Citation for this document

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The Multi-CAST collection has been archived at the University of Bamberg, Germany, and is freely accessible online at multicast.aspra.uni-bamberg.de/.

The entirety of Multi-CAST, including this document, is published under the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), unless noted otherwise. The licence can be reviewed online at creativecommons.org/licenses/by/4.0/.

Multi-CAST Mandarin annotation notes v1.0 last updated 12 January 2020
This document was typeset by NNS with XeLaTeX and the multicast3 class (v3.2.2).
Contents

1 Notes on the GRAID annotations 1
  1.1 Differential object marking 1
  1.2 Serial verb constructions 1
  1.3 Flexible word classes 3
  1.4 Topic constructions 3

References 4

Appendices 6
  A List of corpus-specific GRAID symbols 6
  B List of abbreviated morphological glosses 7
1 Notes on the GRAID annotations

This document describes the implementation of the GRAID (Haig & Schnell 2014) and RefIND (Schiborr et al. 2018) annotation conventions in the Multi-CAST Mandarin corpus. It corresponds to version 2001 of the annotations, published in January 2020. Unless a more recent version of this document exists, it also applies to any later versions of the annotations.

The texts in this corpus were recorded in Modern Standard Mandarin (MSM, officially referred to as Pǔtōnghuà, ‘common speech’), the national language of the People’s Republic China. Standard Mandarin is in many ways an artificial construct; an idealized form of the language has been taught to children in schools nationwide, but actual usage remains highly influenced by regional languages. The narratives in the corpus were recorded in Xi’an in Northwest China; two of the speakers are originally from Northeast China (Dōngběi), the third hails from Xi’an.

Mandarin belongs to the Sinitic branch of the Sino-Tibetan language family. It is a strongly isolating language, and is often described as topic-prominent (Li & Thompson 1981; 1976).

1.1 Differential object marking

In Mandarin, the canonical word order SVO (Iemmolo & Arcodia 2014: 316) may be changed by moving the object in front of the predicate and marking it with a preposition such as bā or gěi (Li & Thompson 1981: 463; Liu 2007).

(1) jǔ gěi ná gè chānpóér xià le yī tiào
  0_he 0.m:obl give DEM CL midwife scare ASP one jump
  ## 0_h:s other adp ln ln np.m:obl v:pred rv rv rv

‘He had already scared the midwife.’

[mc_mandarin_jgz_0105]

In GRAID, preverbal “objects” of this kind are glossed as ⟨obl⟩ ‘oblique arguments’ instead of ⟨p⟩ ‘direct objects’, since, from a strictly formal perspective, they are marked with an adposition and are thus not canonically (un)marked objects. While we do not think this pattern is differential object marking in the narrow sense, that is how it is commonly labelled in the literature (see e.g. Iemmolo & Arcodia 2014), and so we call so here for pragmatic reasons.

1.2 Serial verb constructions

Serial verb constructions are formally very similar to (and often indistinguishable from) topic chains in which multiple predicates occur as a string of verbs, in which co-referential argument(s) are covert. While there are various language-specific means of differentiating serial verb constructions from multiple predicates, often involving the scope of negation or TAM markers over the whole predicate instead of one single verb, in practice most occurrences of serial verb constructions in the Mandarin corpus are formally ambiguous and thus indistinguishable from topic chains with zero arguments:
In (3),  

In these cases, the constructions are thus glossed as multiple predicates with covert arguments and the specifier ⟨svc_⟩ is prefixed to the zero gloss ⟨⟩. This enables GRAID to capture as much information as possible, while still giving researchers the possibility to exclude these contentious zeros and thereby analyze the constructions as serial verb constructions.

Constructions of the kind in (2) and (3) are analyzed as a serial verb constructions only in cases where (A) the string of verbs clearly denotes a single event or action, or (B) analyzing the verbs as multiple predicates in a topic chain would change their meaning in a way that would be contextually incorrect. In these cases, the main verb is glossed ⟨v:pred⟩ and the other verb(s) are glossed ⟨lv_svc⟩ or ⟨rv_svc⟩ (i.e. as subconstituents of the main verb complex), as in the following example:

Here, we know from context that neither of the participants is “coming over”, as the subject of the clause is already in the right place, and the object of the clause is a newborn baby who cannot be the subject of guòlái. We hence treat this construction is a serial verb construction and guòlái changes its semantics to a simple directional meaning. In the three texts in the Mandarin corpus, there are a total of 71 instances of ⟨lv_svc⟩ and ⟨rv_svc⟩, and 60 instances of ⟨svc_0⟩ among all 589 zero arguments.
1.3 Flexible word classes

Mandarin exhibits relatively flexible word classes. Sun (2006: 206) notes that “[n]early all Chinese prepositions can be used as full-fledged verbs.” With regard to the corpus, this poses a problem for prepositions that also act as verbs and are often used in serial verb constructions. In these cases, the question is if they are to be annotated as verbs or as prepositions; and, if they are analyzed as verbs, if they are serial verb constructions or two separate predicates. This also affects the annotation of the argument after the verb, since it would be a direct object (⟨:p⟩) if analysed as a verb, but an oblique argument (⟨:obl⟩) if analyzed as a preposition. An example can be seen in (5):

(5) yùnyōu dào zánmén zhèēr
  0_he travel reach 1PL.INCL here

‘He has travelled to us.’

Here, dào could also be analysed as a fulll verb (receiving the gloss ⟨v:pred⟩), and the clause would then be analyzed as two separate clauses, adding a zero subject in the second clause. The word zhèēr, here annotated as a goal of motion (⟨:g⟩), would then be analysed as an object (⟨:p⟩). For the annotation of this corpus, we have chosen to analyze dào in this context as a preposition, since that is the primary use of the word and closest to the actual formal representation. For comparison, in (6) dào is used independently as a full verb:

(6) dào le dāngpū
  0_they reach ASP pawn_shop

‘[They] came to the pawn shop.’

1.4 Topic constructions

Mandarin is often referred to as a topic-prominent language, in which topic and comment (rather than subject and predicate) could be considered to form the fundamental structure of the clause (see e.g. Li & Thompson 1981: 15, 85). Topics may be separated from the rest of the clause by what Li & Thompson (1981: 86) call pause particles, such as ne in (7). Note that in this example, the topic is repeated as a subject in pronominal form.

(7) ér liāngshānbó ne tā zìjǐ yě juēde
  but Liangshanbo MP 3SG REFL also think

‘And Liangshanbo, he himself thought,...’

Where subjects are separated from the rest of the clause via pause particle and the subject is not repeated a zero gloss ⟨dt_0⟩ is added:
(8) zhè gè dàoji bā píngshí zài siyuàn lǐ niàn  
DEM CL Daoji MP 0_he usually in temple_yard in read  
## ln ln pn_np.h:dt_a other dt_0:h:a other adp np:1 adp v:pred  
niàn jīng  
read scriptures  
rv np:p  
'This Daoji, [he] usually read the scriptures in the temple yard.'  
[mc_mandarin_jgz_0197]

In (8), dàoji is separated from the rest of the clause with the pause particle bā, and is thus analysed as topic. Since the referent is not repeated overtly as a subject, as in (7), a zero gloss is added. No zero is added when the subject is a lexical noun phrase without pause marker (9) even though the subject may still be repeated in the pronounal form as in (10), since there is no formal topic marking on the subject, except for its leftmost placement in the clause.

(9) dànshì zhù yuánwài tā yǒu yì gè nǚ ér  
but Zhu landlord 3sg have one cl daughter  
## other pn_np.h:dt_a rn_np pro.h:a v:pred ln ln np.h:p  
'But Landlord Zhu, he had a daughter.'  
[mc_mandarin_lzh_0010]

When the object is preverbal and in the leftmost position of the clause, it is analyzed as the topic of the clause. In this case, a zero gloss is added (see (12) further below), as the object may be repeated in its usual position when it is the topic of the clause, as in (11), and it would come after the predicate according to canonical word order.

(11) zhè gè jìdiān jiù xiān bù guǎn tā le  
DEM CL Jidian 0_you MP first NEG care_about 3sg ASP  
## ln ln np.h:dt_p 0.2:a other other other v:pred pro.h:p other  
'This Jidian, do not care about him for now.'  
[mc_cypgreek_jgz_0398]

Here, jìdiān is in the leftmost position of the clause and then realized again after the predicate as a pronoun. It is thus the topic of the clause, while the pronoun is the object. In the next example, the topic is not repeated as the object and zero is added in the gloss:

(12) nǐ de lǎilì wǒ zhīdào  
2sg MOD origin 1sg know 0_it  
## ln.pro.2:poss ln np:dt_p pro.1:a v:pred dt_0:p  
'Your origin, I know [it].'  
[mc_mandarin_jgz_0137]

Since the cutoff point for these gradual distinctions is somewhat arbitrary in the corpus, and one might argue that the argument of the verb is in fact overt in the clause, the gloss is extended via a specifier to (dt_0), thus enabling researchers to exclude these instances of zero from analysis or change them in later versions.
References


Appendices

A List of corpus-specific GRAID symbols

The following is a list of the non-standard GRAID symbols used in the annotation of the Multi-CAST Mandarin corpus. Please refer to the GRAID manual (Haig & Schnell 2014: 54–55) for an inventory of basic GRAID symbols.

Form symbols and specifiers

(f0) structurally suppressed argument slot of a predicate
⟨dt_0⟩ zero subject of a topic construction marked by a pause particle
⟨svc_0⟩ omitted argument of a serial verb construction
⟨pn_np⟩ proper name
⟨intrg_other⟩ interrogative pronoun

Function symbols and specifiers

⟨:s_ds⟩ subject of a verb of speech, intransitive
⟨:a_ds⟩ subject of a verb of speech, transitive
⟨:obl_dom⟩ prepositional object with differential object marking

Subconstituent symbols

⟨_adj⟩ attributive adjective; attaches to ⟨ln⟩
⟨_cl⟩ classifier; attaches to ⟨ln⟩ and ⟨rn⟩
⟨_dem⟩ demonstrative determiner; attaches to ⟨ln⟩
⟨_deti⟩ indefinite determiner; attaches to ⟨ln⟩
⟨_num⟩ attributive numeral; attaches to ⟨lv⟩ and ⟨rv⟩
⟨_asp⟩ aspect marker; attaches to ⟨lv⟩ and ⟨rv⟩
⟨_neg⟩ negator; attaches to ⟨lv⟩ and ⟨rv⟩
⟨_svc⟩ secondary verb of a serial verb construction; attaches to ⟨lv⟩ and ⟨rv⟩
### List of abbreviated morphological glosses

<table>
<thead>
<tr>
<th></th>
<th>Abbreviation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>first person</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>second person</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>third person</td>
<td></td>
</tr>
<tr>
<td>ADP</td>
<td>adposition</td>
<td></td>
</tr>
<tr>
<td>ADV</td>
<td>adverb</td>
<td></td>
</tr>
<tr>
<td>ASP</td>
<td>aspect marker</td>
<td>le</td>
</tr>
<tr>
<td>CL</td>
<td>classifier</td>
<td></td>
</tr>
<tr>
<td>COP</td>
<td>copula</td>
<td></td>
</tr>
<tr>
<td>DEM</td>
<td>demonstrative</td>
<td></td>
</tr>
<tr>
<td>INCL</td>
<td>inclusive</td>
<td></td>
</tr>
<tr>
<td>MOD</td>
<td>modifier</td>
<td></td>
</tr>
<tr>
<td>MP</td>
<td>modal particle</td>
<td></td>
</tr>
<tr>
<td>NEG</td>
<td>negator</td>
<td></td>
</tr>
<tr>
<td>PASS</td>
<td>passive</td>
<td></td>
</tr>
<tr>
<td>PL</td>
<td>plural</td>
<td></td>
</tr>
<tr>
<td>REFL</td>
<td>reflexive</td>
<td></td>
</tr>
<tr>
<td>SG</td>
<td>singular</td>
<td></td>
</tr>
<tr>
<td>NC</td>
<td>not classified</td>
<td></td>
</tr>
</tbody>
</table>
Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/