Multi-CAST

Mandarin

corpus counts

Maria Vollmer

January 2021
v1.1

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

Vollmer, Maria. 2021. Multi-CAST Mandarin corpus counts. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts.* (`multicast.aspra.uni-bamberg.de/#mandarin`) (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The Multi-CAST collection has been archived at the *University of Bamberg*, Germany, and is freely accessible online at `multicast.aspra.uni-bamberg.de/`.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 Mandarin corpus counts v1.1 last updated 27 January 2021
This document was typeset by NNS with XeLaTeX and the *multicast3* class (v3.2.3).

Contents

1 Notes on the GRAID counts

This document collects tables with frequency counts for combinations of selected GRAID symbols in version 2101 (from January 2021) of the Multi-CAST Mandarin corpus. Unless a more recent version of this document exists, it also applies to any later versions of the annotations. Note that the tables are intended to offer only 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.

2 The Mandarin corpus

GRAID	⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	*totals*
⟨0 .1⟩	48	30	0	2	3	0	0	0	0	0	83
⟨0 .2⟩	24	18	0	0	1	0	0	0	0	0	43
⟨0 .h⟩	205	223	0	28	0	2	0	0	0	0	458
⟨0 .d⟩	0	1	0	0	0	0	0	0	0	0	1
⟨0⟩	11	47	0	72	2	0	0	0	0	0	132
⟨pro .1⟩	29	20	0	7	5	3	0	20	0	2	86
⟨pro .2⟩	23	20	0	6	5	4	0	5	0	0	63
⟨pro .h⟩	48	50	0	14	11	0	0	37	0	1	161
⟨pro .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩	4	20	0	7	3	1	6	0	0	3	44
⟨np .h⟩	72	146	0	84	48	13	0	39	27	1	430
⟨np .d⟩	0	0	0	1	0	0	0	0	0	0	1
⟨np⟩	9	102	0	227	52	16	86	16	48	156	712
⟨other .h⟩	1	0	0	0	0	0	0	0	0	0	1
⟨other .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other⟩	0	3	0	12	1	3	1	0	156	0	176
totals	474	680	0	460	131	42	93	117	231	163	
⟨##⟩											1111
⟨#⟩											83
totals											1194

Table 1 Summarized GRAID counts for the entire Mandarin corpus.

2.1 *hml*

GRAID	⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	*totals*
⟨0 .1⟩	7	2	0	0	0	0	0	0	0	0	9
⟨0 .2⟩	0	3	0	0	0	0	0	0	0	0	3
⟨0 .h⟩	78	64	0	11	0	1	0	0	0	0	154
⟨0 .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨0⟩	2	8	0	6	0	0	0	0	0	0	16
⟨pro .1⟩	3	5	0	0	0	0	0	1	0	0	9
⟨pro .2⟩	0	0	0	0	1	2	0	0	0	0	3
⟨pro .h⟩	22	18	0	2	0	0	0	10	0	0	52
⟨pro .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩	0	4	0	1	1	0	1	0	0	0	7
⟨np .h⟩	27	32	0	34	18	6	0	20	4	0	141
⟨np .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨np⟩	3	19	0	76	12	4	18	7	11	44	194
⟨other .h⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other⟩	0	0	0	2	0	0	0	0	36	0	38
totals	142	155	0	132	32	13	19	38	51	44	

⟨##⟩	270
⟨#⟩	31
totals	301

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

2.2 *jgz*

GRAID	⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	*totals*
⟨0 .1⟩	39	22	0	2	3	0	0	0	0	0	66
⟨0 .2⟩	24	15	0	0	1	0	0	0	0	0	40
⟨0 .h⟩	99	118	0	15	0	1	0	0	0	0	233
⟨0 .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨0⟩	6	34	0	62	2	0	0	0	0	0	104
⟨pro .1⟩	24	15	0	6	4	3	0	17	0	2	71
⟨pro .2⟩	21	16	0	6	3	2	0	5	0	0	53
⟨pro .h⟩	13	15	0	12	5	0	0	17	0	1	63
⟨pro .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩	4	16	0	6	2	1	5	0	0	3	37
⟨np .h⟩	37	93	0	28	10	5	0	9	15	1	198
⟨np .d⟩	0	0	0	1	0	0	0	0	0	0	1
⟨np⟩	4	58	0	125	31	9	55	7	34	80	403
⟨other .h⟩	1	0	0	0	0	0	0	0	0	0	1
⟨other .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other⟩	0	1	0	9	0	3	0	0	81	0	94
totals	272	403	0	272	61	24	60	55	130	87	
⟨##⟩											674
⟨#⟩											37
totals											711

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

2.3 *lzh*

GRAID	⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	*totals*
⟨0 .1⟩	2	6	0	0	0	0	0	0	0	0	8
⟨0 .2⟩	0	0	0	0	0	0	0	0	0	0	0
⟨0 .h⟩	28	41	0	2	0	0	0	0	0	0	71
⟨0 .d⟩	0	1	0	0	0	0	0	0	0	0	1
⟨0⟩	3	5	0	4	0	0	0	0	0	0	12
⟨pro .1⟩	2	0	0	1	1	0	0	2	0	0	6
⟨pro .2⟩	2	4	0	0	1	0	0	0	0	0	7
⟨pro .h⟩	13	17	0	0	6	0	0	10	0	0	46
⟨pro .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩	0	0	0	0	0	0	0	0	0	0	0
⟨np .h⟩	8	21	0	22	20	2	0	10	8	0	91
⟨np .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨np⟩	2	25	0	26	9	3	13	2	3	32	115
⟨other .h⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other .d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other⟩	0	2	0	1	1	0	1	0	39	0	44
totals	60	122	0	56	38	5	14	24	50	32	
⟨##⟩											167
⟨#⟩											15
totals											182

Table 4 Summarized GRAID counts for the *lzh* text.

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/