Stefan Schnell*, Nils Norman Schiborr, and Geoffrey Haig

Efficiency in discourse processing: Does morphosyntax adapt to accommodate new referents?

Supplementary material

https://doi.org/..., Received ...; accepted ...

Keywords: efficiency, discourse processing, recipient design, information pressure, preferred argument structure, corpus-based typology

Communicated by: ...
Dedicated to ...

Supplementary material

This document contains background information on the analyses presented in the main article, including descriptions of the corpus data and methodologies. For the raw data from Figures 1–5, see Tables 3–9 below. Please refer to the main article for a full list of references.

Multi-CAST corpus data

The results presented in this paper mostly draw from data from the *Multilingual Corpus of Annotated Spoken Texts* (Multi-CAST, Haig and Schnell 2015),[1]

1 *The Multilingual Corpus of Annotated Spoken Texts* (Multi-CAST), accessible online at https://multicast.aspra.uni-bamberg.de/.

***Corresponding author: Stefan Schnell,** Department of General Linguistics, University of Bamberg & ARC Centre of Excellence for the Dynamics of Language, stefan.schnell@uni-bamberg.de
Nils Norman Schiborr, Department of General Linguistics, University of Bamberg, nils-norman.schiborr@uni-bamberg.de
Geoffrey Haig, Department of General Linguistics, University of Bamberg, geoffrey.haig@uni-bamberg.de

a collection of spoken language corpora from a typologically diverse range of languages. The texts in Multi-CAST are all narratives of various kinds, and predominantly monologic. Multi-CAST has been designed to enable crosslinguistic inquiries into referentiality and discourse structure by providing common ground for quantitative analyses. Schnell and Schiborr (2018) provide a basic introduction to the project. The data used in this paper correspond to version 2001 of Multi-CAST, published in January 2020.

The texts in Multi-CAST feature morphosyntactic annotations with the GRAID scheme (*Grammatical relations and animacy in discourse*, Haig and Schnell 2014), referent identification with the RefIND scheme (*Referent indexing in natural-language discourse*, Schiborr et al. 2018), and note the information status of new referents with a reduced version of the RefLex scheme (Riester and Baumann 2017) which we label ISNRef (*Information status of new referents*). The combination of these three layers of annotation allow the data in Multi-CAST to be used for a variety of cross-linguistic and comparative research tasks in referentiality and the intersection between discourse and grammar. We refer readers to the annotation guidelines for details, all of which can be found on the Multi-CAST website.

The following selection criteria have been applied to the data, yielding the sample outlined in Table 1.

- only expressions evoking an identifiable referent
- only referents with at least two mentions in a text (i.e. the introduction and one other occurrence)
- only third person mentions (i.e. excluding first person mentions)
- no clausal references (i.e. headless relative clauses or complement clauses)

For Figure 5, we calculate local information pressure as a function of the number of other referents mentioned in the preceding three clauses. In Du Bois (1987), information pressure is instead measured as a ratio of discourse referents against text length. In Table 2 we provide corresponding values for our data.

Pear story corpus data

Figure 5 is based on data from two corpora of Pear story retellings. The first, from English, is taken from the appendix of Chafe (1980); it contains retellings by 20 speakers. The second, from Persian, is one of the Multi-CAST corpora described above (Adibifar 2016); it contains retellings by 29 speakers.

Tab. 2: Information pressure as a ratio of the number of discourse referents in a text and text length in clause units.

corpus	number of texts	mean ratio refs/clauses	SD
C. Greek	3	0.152	0.022
English	4	0.224	0.026
Mandarin	3	0.179	0.041
Nafsan	9	0.183	0.060
N. Kurdish	2	0.131	0.018
S. Dargwa	8	0.158	0.034
Teop	4	0.108	0.019
Tulil	6	0.171	0.050
Vera'a	10	0.122	0.025

We have coded the introduction strategies used for the five central human characters in the Pear film, this being:

- **picker:** the man picking pears,
- **goatherd:** the man leading a goat past the pear tree with the picker,
- **thief:** the boy with a bike stealing a basket of pears,
- **girl:** the girl on a bike who causes the boy to fall, and
- **three boys:** the three boys who help the first boy pick up the stolen pears,

Our classificatory schema has the following schema, with an additional flag for elaborations (not shown in Tables 8 and 9):

- **A:** the referent is introduced as the subject of a transitive clause (e.g. *a man is picking pears*);
- **S:** the referent is introduced as the subject of an intransitive clause, excluding motion predicates and presentationals (e.g. *three boys appear*)
- **S-motion:** the referent is introduced as the subject of a verb of motion (e.g. *a boy comes along*)
- **existential:** the referent is introduced in an existential/presentational construction (e.g. *there are three other boys*)
- **P:** the referent is introduced as the direct object of a transitive clause (e.g. *he sees a girl*)
- **other:** the referent is introduced in a non-core role (e.g. *it starts with a man picking pears*).

Certain characters (the man leading the goat and the girl on a bike) are not mentioned by some of the speakers; they are excluded from the counts and indicated by a dash in Tables 8 and 9.

Raw values for figures

Tab. 3: Raw values for Figure 1.

	A			S			P		
corpus	**new**	**all**	**%**	**new**	**all**	**%**	**new**	**all**	**P**
C. Greek	8	221	4	20	254	8	79	316	25
English	58	623	9	162	1029	16	293	1058	28
Mandarin	11	324	3	36	495	7	65	292	22
Nafsan	9	298	3	34	447	8	55	293	19
N. Kurdish	5	233	2	38	457	8	55	320	17
S. Dargwa	6	171	4	52	409	13	42	170	25
Teop	7	306	2	25	504	5	38	304	12
Tulil	16	226	7	32	469	7	47	316	15
Vera'a	18	770	2	66	1876	4	135	777	17
totals	138	3172		465	5940		809	3846	

	oblique			goal			other		
corpus	**new**	**all**	**%**	**new**	**all**	**%**	**new**	**all**	**P**
C. Greek	5	49	10	9	102	9	32	267	12
English	66	216	31	81	246	33	142	618	23
Mandarin	14	84	17	4	23	17	44	329	13
Nafsan	4	28	14	10	37	27	34	203	17
N. Kurdish	13	93	14	22	131	17	26	381	7
S. Dargwa	5	43	12	15	75	20	32	143	22
Teop	3	10	30	10	35	29	33	274	12
Tulil	16	92	17	17	79	22	79	501	16
Vera'a	18	99	18	68	345	20	90	903	10
totals	512	3619		236	1073		144	714	

Tab. 4: Raw values for Figure 2, for new (top) and all given (bottom) mentions.

new	A		S		P		oblique		goal		other		
corpus	N	%	N	%	N	%	N	%	N	%	N	%	all
C. Greek	8	5	20	13	79	52	5	3	9	6	32	21	153
English	58	7	162	20	293	37	66	8	81	10	142	18	802
Mandarin	11	6	36	21	65	37	14	8	4	2	44	25	174
Nafsan	9	6	34	23	55	38	4	3	10	7	34	23	146
N. Kurdish	5	3	38	24	55	35	13	8	22	14	26	16	159
S. Dargwa	6	4	52	34	42	28	5	3	15	10	32	21	152
Teop	7	6	25	22	38	33	3	3	10	9	33	28	116
Tulil	16	8	32	15	47	23	16	8	17	8	79	38	207
Vera'a	18	5	66	17	135	34	18	5	68	17	90	23	395
totals	236		144		138		465		512		809		2304

given	A		S		P		oblique		goal		other		
corpus	N	%	N	%	N	%	N	%	N	%	N	%	all
C. Greek	213	20	234	22	237	22	44	4	93	9	235	22	1056
English	565	19	867	29	765	26	150	5	165	6	476	16	2988
Mandarin	313	23	459	33	227	17	70	5	19	1	285	21	1373
Nafsan	289	25	413	36	238	21	24	2	27	2	169	15	1160
N. Kurdish	228	16	419	29	265	18	80	5	109	7	355	24	1456
S. Dargwa	165	19	357	42	128	15	38	4	60	7	111	13	859
Teop	299	23	479	36	266	20	7	1	25	2	241	18	1317
Tulil	210	14	437	30	269	18	76	5	62	4	422	29	1476
Vera'a	752	17	1810	41	642	15	81	2	277	6	813	19	4375
totals	570		837		3107		3034		3037		5475		16060

Tab. 5: Raw values for Figure 3, for 0–1 (top), 2 (middle) and 3 (bottom) other referents mentioned in the previous three clauses.

0–1 refs.	A		S		P		non-core		
corpus	N	%	N	%	N	%	N	%	all
C. Greek	0	0	4	29	6	43	4	29	14
English	3	8	13	36	11	31	9	25	36
Mandarin	0	0	4	27	5	33	6	40	15
Nafsan	2	8	10	38	1	4	13	50	26
N. Kurdish	0	0	7	64	1	9	3	27	11
S. Dargwa	1	5	9	41	4	18	8	36	22
Teop	3	15	6	30	6	30	5	25	20
Tulil	1	11	1	11	2	22	5	56	9
Vera'a	3	7	8	20	9	22	21	51	41

2 refs.	A		S		P		non-core		
corpus	N	%	N	%	N	%	N	%	all
C. Greek	3	10	2	6	15	48	11	35	31
English	10	8	36	30	41	34	35	29	122
Mandarin	0	0	6	24	7	28	12	48	25
Nafsan	1	2	7	18	15	38	17	42	40
N. Kurdish	0	0	11	41	11	41	5	19	27
S. Dargwa	2	5	10	23	16	36	16	36	44
Teop	1	4	5	21	6	25	12	50	24
Tulil	3	10	8	27	4	13	15	50	30
Vera'a	3	3	24	22	39	35	45	41	111

3 refs.	A		S		P		non-core		
corpus	N	%	N	%	N	%	N	%	all
C. Greek	2	4	7	16	25	56	11	24	45
English	15	7	38	18	84	39	76	36	213
Mandarin	5	13	8	21	12	32	13	34	38
Nafsan	3	8	8	20	19	48	10	25	40
N. Kurdish	3	6	9	17	20	38	21	40	53
S. Dargwa	1	3	18	46	10	26	10	26	39
Teop	2	6	6	18	12	35	14	41	34
Tulil	4	6	14	23	17	27	27	44	62
Vera'a	6	5	19	15	43	34	58	46	126

Tab. 6: Raw values for Figure 3, for 4 (top), 5 (middle) and 6+ (bottom) other referents mentioned in the previous three clauses.

4 refs.	A		S		P		non-core		
corpus	N	%	N	%	N	%	N	%	all
C. Greek	1	3	3	8	21	55	13	34	38
English	17	9	32	17	74	40	64	34	187
Mandarin	3	6	6	13	19	40	19	40	47
Nafsan	2	7	6	22	15	56	4	15	27
N. Kurdish	1	3	6	17	12	34	16	46	35
S. Dargwa	2	7	11	37	9	30	8	27	30
Teop	0	0	7	29	9	38	8	33	24
Tulil	5	11	4	9	12	27	23	52	44
Vera'a	3	5	10	15	25	38	27	42	65

5 refs.	A		S		P		non-core		
corpus	N	%	N	%	N	%	N	%	all
C. Greek	2	12	2	12	9	53	4	24	17
English	8	6	18	14	46	36	56	44	128
Mandarin	1	4	7	28	10	40	7	28	25
Nafsan	1	10	3	30	4	40	2	20	10
N. Kurdish	0	0	3	18	6	35	8	47	17
S. Dargwa	0	0	2	25	0	0	6	75	8
Teop	0	0	0	0	3	75	1	25	4
Tulil	2	5	3	8	9	24	24	63	38
Vera'a	1	3	5	13	15	39	17	45	38

6+ refs.	A		S		P		non-core		
corpus	N	%	N	%	N	%	N	%	all
C. Greek	0	0	2	25	3	38	3	38	8
English	5	4	25	22	37	32	49	42	116
Mandarin	2	8	5	21	12	50	5	21	24
Nafsan	0	0	0	0	1	33	2	67	3
N. Kurdish	1	6	2	12	5	31	8	50	16
S. Dargwa	0	0	2	22	3	33	4	44	9
Teop	1	10	1	10	2	20	6	60	10
Tulil	1	4	2	8	3	12	18	75	24
Vera'a	2	14	0	0	4	29	8	57	14

Tab. 7: Raw values for Figure 4.

	A			S			P		
corpus	**br.new**	**all**	**%**	**br.new**	**all**	**%**	**br.new**	**all**	**P**
C. Greek	4	8	50	11	20	55	41	79	52
English	14	58	24	48	162	30	111	293	38
Mandarin	8	11	73	21	36	58	47	65	72
Nafsan	6	9	67	22	34	65	43	55	78
N. Kurdish	3	5	60	20	38	53	23	55	42
S. Dargwa	1	6	17	23	52	44	18	42	43
Teop	3	7	43	9	25	36	15	38	39
Tulil	11	16	69	16	32	50	28	47	60
Vera'a	7	18	39	23	66	35	75	135	56
totals	57	138		193	465		401	809	

	oblique			goal			other		
corpus	**br.new**	**all**	**%**	**br.new**	**all**	**%**	**br.new**	**all**	**P**
C. Greek	1	5	20	4	9	44	13	32	41
English	29	66	44	41	81	51	54	142	38
Mandarin	12	14	86	4	4	100	32	44	73
Nafsan	2	4	50	9	10	90	26	34	76
N. Kurdish	6	13	46	4	22	18	7	26	27
S. Dargwa	1	5	20	7	15	47	14	32	44
Teop	0	3	0	4	10	40	16	33	48
Tulil	9	16	56	8	17	47	46	79	58
Vera'a	9	18	50	44	68	65	48	90	53
totals	69	144		125	236		256	512	

Tab. 8: Raw values for Figure 5, introduction strategies in the English Pear stories (Chafe 1980).

corpus	text	picker	goatherd	thief	girl	three boys
English	01	exist	S-mot	S-mot	P	P
English	02	exist	S-mot	S-mot	P	P
English	03	other	other	S-mot	P	exist
English	04	other	S-mot	S-mot	other	S-mot
English	05	exist	S-mot	S-mot	S-mot	S-mot
English	06	other	P	S-mot	S-mot	S-mot
English	07	exist	S-mot	S-mot	S-mot	exist
English	08	exist	—	S-mot	P	exist
English	09	exist	P	other	exist	S-mot
English	10	exist	S-mot	S-mot	S-mot	S-mot
English	11	exist	A	S-mot	S-mot	P
English	12	other	S	other	S-mot	exist
English	13	exist	P	other	S-mot	S-mot
English	14	exist	S-mot	other	S-mot	exist
English	15	other	other	S-mot	S-mot	S-mot
English	16	A	—	S-mot	P	S-mot
English	17	P	P	other	P	S
English	18	exist	S-mot	S-mot	P	exist
English	19	A	S-mot	S-mot	P	S
English	20	exist	S-mot	other	other	exist

Tab. 9: Raw values for Figure 5, introduction strategies in the Persian Pear stories (Adibifar 2016).

corpus	text	picker	goatherd	thief	girl	three boys
Persian	g1-f-01	A	S-mot	S-mot	—	S
Persian	g1-f-02	A	—	P	—	S
Persian	g1-f-05	exist	other	A	P	S
Persian	g1-f-07	exist	S-mot	S-mot	S-mot	S-mot
Persian	g1-f-08	exist	S-mot	S-mot	other	S
Persian	g1-f-09	P	S-mot	P	S-mot	P
Persian	g1-f-10	S-mot	S-mot	S-mot	S-mot	S
Persian	g1-f-11	P	—	S	P	—
Persian	g1-f-12	P	S-mot	S-mot	—	S-mot
Persian	g1-f-14	A	S-mot	S-mot	P	P
Persian	g1-m-03	A	—	S-mot	—	S-mot
Persian	g1-m-04	P	S	S-mot	P	exist
Persian	g1-m-06	A	—	S-mot	—	S-mot
Persian	g1-m-13	A	S-mot	S-mot	S-mot	S
Persian	g2-f-01	exist	other	S-mot	S-mot	S-mot
Persian	g2-f-02	exist	exist	S-mot	S-mot	S
Persian	g2-f-03	other	S-mot	S-mot	—	P
Persian	g2-f-04	A	—	S-mot	—	S-mot
Persian	g2-f-05	A	S-mot	S-mot	other	S-mot
Persian	g2-f-06	exist	—	S-mot	other	P
Persian	g2-f-07	exist	S-mot	S-mot	S	S-mot
Persian	g2-m-08	exist	P	S-mot	S-mot	S
Persian	g2-m-09	other	—	S-mot	S-mot	exist
Persian	g2-m-10	exist	S-mot	S-mot	—	exist
Persian	g2-m-11	S	—	exist	—	S-mot
Persian	g2-m-12	A	S-mot	S-mot	S-mot	S
Persian	g2-m-13	S-mot	—	S-mot	P	S-mot
Persian	g2-m-14	exist	—	S-mot	S-mot	S
Persian	g2-m-15	exist	—	S-mot	other	S

References

Adibifar, Shirin. 2016. Multi-CAST Persian. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#persian.

Chafe, Wallace (ed.). 1980. *The Pear Stories*. Norwood, NJ: Ablex.

Du Bois, John. 1987. Absolutive zero. *Lingua* 71(2). 203–222.

Forker, Diana & Nils N. Schiborr. 2019. Multi-CAST Sanzhi Dargwa. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#sanzhi.

Hadjidas, Harris & Maria C. Vollmer. 2015. Multi-CAST Cypriot Greek. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#cypgreek.

Haig, Geoffrey & Stefan Schnell. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse)*. https://multicast.aspra.uni-bamberg.de/#annotations.

Haig, Geoffrey & Stefan Schnell (eds.). 2015. *Multi-CAST*. https://multicast.aspra.uni-bamberg.de/.

Haig, Geoffrey, Maria C. Vollmer & Hanna Thiele. 2019. Multi-CAST Northern Kurdish. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#nkurd.

Meng, Chenxi. 2019. Multi-CAST Tulil. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#tulil.

Mosel, Ulrike & Stefan Schnell. 2015. Multi-CAST Teop. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#teop.

Riester, Arndt & Stefan Baumann. 2017. *The RefLex scheme — Annotation guidelines* (SinSpeC: Working papers of the SFB 732 14). Stuttgart: University of Stuttgart. http://elib.uni-stuttgart.de/handle/11682/9028.

Schiborr, Nils N. 2015. Multi-CAST English. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#english.

Schiborr, Nils N. 2018. multicastR. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://cran.r-project.org/package=multicastR.

Schiborr, Nils N., Stefan Schnell & Hanna Thiele. 2018. *RefIND — Referent Indexing in Natural-language Discourse*. Bamberg / Melbourne: University of Bamberg. https://multicast.aspra.uni-bamberg.de/#annotations.

Schnell, Stefan. 2015. Multi-CAST Vera'a. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#veraa.

Schnell, Stefan & Nils N. Schiborr. 2018. Corpus-based typological research in discourse and grammar. *Asian and African Languages and Linguistics* 12. 1–16. http://hdl.handle.net/10108/91145.

Thieberger, Nick & Timothy Brickell. 2019. Multi-CAST Nafsan. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#nafsan.

Vollmer, Maria. 2020. Multi-CAST Mandarin. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST*, https://multicast.aspra.uni-bamberg.de/#mandarin.