thresholds



Our extraction thresholds

- length of p-frame: 4-word (internal slots only)
  - comparability, similarity to phrasal cores (Vincent 2013), manageability
- minimum frequency: 40 hits pmw
  - i.e. at least 12 instances in our corpus
- range/dispersion
  - at least 10 texts (3 out of 4 sub-disciplines)

Yields 542 p-frames: how long do you think our final list is?

Manual filtering

Once you have extracted 542 ‘candidate’ frames, how to whittle it down to a useful list of phrases?

	Type	Rank	Freq	Range
1	al + et al	1	1042	199
2	et al + et	2	1041	199
3	the + of the	3	494	238
4	in the + of	4	320	186
5	on the + of	5	211	132
6	et al + the	6	204	131
7	the + of this	7	166	140
8	et al + and	8	160	76
9	to the + of	8	160	121
10	and the + of	10	158	118
11	the + of a	11	133	104
12	of the + of	12	126	100
13	this study + to	13	122	116
14	for the + of	14	114	91
15	has been + to	15	109	85

Manual filtering criteria

i) crossing phrasal/clausal/punctuation boundaries

ii) including proper names, symbols etc.

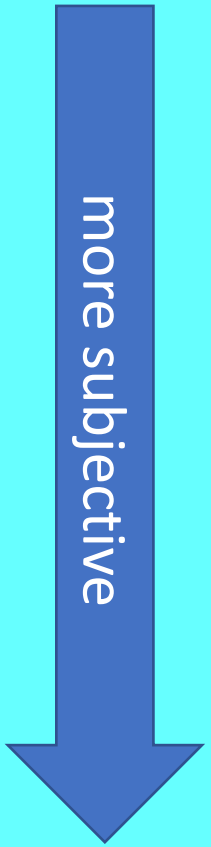
iii) high variability / entropy

iv) overlap with others

v) semantic incoherence of fillers

vi) ‘incomplete’ phrases

vii) not pedagogically useful



Manual filtering criteria

Starting point – **542** 4-word p-frames

		Example(s)	Running total
i)	crossing phrasal/clausal boundaries	*et al * and, to be * and*	417
ii)	including proper names, symbols etc.	*in * united states*	412
iii)	high variability / entropy (of fillers)		

Variability and entropy

Variability – TTR of slot; higher = more variable (compare *in the * in* and *in * current study*)

Entropy – 'the distribution of variant types in a … slot, … rang[ing] from 0 to 1. A value closer to 1 indicates a more even distribution in which all variants are equally likely to occur' (Tan & Römer 2022).

high variability + high Entropy = little to no patterning of fillers; we found if either figure is >0.9 and the other is > 0.75 then p-frame wasn't worth including

NB this is not used directly in previous research but it saves time by removing frames not of interest

Manual filtering criteria

		Example(s)	Running total
i)	crossing phrasal/clausal boundaries	*et al * and, to be * and*	417
ii)	including proper names, symbols etc.	*in * united states*	412
iii)	high variability / entropy	*the * has been*	297
iv)	overlap with other frames		

Overlap

Instances overlap with frame of same length (so one can be removed)

e.g.

*on the * hand* *on * other hand*

and

*one * the most* *one of * most*

Manual filtering criteria

		Example(s)	Running total
i)	crossing phrasal/clausal boundaries	*et al * and, to be * and*	417
ii)	including proper names, symbols etc.	*in * united states*	412
iii)	high variability / entropy	*the * has been*	297
iv)	overlap with others	*on * other hand*	181
v)	semantic coherence of fillers		

Semantic coherence of fillers

examine fillers, noting POS and coherence of meanings. Exclude frames if more than 50% of fillers are categorized as semantically incoherent (Nekrasova-Beker 2019)

Cf 'semantic preference' (Sinclair 2004) also also Renouf & Sinclair (1991)

Also exclude frames if this process takes overall freq below 12 (40 pmw) – not generally mentioned in previous research

Semantic coherence: are these ‘coherent’?

ow-income country and	it is estimated that	six million people have CKD (Hyo
Arjmandzadegan, 2013).	It is estimated that 350	to 400 million people in the world
udies. <1. Introduction>	It is estimated that 43–60%	of persons with Multiple Sclerosis
ical stimulation. Finally,	it is hypothesized that	noise band stimulation is superior
with chronic back pain.	It is hypothesized that	the additional PNE will produce su
99; Nilsson et al., 2001),	it is possible that	the high oestrogen concentration
ed healthy participants.	It is possible that	the inhibitory or facilitatory effect
008; Tanaka et al., 2000),	it is possible that	in response to an exercise challen
al CA-UTI rate in Turkey,	it is reported that	these infections are the most freq
Collard, & Saint, 2005).	It is reported that	approximately 85% of hospital-ac

Judgement of ‘coherence’ may not be solely based on (meanings of) fillers but on other aspects overlapping with next criterion

Rather subjective procedure – IRR important

Manual filtering criteria

		Example(s)	Running total
i)	crossing phrasal/clausal boundaries	*et al * and, to be * and*	417
ii)	including proper names, symbols etc.	*in * united states*	412
iii)	high variability / entropy	*the * has been*	297
iv)	overlap with others	*the * of this*	181
v)	semantic coherence of fillers	*it is*	*hypothesised* / *expected* / *believed*	*that*	73
vi)	‘complete’ phrases		

‘Completeness’ – function of phrase

Not addressed in p-frame literature with exception of Marco (2000); lack of awareness of phraseological work on unit of meaning (Sinclair 2004)?

If we consider semantic coherence of fillers, why not also patterning *surrounding* the p-frame (collocation, colligation, semantic preference, semantic prosody reflecting function of phrase) *in conjunction with* semantic coherence?

Some questionable choices e.g. excluding p-frames composed solely of function words or p-frames that are ‘linguistically incomplete’ without definition of ‘linguistic completeness’

Example: *a * impact on* (24)

After semantic coherence analysis

a [NEGATIVE/BIG] *impact on* (20)

Is this a ‘complete’ phrase? How could we decide on this?

What (HS-relevant) patterns can you see?

disease creates a negative feedback cycle, which eventually has	a negative impact on	surgical recovery. Sleep disturbances and a
repeated, transient exposure to high glucose concentrations has	a negative impact on	the vasculature. One aspect of this might b
t on their QOL.13,16,17 Cheng16 indicated that depression has	a greater impact on	patients' QOL compared with other syndro
e utilisation of healthcare services. Higher OOP expenditure has	a greater impact on	poverty and makes them more vulnerable
health and other physical health problems.1,2 But FAP also has	a big impact on	parents' health and wellbeing. 3 Finally, its
ase an individual's overall vulnerability, which consequently has	a detrimental impact on	his or her physical health. Additionally, incr
fatigue and decreased performance among nurses. Thus, it has	a serious impact on	QOL and renders it impossible for patients
004; Ciocca et al., 2002; Wei et al., 2004). This limitation also has	a strong impact on	the perception and production of melody b
Witter et al., 2013). It is however likely that the strategy can have	a negative impact on	equity in access to services as it can encou
irment. We expected hearing impairment as well as age to have	a negative impact on	masked speech perception performance. H
d negatively by doctors and health professionals and may have	a negative impact on	the doctor-patient relationship as well.5 Th
dromes associated with breast cancer patients' treatments have	a negative impact on	their QOL.13,16,17 Cheng16 indicated that

QOL: quality of life

Looking to the left

disease creates a negative feedback cycle, which eventually has	a negative impact on	surgical recovery. Sleep disturbances and a
repeated, transient exposure to high glucose concentrations has	a negative impact on	the vasculature. One aspect of this might b
t on their QOL.13,16,17 Cheng16 indicated that depression has	a greater impact on	patients' QOL compared with other syndro
e utilisation of healthcare services. Higher OOP expenditure has	a greater impact on	poverty and makes them more vulnerable
health and other physical health problems.1,2 But FAP also has	a big impact on	parents' health and wellbeing. 3 Finally, its
ase an individual's overall vulnerability, which consequently has	a detrimental impact on	his or her physical health. Additionally, incr
fatigue and decreased performance among nurses. Thus, it has	a serious impact on	QOL and renders it impossible for patients
004; Ciocca et al., 2002; Wei et al., 2004). This limitation also has	a strong impact on	the perception and production of melody b
Witter et al., 2013). It is however likely that the strategy can have	a negative impact on	equity in access to services as it can encou
irment. We expected hearing impairment as well as age to have	a negative impact on	masked speech perception performance. H
d negatively by doctors and health professionals and may have	a negative impact on	the doctor-patient relationship as well.5 Th
dromes associated with breast cancer patients' treatments have	a negative impact on	their QOL.13,16,17 Cheng16 indicated that

Looking to the right

disease creates a negative feedback cycle, which eventually has	a negative impact on	surgical recovery. Sleep disturbances and a
repeated, transient exposure to high glucose concentrations has	a negative impact on	the vasculature. One aspect of this might b
t on their QOL.13,16,17 Cheng16 indicated that depression has	a greater impact on	patients' QOL compared with other syndro
e utilisation of healthcare services. Higher OOP expenditure has	a greater impact on	poverty and makes them more vulnerable
health and other physical health problems.1,2 But FAP also has	a big impact on	parents' health and wellbeing. 3 Finally, its
ase an individual's overall vulnerability, which consequently has	a detrimental impact on	his or her physical health. Additionally, incu
fatigue and decreased performance among nurses. Thus, it has	a serious impact on	QOL and renders it impossible for patients
004; Ciocca et al., 2002; Wei et al., 2004). This limitation also has	a strong impact on	the perception and production of melody b
Witter et al., 2013). It is however likely that the strategy can have	a negative impact on	equity in access to services as it can encou
irment. We expected hearing impairment as well as age to have	a negative impact on	masked speech perception performance. H
d negatively by doctors and health professionals and may have	a negative impact on	the doctor-patient relationship as well.5 Th
dromes associated with breast cancer patients' treatments have	a negative impact on	their QOL.13,16,17 Cheng16 indicated that

‘complete’ phrase from instances of *a* impact on*

[CONDITION] HAVE *a* [NEGATIVE/BIG] *impact on* [(ASPECT OF) HEALTH]

For example

FAP also has a big impact on parents’ health and wellbeing

According to the model of the lexical unit (Sinclair 2004), this should fulfil a ‘function’ (i.e. have a ‘semantic prosody’); maybe here associated with the move ‘establishing the territory’

Again this stage involved comparison of ratings to achieve final agreed list, with some exclusions on basis of low frequency/range

Manual filtering criteria

i) crossing phrasal/clausal boundaries

ii) including proper names, symbols etc.

iii) high variability / entropy

iv) overlap with others

v) semantic coherence of fillers

vi) ‘complete’ phrases

Example(s)	Running total
*et al * and, to be * and*	417
*in * united states*	412
*the * has been*	297
*the * of this*	181
it is	
[CONDITION] HAVE *a* [NEGATIVE/BIG] *impact on* [(ASPECT OF) HEALTH]	34

Are they likely to be ‘useful’, though?

Pedagogical ‘usefulness’

Is the list ‘useful’: are they likely to help aspiring researchers in Health Sciences write introductions more effectively? How to find this out?

Our approach was to survey stakeholders – i.e. academics working and publishing in the field of Health Sciences

kindly piloted by Elena Mazzeri (BA student at Coventry University)

Survey

Please evaluate the pedagogical usefulness of the phrases provided considering following questions:

- Do you recognise the phrase?
- Does it have a clear function/use in your field?
- Is it an expression that you would use when preparing papers?
- Is it a phrase that new researchers in the area might struggle to use?

From 1: definitely not useful to 5: definitely useful; worth teaching

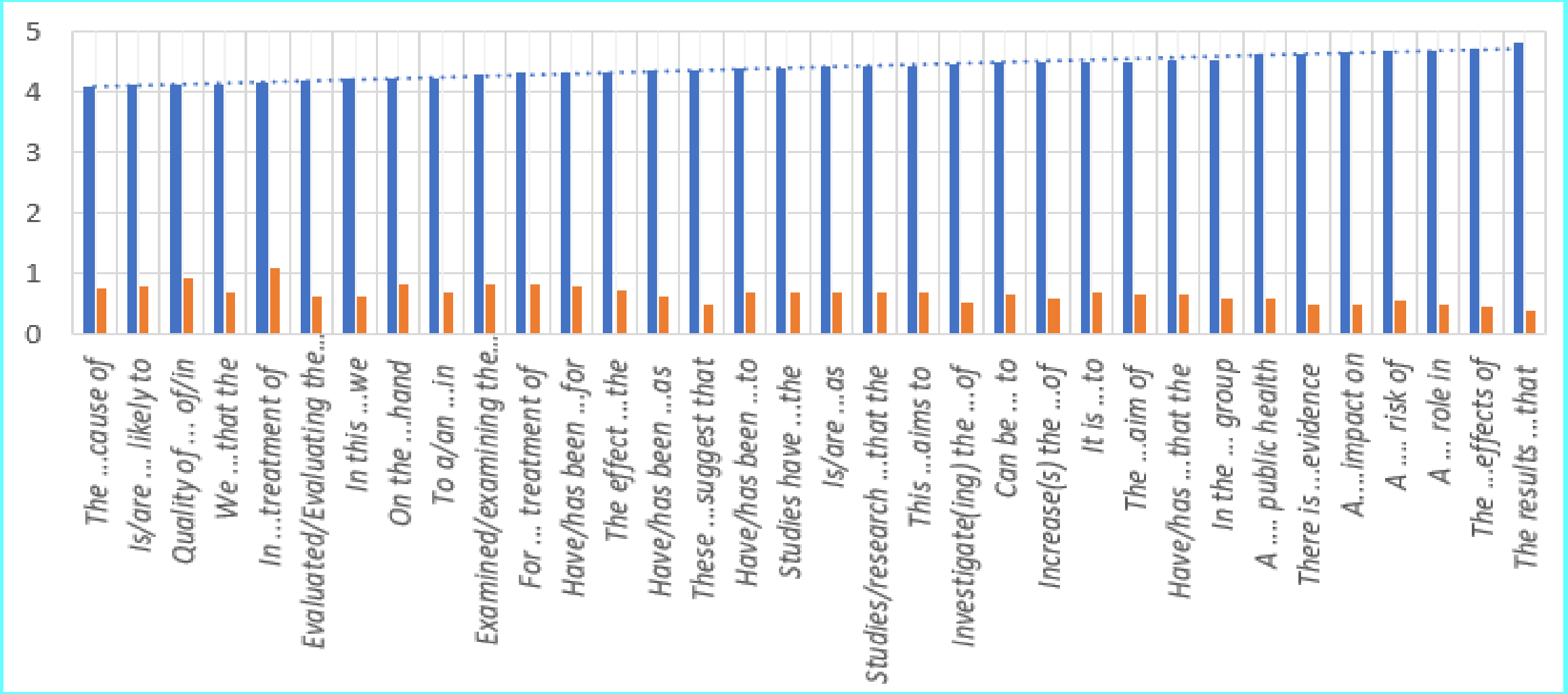
Survey item

Core Phrase			Structure	Example Sentence
A	negative detrimental serious severe	impact on	condition/illness HAVE a (negative) impact on [well-being/quality of life]	Long-term illnesses are highly prevalent and have a severe impact on well-being.

29 participants: 7 EAP/ESP teachers + 22 academics in Health Sciences in Turkey

Survey results

NB in the opinion of researchers, **not** novice writers – we still need to actually try these out!



Conclusions

- Attraction of p-frame research: easy to produce a list
- In practice producing a *pedagogical list* is hard; we still don't know for sure if we've done this
- Important not to lost sight of the final goal – what's the aim?
- Also not to forget that p-frames is a method, a starting point. Nobody 'uses' a p-frame; they use a phrase, a linguistic item

Paper to be submitted to RMAL for our SI at end of Jan

Thanks for coming! Any questions?

More project info here: https://clac.coventry.domains/

References

Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bestgen, Y. (2019) Comparing Lexical Bundles across Corpora of Different Sizes: the Zipfian Problem. *Journal of Quantitative Linguistics.* https://doi.org/10.1080/09296174.2019.1566975

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). Longman.

Casal, J. E., & Kessler, M. (2020). Form and rhetorical function of phrase-frames in promotional writing: A corpus- and genre-based analysis. *System*, *95*, 102370. https://doi.org/10.1016/j.system.2020.102370

Cunningham, K. J. (2017). A phraseological exploration of recent mathematics research articles through key phrase frames. *Journal of English for Academic Purposes*, *25*, 71. https://doi.org/10.1016/j.jeap.2016.11.005

Eeg-Olofsson, M. & Altenberg, B. (1994) Discontinuous recurrent word combinations in the London-Lund Corpus. In U. Fries, G. Tottie, & P. Schneider (Eds.), *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora*. Rodopi, pp. 64-77.

Fletcher, W. (2002-2007). KfNgram. Annapolis: USNA. (10 January, 2024)

Geluso, J. (2022). Grammatical and functional characteristics of preposition-based phrase frames in English argumentative essays by L1 English and Spanish speakers. *Journal of English for Academic Purposes*, *55*, 101072.

Goldberg, A. (1995). *Constructions : A construction grammar approach to argument structure*. University of Chicago Press.

Golparvar, S. E., & Barabadi, E. (2020). Key phrase frames in the discussion section of research articles of higher education. *Lingua*, *236*, 102804. https://doi.org/10.1016/j.lingua.2020.102804

Grabowski, Ł. (2015). Phrase frames in English pharmaceutical discourse: a corpus-driven study of intra-disciplinary register variation. *Research in Language*, *13*(3), 266-291.

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, *18*(1), 109–136. https://doi.org/10.1075/ijcl.18.1.08gra

He, M., Ang, L. H., & Tan, K. H. (2021). A corpus-driven analysis of phrase frames in research articles on business management. *Southern African Linguistics and Applied Language Studies*, *39*(2), 139–151. https://doi.org/10.2989/16073614.2021.1920438

Hunston, S. and Francis, G., (2000) *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins Publishing.

Juknevičienė, R., & Grabowski, Ł. (2018). Comparing formulaicity of learner writing through phrase-frames: A corpus-driven study of Lithuanian and Polish EFL student writing. *Research in language*, *16*(3), 303-323.

Liu, C.-Y., & Chen, H.-J. H. (2022). A phraseological exploration of university lectures through phrase frames. *Journal of English for Academic Purposes*, *58*(December 2021), 101135. https://doi.org/10.1016/j.jeap.2022.101135

Liu, L., Jiang, F. K., & Du, Z. (2023). Figure legends of scientific research articles: Rhetorical moves and phrase frames. *English for Specific Purposes*, *70*, 86-100.

Lu, X., Casal, J. E., Liu, Y., Kisselev, O., & Yoon, J. (2021). The relationship between syntactic complexity and rhetorical move-steps in research article introductions: Variation among four social science and engineering disciplines. *Journal of English for Academic Purposes*, *52*(April), 101006. https://doi.org/10.1016/j.jeap.2021.101006

Lu, X., Yoon, J., & Kisselev, O. (2018). A phrase-frame list for social science research article introductions. *Journal of English for Academic Purposes*, *36*, 76–85. https://doi.org/10.1016/j.jeap.2018.09.004

Lu, X., Yoon, J., & Kisselev, O. (2021). Matching phrase-frames to rhetorical moves in social science research article introductions. *English for Specific Purposes*, *61*, 63–83. https://doi.org/10.1016/j.esp.2020.10.001

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: A genre-based study. *English for specific purposes*, *19*(1), 63-86.

Mbodj, N. B. (2021). Writing in the Disciplines and Within-discipline Variations: A Comparison of the Formulaic Profiles of the Medical Research Article and the Medical Case Report.

Nekrasova-Beker, T. M. (2019). Discipline-specific use of language patterns in engineering: A comparison of published pedagogical materials. *Journal of English for Academic Purposes*, *41*, 100774.

Nuttall, C. (2021). Profiling lexical frame use in NSF grant proposal abstracts. *Applied Corpus Linguistics*, *1*(3), 100009.

Ren, J. (2022). A comparative study of the phrase frames used in the essays of native and nonnative English students. *Lingua*, *274*, 103376.

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. *English corpus linguistics*, 128-143.

Simpson-Vlach, R., & Ellis, N. (2010) An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics 31*(4), 487-512

Sinclair, J. (2004) *Trust the text: language, corpus and discourse.* Routledge.

Stubbs, M. (2007). An example of frequent English phraseology: Distribution, structures and functions. In Facchinetti, R. (ed.) *Corpus Linguistics 25 Years on*, (pp. 89–105). Rodopi.

Tan, Y., & Römer, U. (2022). Using phrase-frames to trace the language development of L1 Chinese learners of English. *System*. https://doi.org/10.1016/j.system.2022.102844

Walcott, K. (2021). Informing academic writing pedagogy through the study of phrase-frames. *Journal of Language Teaching and Research*, *12*(1), 158–171. https://doi.org/10.17507/jltr.1201.1

Yoon, J., & Casal, J. E. (2020). P-frames and rhetorical moves in applied linguistics conference abstracts. In U. Römer, V. Cortes, & E. Friginal (Eds.), *Advances in corpus-based research on academic writing: Effects of discipline, register, and writer expertise* (pp. 282–305). John Benjamins. https://doi.org/10.1075/scl.95.12yoo