Université Paris Nanterre

CREA EA 370

PAC THE PHONOLOGY OF CONTEMPORARY ENGLISH

Theoretical and methodological issues in the description of spoken English syntax: Perspectives from the PAC New Zealand syntactic annotation project

Sophie Raineri, Romain Delhem*, Cécile Viollain & Hugo Chatellier

Université Paris Nanterre (CREA), France

***Université Clermont Auvergne (LRL), France**

PAC 2023 – Université Paris Nanterre – April 12, 2023

UNIVERSITÉ Clermont Auvergne

Laboratoire de Recherche sur le Langage

Starting point

<K:> mmmm that's another movie I'd like to see that one about the er

<B:> Mark Mark Zuckerberg

<K:> yeah what's it The Social Experiment or what's it called something like that can't remember

<B:> yeah

<K:> it's supposed to be a very good movie I heard it being reviewed in the er radio erm the other day and the the woman who reviewed it her uncle who is quite elderly it sounds like and has nothing to do with Facebook or computer technology or any sort of information technology he went and thought it was great a great movie

<B:> mm mm

<K:> it seems like you don't have to be tied up in the whole in the whole phenomena yourself in order to appreciate the story of it

Illustration of main features of oral syntax

"[C]onversational grammar is **non-sentence-based**, **co-constructed** and **highly interactive**."

(Carter & McCarthy, 2015: 1)

"[G]rammatical features in spoken utterances reflect **the creation of discourse rather than just the internal construction of phrases, clauses and sentences**."

(Carter & McCarthy, 2006: 177)

"[M]uch spontaneous language does not even have a syntactic structure in which phrases combine into clauses or clauses into integrated clause complexes. Rather, the structure consists of **blocks of syntax with little or no syntactic linkage and requiring from the listener a larger than usual inference** based on contextual and world knowledge."

(Miller & Weinert, 1998: 28)

PAC data

PAC in figures

- ✓ more than 600 speakers (native & non-native)
- ✓ from 15 countries (Ireland, England, Scotland, Canada, USA, NZ, Aus., India, Singapore, Italy, France, Spain, China…)
- ✓ and more than 30 survey locations (Birmingham, Manchester, Sydney, Delhi, Cork, Ottawa, Montreal, Boston…)
- ✓ that is approx. 750 hours of recordings

Data & metadata

- ➢ A shared protocol (LVTI - langue, ville, travail, identité)
  - ❑ reading tasks (2 wordlists & 1 text)
  - ❑ 2 conversations (1 interview with fieldworker & 1 informal conversation)
  - ❑ 1 sociolinguistic questionnaire (age, origins, interests, class, gender, etc.)

Outline

1. **(Quick) review of the literature on oral syntax**
   - Advances in the description of syntax of spoken English
   - Alternatives to traditional sentences in anglophone literature
   - Alternatives to traditional sentences in francophone literature
   - Existing annotated corpora of spoken English

2. **Goals & methods**
   - Research questions
   - Syntactic segmentation: principles
   - Syntactic segmentation: procedure

3. **Preliminary results, challenges & discussion**
   - Characterization of Macro-Syntactic Units
   - Focus on co-construction
   - Focus on syntactic indeterminacy

Advances in the description of the specificities of oral syntax

Biber *et al.* (1999), Carter & McCarthy (2006), Corley & Stewart (2008), Haselow & Hancil (2021), Kirjavainen *et al.* (2022)

- **Syntactically incomplete sentences**
- Unembedded dependent clauses
- Clausal blends
- **Low degree of syntactic elaboration**
- Low type/token ratio
- **Syntactic ellipsis**
- Anaphoric expressions
- Disjunctive elements
- Vague expressions
- High frequency of imperatives and questions
- High frequency of negative and adversative markers
- Pragmatic markers
- **Non-standard grammar**
- Dysfluence

"[w]e may wonder what (if anything) should replace the sentence as a basis for dividing the spoken discourse into maximal grammatically analyzable (parsable) units?"

Biber *et al.* (1999: 1038)

Alternatives to traditional sentences in the anglophone literature

Prosodic units (Izre'el *et al.* 2020; Chafe 1987)

Turn constructional units (Schegloff 2007)

Larger (functional) discourse units (Egbert *et al.* 2021)

Syntactic units (Miller & Weinert 1998; Biber *et al.* 1999; Foster *et al.* 2000; Degand & Simon 2009; Benzitoun 2010; Pietrandrea *et al.* 2014; Lacheret-Dujour *et al.* 2019)

Clauses & C-units

"[T]he clause should be taken as **the major locus of distributional and dependency relations and not the (system) sentence**."

(Weinert & Miller, 1998: 46)

"Clausal and non-clausal units are maximal grammatical units in the sense that they cannot be syntactically integrated with the elements which precede or follow them (...). We will use **the umbrella term C-unit for both clausal and non-clausal units, *i.e.* for syntactically independent pieces of speech**."

(Biber *et al.*, 1999: 1070)

B: || So this was your mother's? ||

A: || No, || my father's. ||

B: ||Your father's mother? ||

A: || Yeah. || Her name was Martha <name> ||

B: || Uh huh. ||

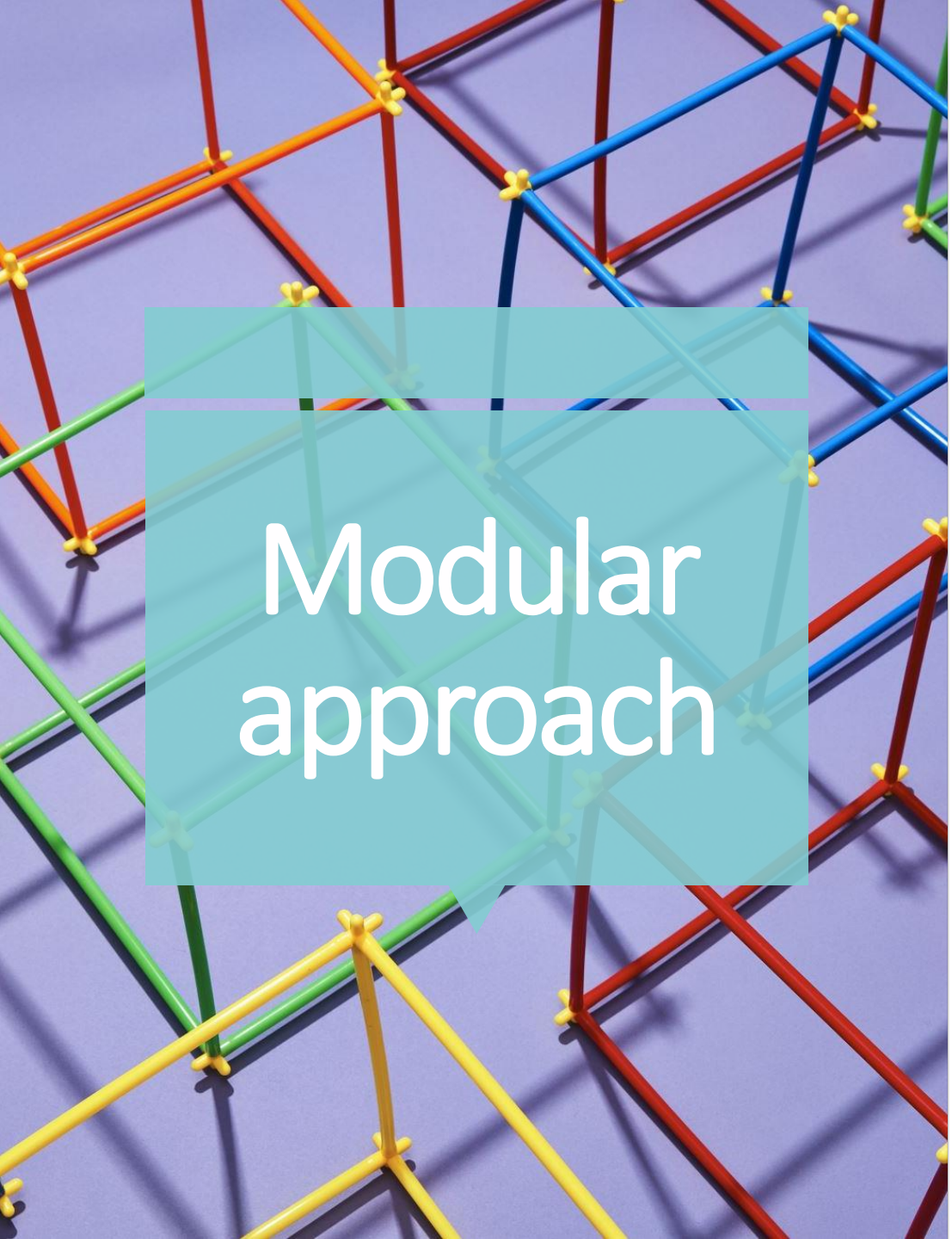
Macro & microsyntax

- **2 levels of syntactic structure** (Blanche-Benveniste *et al.* 1990; Degand & Simon 2009; Lacheret-Dujour *et al.* 2019)

"Macrosyntax describes the relations holding between a number of syntactic constructions typical of spoken language and particularly frequent in spoken French – such as paratactic structures, detachments, discourse markers – and the rest of the production."
(Lacheret-Dujour et al., 2019: 38)

"This micro-syntactic analysis is then expanded to the macro-syntactic level which includes so-called 'associés' ('adjuncts') which are not governed by the main verb (hence offering no possibility for clefting), but are semantically or pragmatically linked to the whole dependency clause (in a 'préfixe' or 'postfixe' establishing a pragmatic relationship to the main clause). They have a non-autonomous status in discourse."
(Degand & Simon 2009: 7)

- **2 types of units**
  - Government Unit (GU) at the microsyntactic level
  - Illocutionary Unit (IU) at the macrosyntactic level



Modular approach

"[L]anguages are organized in a number of autonomous mechanisms of linguistic cohesion operating simultaneously and independently from one another in discourse."

(Pietrandrea *et al*., 2014: 335)

Existing annotated corpora of spoken English

	Dialectal variation	Communication setting & style variation	Speakers' sociolinguistic profiles	Syntactic annotation	Prosodic annotation	Access to sound files
Switchboard Corpus (Meteer *et al.* 1995)	✗	✗	✗	✓	✓	✓
Christine Corpus (Rahman & Sampson 2000)	(✓)	✗	(✓)	✓	✗	✗
DCPSE (Aarts & Wallis 2006)	✗	✓	✓	✓	✗	✓
Lancaster/IBM SEC (Knowles *et al.* 1996)	✗	(✓)	(✓)	✓	✓	✓
ICE (Greenbaum 1991)	✓	✓	(✗)	✓	(✗)	✗
HKCSE (Cheng et al. 2008)	✗	✓	✗	✗	(✓)	✗
PAC corpora (Durand *et al.* 2015)	✓	✓	✓	✓	✓	✓

Research questions

- "Is there **a distinctive grammar of spoken language, operating by laws different from those of the written language?** If so, what are these laws, and what are the functional or other principles underlying them?"

  (Biber *et al*., 1999: 1038)

- What are the **appropriate analytical units** for the syntax of spontaneous spoken language?

- "To what extent do prosodic and syntactic structures interact? **To what extent are they autonomous from one another** in creating discourse units?"

  (Lacheret-Dujour *et al.*, 2019: 1)

- To what extent does the syntax of spoken English **vary**?

PAC New Zealand database

3 corpora compiled in 2010 (Viollain 2014) in:

- Wellington (North Island)
- Christchurch (South Island)
- **Dunedin (South Island)**

13 speakers in total: 5 male & 8 female

- 3 generations:
  - 3 informants under 20
  - 5 informants between 43 & 51
  - 5 informants between 65 & 76
- social class: homogeneous (middle / upper middle)
- ethnicity: all Pākehā except for one informant of both Māori and Pākehā origins
- all raised in New Zealand

Data for our test study

- 9 informal conversations (≅ 90 min of oral speech)

Principles of syntactic segmentation

Maximal Syntactic Unit (MSU) : ungoverned head (verb or other) + dependent elements + peripheral non-autonomous elements

- Dependency relations + piles (repetition of same structure, dislocation)

- Unfinished clauses treated as distinct MSUs

| *he worked as a* | *erm he was a teacher* |

- Independent coordinated clauses treated as distinct MSUs

| *no no I didn't do UE at all* | *erm I did NCEA level two* | *and then we moved down here*

- Coordinated clauses within subordinate clauses NOT treated as distinct MSUs

| *is there a physics for people who need some physics generally but don't think they're going to go on in it* |

- Question tags treated as distinct MSUs

you didn't do 141 this year | *did you* | *you did it*

| for maximal syntactic boundary

_ for potential maximal syntactic boundary (cases of syntactic indeterminacy):

<W:> | well you should email them back anyway _ and just _ because she said she wouldn't |

<W:> | well I tried to do that | I didn't press anything |

<H:> | I wonder _ sometimes _ for some reason it sometimes it goes on mute |
<W:> | yeah oh I wonder if it's second mute or something |

Annotation system

Annotation system

§ for co-construction of speech by 2 speakers across turns:

<OH:> | you see | and so we take all the good files | and put them on to the erm
<OW:> | oh right |
<OH:> on to the new one you see and
<OW:> | right |
<OH:> and what not and | and then I discovered the shredder wasn't
<OW:> § working |
<OH:> working |

Annotation system

\+ for partial maximal syntactic boundary:

Inserts / parentheticals

| it's supposed to be a very good movie | I heard it being reviewed in the er radio erm the other day | and the the woman who reviewed it her uncle who is quite elderly + it sounds like + and has nothing to do with Facebook or computer technology or any sort of information technology he went | and thought it was great a great movie |

Reported speech

| James does know because Emma texted him saying + you've got have you got my phone by mistake | and James then texted her back to say yes he did.

Annotation procedure

- ✓ Standard Orthographic Transcription (SOT) done in Praat using PAC transcription conventions (which notably included prosodic information about pauses etc.)
- ✓ Conventions/punctuation (, . ?), capitalization removed to obtain raw orthographic script
- ✓ Double coding
- ✓ Merger of the 2 codings after group discussions on disagreements

→ Adjusted annotation system

Analysis of MSUs

Total word count	21 032
Word count excluding indeterminate MSUs	17 267 (82,09%)
Number of MSUs (excluding indeterminate MSUs)	2 364
Average length of MSUs (number of words)	7,58
Shortest MSU (number of words)	1
Longest MSU (number of words)	63
Number of 1-word MSUs	410 (17,34%)

Analysis of MSUs

Focus on 1-word MSUs

hm/mmm

yes/yeah/yup/no

ok

well

oh/wow

really

great/cool/sweet

+ 1-word answer to previous question

Focus on longest MSU

| me and Chris are meeting at least twice to mark 30 papers look at the marking schedule to guinea pig mark some papers and change the marking schedule if we need to and then I think mark 30 of our own papers sort of sitting together and check if we've got any queries with each other and see how that schedule fits |

Analysis of MSUs

Number of SV(+) MSUs	1530 (64,72%)
Number of complex MSUs	550 (23,27%)

Number of MSUs prefaced by *and*	257 (10,87%)
Number of MSUs prefaced by *but*	155 (6,56%)

<MAH:> | so as long as the weather forecast is good you'll go |

<MAW:> | yes | but if the weather forecast is windy or something like that we're not going to go |

<MAH:> | so what | you're going to wait wait for booking accommodation |

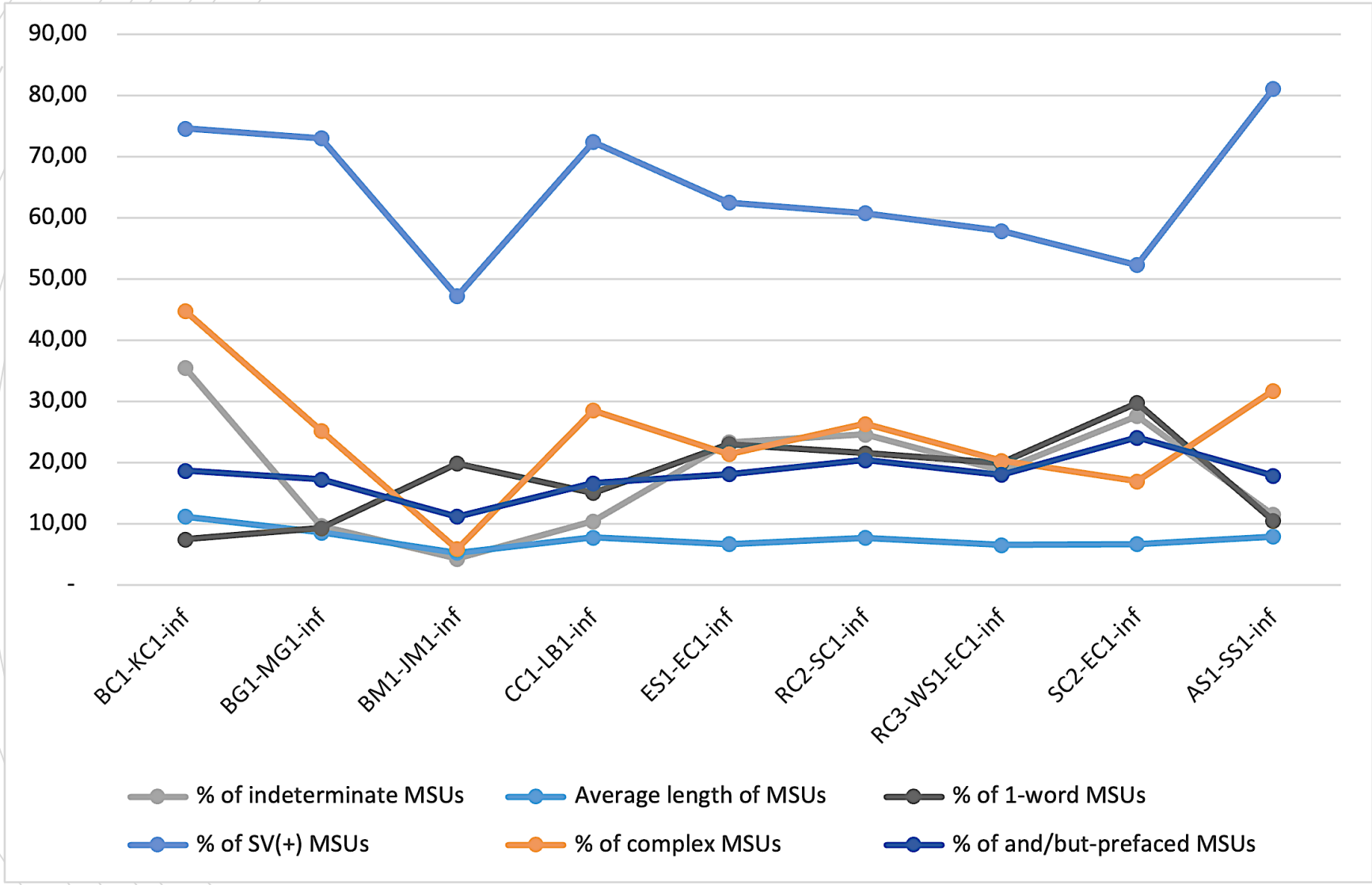
<MAW:> | yeah | but there's not going to be a demand for booking at that time of year | is there |

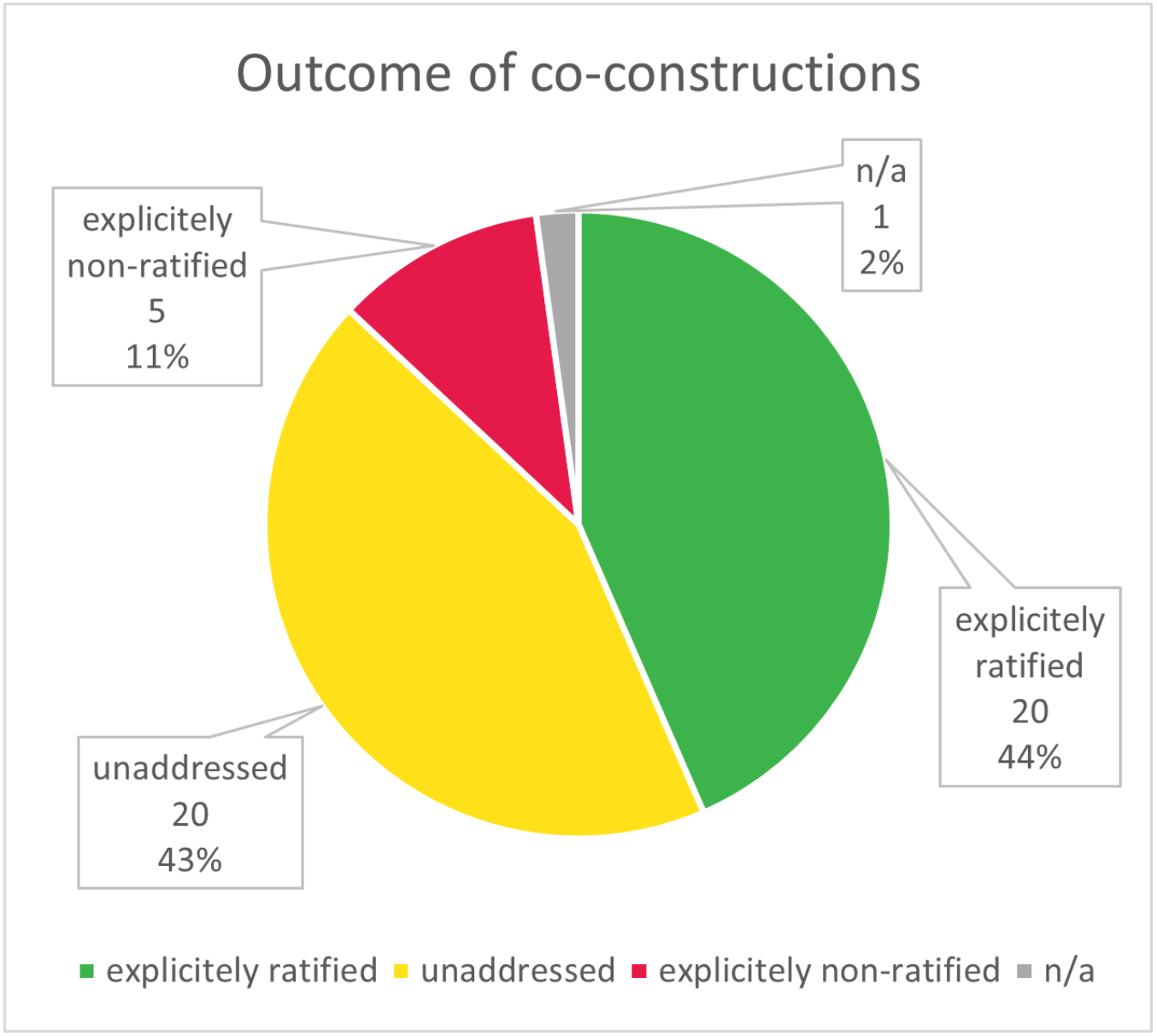
<MAH:> | probably isn't |

<MAW:> | but we do need cabins or something | don't want to stay in a tent |

Analysis of MSUs

Analysis of MSUs





Focus on co-construction

Ratified

W: | mmmm that's a another movie I'd like to see the one about the er
H: § Mark Zuckerberg |
W: | yeah is it The Social Experiment | what's it called | something like that | can't remember |

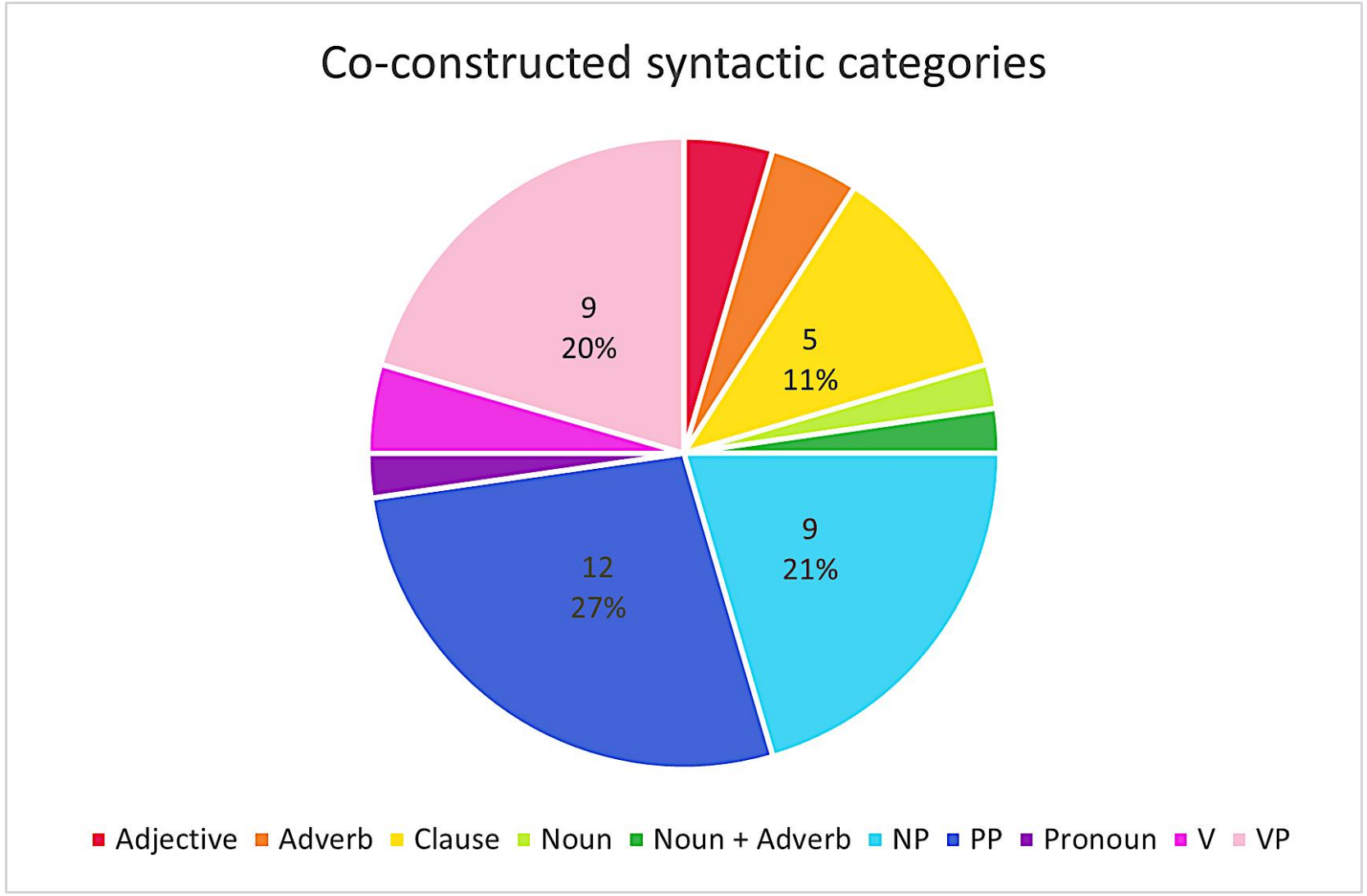
Non-ratified

F1: | and then Bruce was very keen to do er some camping _ that _ you can freedom camp in er Scotland |
F2: § with the right kind of camper then |
F1: | oh no no no | we intend to er

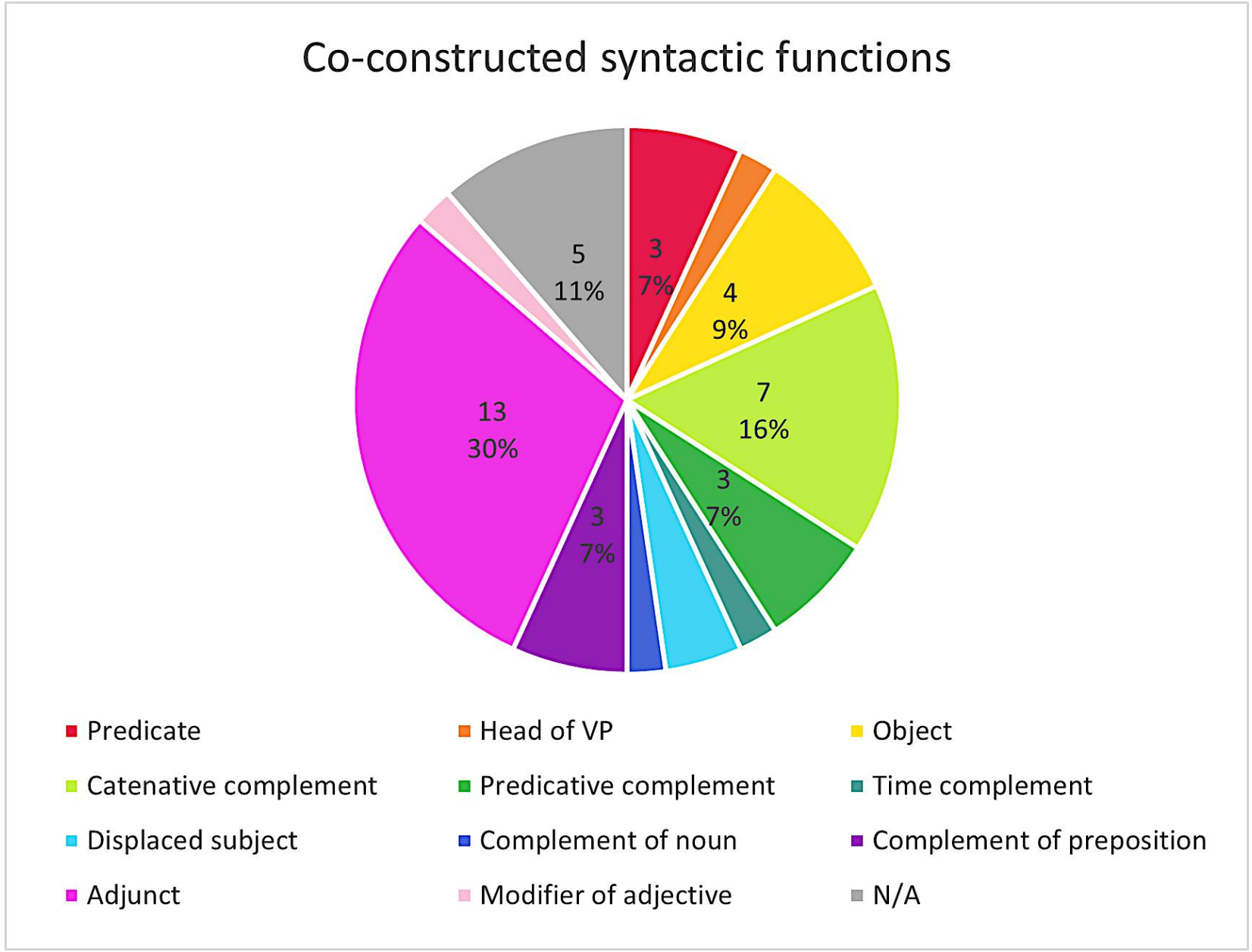
Unaddressed

OH: she's a Scottish fold apparently which is _ you know _ they have their ears folded back
OW: § folded down |
OH: and the round owl-like face |

Focus on co-construction



Focus on co-construtions



Focus on co-construtions

Object
OF: the various domains that want to have +
Y: § yeah an input |

Time adjunct
MAH: | so he's got physics tomorrow |
MAW: § in the afternoon | but that's good if say Hans helped him |

Catenative complement
OH: | so she probably she wouldn’t be any good to interview | she’d probably | being a Scottish fold she’d probably speak |
OW: § speak the wrong the wrong dialect |

Focus on co-construction

Focus on syntactic indeterminacy

Prosodic information solves ambiguity

<MS:> | yeah I did | I only did one 141 this | you didn't do 141 this year | did you | you did it last year |
<FS:> | 141 _ what _ for _ French |
<MS:> | French yeah |
<FS:> | I did that last year |
<MS:> | yeah |
<FS:> | yeah |

Prosodic information does not solve ambiguity

<F:> | oh so where where would that be | like are there specific regions that are good for wind |
<S:> | er almost everywhere in the world around the coast is good
<F:> | ok |
<S:> because you get sea-breezes |
<F:> | hm |
<S: | and other than that it depends on | as you say it depends on the geography _ so _
<F:> | ok |
<S:> _ some places have you know mountain passes which cause winds to flow | erm islands are usually pretty good because they are usually not too far from the sea |

Selected references

Benzitoun C. (2010). Quelle(s) unité(s) syntaxique(s) maximale(s) en français parlé? Discussions autour de quelques problèmes rencontrés. Travaux de Linguistique 60, 109-126.

Biber, D., Johansson, S., Leech, G. Conrad, S. & Finegan, E. (1999). Longman grammar of spoken and written English. Edinburgh: Pearson Education Ltd.

Carter, R., & McCarthy, M. (2006). Cambridge grammar of English: A comprehensive guide: Spoken and written English grammar and usage. Cambridge: CUP.

Carter, R. & McCarthy, M. (2017). Spoken Grammar: Where Are We and Where Are We Going?. Applied Linguistics 38, 1-20. Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. Language and Linguistics Compass 2(4), 589-602.

Degand, L. & Simon, A. C. (2009). On identifying basic discourse units in speech: theoretical and empirical issues. Discours 4. https://doi.org/10.4000/discours.5852.

Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T. and Baker, P. (2021). Identifying and describing functional discourse units in the BNC Spoken 2014. Text & Talk 41 (5-6), 715-737.

Foster, P., Tonkyn, A. & Wigglesworth, G. (2001). Measuring spoken language: a unit for all reasons. Applied Linguistics 21(3), 354-375.

Haselow, A. & Hancil, S. (2021). Studies at the Grammar-Discourse Interface. Discourse Markers and Discourse-Related Grammatical Phenomena. Amsterdam: Benjamins.

Izre'el, S., Mello, H., Panunzi, A., & Raso, T. (Eds.). (2020). In Search of Basic Units of Spoken Language: A Corpus-Based Approach. Amsterdam: John Benjamins.

Kirjavainen, M., Crible, L., & Beeching, K. (2022). Can filled pauses be represented as linguistic items? Investigating the effect of exposure on the perception and production of um. Language and Speech 65(2), 263-289.

Miller, J. & Weinert, R. (1998). Spontaneous Spoken Language: Syntax and Discourse. Oxford: OUP.

Pietrandrea, P., Kahane, S., Lacheret, A. & Sabio, F. (2014). The notion of sentence and other discourse units in spoken corpus annotation. In: Mello, H. & Raso, T. (Eds). Spoken corpora and Linguistic Studies. Amsterdam-Philadelphia: John Benjamins, 331-364.

Schegloff, E.A. (2007). Sequence organization in interaction: A primer in conversation analysis. Cambridge: CUP.

Viollain, C. (2014). Sociophonologie de l'anglais contemporain en Nouvelle-Zélande : corpus et dynamique des systèmes. PhD dissertation. Université Toulouse II.