Multi-CAST

English

corpus counts

Nils Norman Schiborr
University of Bamberg

May 2019
v2.0

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

Australian Government
Australian Research Council

University of Bamberg

DFG

Multi-CAST *Multilingual Corpus of Annotated Spoken Texts*

Citation for this document

Schiborr, Nils N. 2019. Multi-CAST English corpus counts. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts.* (`multicast.aspra.uni-bamberg.de/`) (date accessed)

Citation for the Multi-CAST collection

Haig, Geoffrey & Schnell, Stefan (eds.). 2015. *Multi-CAST: Multilingual corpus of annotated spoken texts.* (`multicast.aspra.uni-bamberg.de/`) (date accessed)

The Multi-CAST collection has been archived at the *University of Bamberg*, Germany, and is freely accessible online at `multicast.aspra.uni-bamberg.de/`.

Multi-CAST English corpus counts v2.0 last updated 8 May 2019
This document was typeset by NNS with XeLaTeX and the *multicast3* class (v3.0.9045).

Contents

1 Notes on the GRAID counts

This document collects tables with frequency counts for combinations of selected GRAID symbols in version 1905 (from May 2019) of the Multi-CAST English corpus. The tables are intended to offer cursory impressions of the relative proportions between different types of referring expression; they do not provide exact summaries of the annotations.

Only a small number of basic GRAID symbols are counted:

Function symbols

⟨0⟩	zero
⟨pro⟩	definite pronoun
⟨np⟩	full noun phrase
⟨other⟩	form not further specified

Person/Animacy symbols

⟨.1⟩	first person
⟨.2⟩	second person
⟨.h⟩	third person, human
⟨.d⟩	third person, anthropomorphic
ø	third person, non-human

Function symbols

⟨:a⟩	subject of a transitive clause
⟨:s⟩	subject of an intransitive clause
⟨:ncs⟩	non-canonical subject
⟨:p⟩	direct object
⟨:obl⟩	oblique argument
⟨:g⟩	goal argument
⟨:l⟩	locational argument
⟨:poss⟩	possessive
⟨:pred⟩	predicate
⟨:other⟩	function not further specified

Clause boundary symbols

⟨##⟩	independent clause
⟨#⟩	other clause

Only basic categories are listed; categories represented by complex symbols with additional specifiers (e.g. ⟨dem_pro⟩ ‘demonstrative pronoun’) have been subsumed under the more basic category (e.g. ⟨pro⟩ ‘definite pronoun’). Please refer to the annotation notes for this corpus for information on all annotated categories, including those not listed here.

The tables in this document can be recreated with the `mc_table` function from *multicastR*, the companion package to Multi-CAST for the statistical computing language R. Please refer to the package documentation (`?multicastR`, `?mc_table`) for more information.

2 The English corpus

GRAID	⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	*totals*
⟨0 .1⟩	74	22	0	0	0	0	0	0	0	0	96
⟨0 .2⟩	27	3	0	0	0	0	0	0	0	0	30
⟨0 .h⟩	70	26	0	1	0	0	0	0	0	0	97
⟨0 .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨0⟩	10	24	0	39	0	1	0	0	38	1	113
⟨pro .1⟩	369	291	1	42	19	14	1	111	1	1	850
⟨pro .2⟩	104	51	1	8	7	9	0	20	0	0	200
⟨pro .h⟩	216	237	2	35	15	16	0	59	0	0	580
⟨pro .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩	35	234	2	325	43	21	7	22	17	17	723
⟨np .h⟩	63	81	0	38	28	20	2	11	36	4	283
⟨np .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨np⟩	19	79	0	525	149	131	161	9	133	167	1373
⟨other .h⟩	7	3	0	0	0	0	0	0	1	0	11
⟨other .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other⟩	2	15	0	42	4	65	60	0	277	25	490
totals	996	1066	6	1055	265	277	231	232	503	215	
⟨##⟩											990
⟨#⟩											1255
totals											2245

Table 1 Summarized GRAID counts for the entire English corpus.

2.1 kent01

GRAID	⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	*totals*
⟨0 .1⟩	19	6	0	0	0	0	0	0	0	0	25
⟨0 .2⟩	9	2	0	0	0	0	0	0	0	0	11
⟨0 .h⟩	20	4	0	0	0	0	0	0	0	0	24
⟨0 .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨0⟩	3	8	0	7	0	1	0	0	10	1	30
⟨pro .1⟩	85	55	1	10	4	4	1	31	0	1	192
⟨pro .2⟩	26	13	0	0	0	0	0	1	0	0	40
⟨pro .h⟩	80	61	1	14	5	5	0	31	0	0	197
⟨pro .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩	9	81	0	93	6	8	2	5	5	5	214
⟨np .h⟩	19	20	0	15	8	7	0	6	11	2	88
⟨np .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨np⟩	11	32	0	142	43	50	59	4	27	40	408
⟨other .h⟩	1	1	0	0	0	0	0	0	0	0	2
⟨other .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other⟩	1	1	0	9	0	14	18	0	92	6	141
totals	283	284	2	290	66	89	80	78	145	55	
⟨##⟩											378
⟨#⟩											244
totals											622

Table 2 Summarized GRAID counts for the *kent01* text.

2.2 *kent02*

GRAID	⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	*totals*
⟨0 .1⟩	55	16	0	0	0	0	0	0	0	0	71
⟨0 .2⟩	18	1	0	0	0	0	0	0	0	0	19
⟨0 .h⟩	50	22	0	1	0	0	0	0	0	0	73
⟨0 .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨0⟩	7	16	0	32	0	0	0	0	28	0	83
⟨pro .1⟩	284	236	0	32	15	10	0	80	1	0	658
⟨pro .2⟩	78	38	1	8	7	9	0	19	0	0	160
⟨pro .h⟩	136	176	1	21	10	11	0	28	0	0	383
⟨pro .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩	26	153	2	232	37	13	5	17	12	12	509
⟨np .h⟩	44	61	0	23	20	13	2	5	25	2	195
⟨np .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨np⟩	8	47	0	383	106	81	102	5	106	127	965
⟨other .h⟩	6	2	0	0	0	0	0	0	1	0	9
⟨other .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other⟩	1	14	0	33	4	51	42	0	185	19	349
totals	713	782	4	765	199	188	151	154	358	160	
⟨##⟩											612
⟨#⟩											1011
totals											1623

Table 3 Summarized GRAID counts for the *kent02* text.

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/