Multi-CAST

collection overview

Nils Norman Schiborr

January 2020
v2.1

University of Bamberg

DFG

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

Citation for this document

Schiborr, Nils N. 2020. Multi-CAST collection overview. 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The Multi-CAST collection has been archived at the *University of Bamberg*, Germany, and is freely accessible online at `multicast.aspra.uni-bamberg.de/`.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 collection overview v2.1 last updated 12 January 2020
This document was typeset by NNS with XeLaTeX and the *multicast3* class (v3.2.2).

Contents

1 Introduction

Multi-CAST, the *Multilingual Corpus of Annotated Spoken Texts* (Haig & Schnell 2015),[1] is a collection of annotated texts from a typologically diverse set of languages. The texts in the collection are chiefly non-elicited and monologic narratives. Multi-CAST has been designed to enable cross-linguistic inquiries into referentiality and discourse structure by providing common ground for quantitative analyses,[2] in an effort to address questions posed by notions such as preferred argument structure (Du Bois 1987; 2003; 2017), referential density (Bickel 2003; Noonan 2003), and accessibility theory (Ariel 1988; 1990; 2004), among many others.

The collection is being compiled under the supervision of Geoffrey Haig and Stefan Schnell, and is freely accessible online from the servers of the University of Bamberg under an *Attribution Creative Commons* licence. Alongside transcriptions, an idiomatic English translation, and morphological glossing, the texts in Multi-CAST have been annotated within a shared framework, yielding a multi-level structure that lends itself to a variety of complex queries. The GRAID annotation scheme (Haig & Schnell 2014) has been designed for investigations at the intersection between discourse and grammar, and aims to be applicable to a typologically diverse spectrum of languages, offering a uniform set of tags and a simple combinatory syntax. A growing subset of texts in the collection additionally features referent identification with the RefIND scheme (Schiborr et al. 2018).

As of January 2020, the collection comprises data from 12 languages, encompassing roughly 14 hours of recordings, 21 500 clause units, and 95 000 words across 99 individual texts. Each corpus in the collection is treated as its own contribution, and is hence an individually citable resource. For all texts, extensive background information on the recordings and annotations is provided. The transcriptions, translations, and annotations are available in a variety of digital formats, including as EAF files, a file format used by the free linguistic annotation software ELAN,[3] and as XML and tab-separated values (TSV) files. The *multicastR* package (Schiborr 2018) provides a simple interface for accessing the Multi-CAST annotation data directly in the statistical computing language R.[4]

This collection overview serves to document the contents of Multi-CAST, its structure, and some of the decisions that went into its design: it provides descriptions of the annotations applied to the texts (Section 2), top-level summaries of the various corpora and texts therein (Sections 3 and 4), as well as technical outlines of the available file formats (Section 5) and a short user's guide to the *multicastR* package (Section 6). Appendices A and B contain lists of metadata on the texts and speakers, and a timeline of changes and additions to the collection can be found in Appendix C.

2 Annotations

The texts in Multi-CAST comprise spoken language, and have been annotated in accordance with the basic standards of spoken corpus annotations, including an orthographic transcription, a free translation into English, and morpheme-by-morpheme glossing as per the *Leipzig Glossing Rules*.

1 `multicast.aspra.uni-bamberg.de/`
2 See Mettouchi et al. (2015) for a similar approach towards developing corpora from African languages for comparative purposes.
3 `tla.mpi.nl/tools/tla-tools/elan`
4 `cran.r-project.org/`

A specific feature of Multi-CAST are the following three layers of annotations, which extend the standard morphological annotation of the texts:

- morphosyntactic annotations with the GRAID scheme (*Grammatical relations and animacy in discourse*, Haig & Schnell 2014),
- referent identification with the RefIND scheme (*Referent indexing in natural-language discourse*, Schiborr et al. 2018), and
- information status of referents with the ISNRef scheme (*Information status of new referents*, Schiborr et al. 2018: 15), an adaptation of the RefLex scheme (Riester & Baumann 2017).

These three layers of annotation make the data in Multi-CAST particularly valuable for cross-linguistic and comparative research into the realm of referentiality and the intersection between discourse and grammar.

This section briefly outlines the principles of the annotation schemes. For comprehensive descriptions, please refer to to the manuals for the three systems, all available from the Multi-CAST website. A detailed discussion of research motivations can be found in the *Multi-CAST research context* (Haig & Schnell 2016a).

2.1 GRAID: morphosyntactic annotations

GRAID stands for *Grammatical Relations and Animacy in Discourse* (Haig & Schnell 2014). This annotation scheme has been designed specifically for the quantitative corpus-based investigation of discourse and grammar and the interrelations between them. The focus of GRAID is the expression of participants in states of affairs as they occur in connected discourse features, marking their form, semantics, and syntactic function. The general template of a GRAID "word" is as follows:

(1) ⟨ *form* . *person/animacy* : *function* ⟩

In example (2) below, the subject of the intransitive predicate ⟨:s⟩ is a full noun phrase ⟨np⟩ with a human referent ⟨.h⟩. A further oblique argument is the PP that expresses the semantic role of a goal, and it is the complement that carries the GRAID gloss. Its form is also that of an NP ⟨np⟩, and ⟨:g⟩ is its syntactic function gloss. Note also that GRAID glosses are aligned with words, but functionally apply to whole phrases, and where these consist of more than one word this is also noted, see Haig & Schnell (2014: 28–30).

(2) *A stranger went into the garden*
ln np.h:s v:pred adp ln np:g

A crucial aspect of GRAID annotations is that they also register zero arguments in cases where a specific referent that is (i) licensed by the predicate, (ii) retrievable from context, and (iii) in principle be expressible is nevertheless left unexpressed. This is the case in (3), the putative continuation of (2), where the subject of the second, transitive clause is left unexpressed, being coreferential with the preceding subject.

(3) a. *A stranger went into the garden*
ln np.h:s v:pred adp ln np:g

b. *and picked some flowers*
other 0.h:a v:pred ln np:p

GRAID has a set of three basic form classes: full descriptions or lexical NPs ⟨np⟩, proforms or person indices ⟨pro⟩, and zero anaphora ⟨0⟩. In terms of syntactic functions, we distinguish between transitive and intransitive clauses, treating the subject of the latter as ⟨:s⟩ and the agent-like argument of the former as ⟨:a⟩; the patient-like arguments of transitive clauses are glossed ⟨:p⟩. Non-core arguments receive certain other function glosses; see Haig & Schnell (2014: 14–16) for a discussion of non-core arguments and their annotation. With regards to semantics, we also distinguish person and humanness, as shown in (4):

(4)
```
      and   then  she      gave    them  to   me.
## other other pro.h:a v:pred pro:p adp pro.1:g
```

GRAID has symbols for first ⟨.1⟩ and second ⟨.2⟩ person; if neither is used, we assume the third person. Within the third person, we distinguish human ⟨.h⟩ and non-human referents; for the latter no symbol is used, meaning the slot is left blank. Note that humanness is a feature value that is assumed to be entailed in the first and second person. In addition to the symbols outlined here, GRAID defines a small number of additional categories that enable further distinctions, see Haig & Schnell (2014) for a full listing.

A crucial aspect of GRAID annotations is that they have been designed to be applicable and comparable across diverse languages. This is achieved by drawing formal distinctions on a fairly general level. GRAID otherwise relies on functional distinctions that can be assumed to be identifiable across languages. For instance, the form gloss ⟨pro⟩ is used for various types of person indices such as personal pronouns, but also where a subset of demonstratives is used pronominally as an anaphoric expression (see Haig & Schnell 2016a: 9 for further explanation). In terms of syntactic functions, we adapt Andrews' (2007: 135–140) definition of the core argument functions S, A, and P, which in essence combines cross-linguistically determinable semantic prototype features like proto-agent and proto-patient, the identity of encoding of arguments (but not the particular encoding itself), and their number. See Haig & Schnell (2014: 12–14) for a discussion.

Once a sufficient stretch of discourse from a language or set of languages has been annotated, the GRAID system enables the analysis of relevant combinations of values in the three slots. For instance, one can easily determine the overall number of transitive subjects based on the function gloss ⟨:a⟩, and then determine the proportion of lexical NPs within this function by taking the number of glosses that contain both the form gloss ⟨np⟩ and the function ⟨:a⟩ and relating it to the total. This process can be repeated for all core argument functions A, S, and P, and the proportions then compared to each other. This is the essence of the procedure of Haig & Schnell's (2016b) critical assessment of Du Bois (1987) preferred argument structure hypothesis. Similarly, the proportion of zero forms in S and A functions (i.e. "subjects") can be identified, which then allows estimation of the degree to which a language displays what has traditionally been called "pro-drop".

2.2 RefIND: referent identification

RefIND stands for *Referent indexing in natural-language discourse* (Schiborr et al. 2018). This annotation scheme is at a glance comparatively simple in its design, comprising solely of numerical identifiers for unique discourse referents. The central idea of RefIND is that the same identifier is used for all mentions of a specific discourse referent in a text. Whenever a referent is newly introduced into the discourse, it is assigned a new identifier. (5) extends the earlier examples (2–4) with referent indices:

(5) a. *A stranger went into the garden*
```
## ln np.h:s  v:pred adp ln  np:g
      0001                  0002
```

b. *and picked some flowers,*
```
## other 0.h:a v:pred ln   np:p
         0001              0003
```

c. *and then she gave them to me.*
```
## other other pro.h:a v:pred pro:p adp pro.1:g
               0001           0003      0000
```

Discourse referents receive a four-digit number in the order of their first mention in a given text. Hence in (5), the stranger receives the index ⟨0001⟩ on each of the three occasions of being mentioned. Likewise, the flowers are taken up again by a pronoun, and hence both mentions receive the identifier ⟨0004⟩. Lastly, the index ⟨0000⟩ is by convention assigned to the narrator of the text.

The crucial challenge in annotating with RefIND involves determining whether a given nominal expression encodes a reference to an entity that is likely to be tracked in the following discourse. The RefIND guidelines (Schiborr et al. 2018) provide relevant guidelines for this issue.

Referent indexes are aligned with the respective GRAID glosses, and texts annotated with both can thus be analysed for various anaphoric relationships and expressions. Referential forms and syntactic functions can thus easily be read the associated GRAID glosses. For instance, by using RefIND in conjunction with GRAID, it is possible to calculate anaphoric distances and chart the continuity of specific syntactic functions across mentions, in order to then determine the association of these properties with different types of referring expression. Moreover, RefIND annotations enable identification of the first mention of referents, a point again relevant to the aforementioned notion of preferred argument structure (Du Bois 1987). The capture of the referential properties of first mentions is enhanced by combining RefIND with ISNRef, which we turn to in the next section.

2.3 ISNRef: information status of new referents

ISNRef stands for *Information Status of New Referents*. ISNRef extends the annotations with RefIND by charting the information status of referents at the point of their introduction into discourse (see Schiborr et al. 2018: 15). It is in essence a drastically reduced version of the RefLex annotation scheme designed by Riester & Baumann (2017) for the purpose of addressing questions of referential choice and referent tracking in the tradition of Halliday & Hasan (1976), Prince (1981), and many others.

The ISNRef scheme notes whether a newly mentioned referent is in some way evoked by the context, in which case it is labelled as a ⟨bridging⟩ anaphor, or not. In the latter case, we label it either as brand ⟨new⟩ or as known but ⟨unused⟩, depending on the assumptions of general knowledge in a given speech community.

(6) a.
```
      A  stranger went   into the garden
## ln np.h:s   v:pred adp  ln  np:g
      0001                     0002
      new                      bridging
```
b.
```
      and         picked some flowers,
## other 0.h:a v:pred ln   np:p
         0001             0003
                          bridging
```
c.
```
      and   then  she      gave   them  to  me.
## other other pro.h:a v:pred pro:p adp pro.1:g
               0001           0003      0000
```

As can be seen in (6), ISNRef glosses are aligned with the corresponding referent indices, and hence indirectly also with the GRAID glosses.

3 Corpus languages

As of January 2020, the Multi-CAST collection comprises data from 12 languages: Arta, Cypriot Greek, English, Mandarin, Nafsan, Northern Kurdish, Persian, Sanzhi Dargwa, Teop, Tondano, Tulil, and Vera'a. It encompasses 99 individual texts and roughly 14 hours of recordings, 21 500 clause units, and 95 000 words. Each corpus in the collection is treated as its own contribution, and is hence an individually citable resource with the annotators as authors.

This section provides a brief outline of the various Multi-CAST corpora. Table 1 summarizes selected corpus statistics, and the map in Figure 1 offers a geographical overview of the included languages. Comprehensive metadata on the texts and speakers can be found in Appendices A and B. The 'glottocodes' listed below reference entries within *Glottolog* (Hammarström et al. 2019).[5] 'Identifiers' are the corpus labels used internally in Multi-CAST. For an explanation of the versioning system used by Multi-CAST, see Section 4.3.

3.1 Arta

Yukinori Kimoto

glottocode	arta1239
affiliation	Austronesian, Malayo-Polynesian, Northern Luzon
area spoken	Quirino Province, Luzon, the Philippines
varieties rec'd	Arta
text types	traditional narratives, autobiographical narratives
sources	Kimoto 2017
identifier	arta
availability	since August 2019, version 1908
GRAID	7.0 (≥ 1908)
RefIND	✘
ISNRef	✘

5 glottolog.org/

Figure 1 The Multi-CAST corpora.

corpus	*identifier*	*text types* TN	AN	SN	*no. of texts*	*length in h:mm:ss*	*clause units*
Arta	`arta`	◆	◆	–	11	1:21:30	1 030
Cypriot Greek	`cypgreek`	◆	–	–	3	—	1 071
English	`english`	–	◆	–	5	3:55:55	5 649
Mandarin	`mandarin`	◆	–	–	3	0:39:53	1 194
Nafsan	`nafsan`	◆	–	–	9	0:38:11	1 012
Northern Kurdish	`nkurd`	◆	–	–	3	0:51:57	1 841
Persian	`persian`	–	–	◆	29	0:52:32	1 418
Sanzhi Dargwa	`sanzhi`	◆	◆	–	8	0:40:09	1 066
Teop	`teop`	◆	–	–	4	0:46:35	1 302
Tondano	`tondano`	–	◆	◆	8	1:15:58	1 085
Tulil	`tulil`	◆	◆	–	6	1:13:51	1 264
Vera'a	`veraa`	◆	–	–	10	2:01:48	3 608
collection totals					99	14:18:19	21 540

Table 1 Overview of the Multi-CAST corpora as of January 2020, version `2001`.
TN = traditional narratives, AN = autobiographical narratives, SN = stimulus-based narratives.

citation Kimoto, Yukinori. 2019. Multi-CAST Arta. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (`multicast.aspra.uni-bamberg.de/#arta`)

Arta is a critically endangered Austronesian language spoken by a group of hunter-gatherers living in Luzon, the Philippines. The number of fluent speakers is between nine and eleven, most of which are over the age of forty. Since all speakers have settled down in the communities of neighbouring Negrito groups (Casiguran/Nagitupunan Agta people), the language is not in active use and no longer taught to children. All of the speakers are multilingual with Casiguran/Nagtipunan Agta and Ilokano.

The texts in this corpus were collected by Yukinori Kimoto during fieldwork in the Quirino and Aurora provinces in Luzon between 2012 and 2018. See Kimoto (2017) for a description of the language.

Background to the recordings

alisiya Speaker AR01. The speaker talks about how she fell ill when she was young, and how her illness and the lifelong paralysis that resulted from have affected her life.

arsenyo Speaker AR02. The speaker talks about his best friend Arsenyo (AR03), telling an impressive story about him, who, among others, took care of the speaker during their hunting trips together.

child Speaker AR03. The speaker talks about the difficulties he and his wife faced raising their children as a result of poverty, lack of schooling, and insufficient medical care.

delia Speaker AR01. An autobiography. The speaker tells stories about how badly he behaved when he was young, about how he married the present wife, and about the influence of religious missionaries.

disubu Speaker AR03. A description of the food the speaker and his contemporaries used to eat in their childhood. It also includes a description about the different activities conducted by men and women in their own hunting and gathering societies.

hapon Speaker AR02. The speaker shares his father's stories of the hardships endured during Japanese occupation of the Philippines and the Pacific War, when the Arta people were forced to hide in the forests near their villages for fear of their lives.

husband Speaker AR01. The speaker talks about her late husband, telling several stories about him, including one involving the New People's Army.

marry Speaker AR02. A message to a newly married couple. The speaker speaks about the social norms they should follow, and advises them to always be considerate of each other.

swateng Speaker AR03. A folk story about a man called Sanuwateng, who came to the lowlands to marry an Arta girl. Because of his prolonged absence following his courtship, she decides to marry another man, which leads to a tragic and bloody ending.

typhoon Speaker AR01. A narrative about the typhoon that hit the Arta community in August 2013. The speaker is telling how the whole community dealt with the natural disaster during and after the typhoon.

udulan Speaker AR03. Two short folk stories about two men: Udulan is the main character of the story of a marriage between two different Negrito groups from the eastern and western sides of Sierra Madre, and Sanuwateng is the villain of a tragic story of intertribal marriage, the longer version of which is found in the text *swateng*.

3.2 Cypriot Greek

Harris Hadjidas, Maria Vollmer

glottocode	`cypr1239`
affiliation	Indo-European, Greek, Attic
area spoken	Cyprus
varieties rec'd	Yeri-Pyroi
text types	traditional narratives
identifier	`cypgreek`
availability	since May 2015, version `1505`
GRAID	7.0 (≥ `1505`)
RefIND	✓ (≥ `1905`)
ISNRef	✓ (≥ `1905`)
citation	Hadjidas, Harris & Vollmer, Maria C. 2015. Multi-CAST Cypriot Greek. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (`multicast.aspra.uni-bamberg.de/#cypgreek`)

Cypriot Greek is the variety of Greek spoken in Cyprus. The three texts in this subcorpus, all of which are traditional narratives, were originally recorded in the 1960s, and later compiled and published by Konstantinos Giangoullis as part of a book of traditional Cypriot tales (Giangoullis 2009):

- *jitros* from pages 51–53,
- *minaes* from pages 47–51, and

- *psarin* from pages 84–88.

The speaker in these texts, Elenis Mich (CG01), grew up and spent her life in the village of Yeri-Pyroi, near Nicosia. Unfortunately, no recordings are available for the texts. They appear to have been only minimally edited, and reflect reasonably faithfully the spoken language used in traditional narratives. The author of the text collection, Konstantinos Giangoullis, has kindly given his permission for the three texts to be made freely available in Multi-CAST.

The texts were originally transliterated into the roman alphabet and translated into English by a native speaker, Harris Hadjidas, who also conducted an initial round of syntactic annotation with an earlier version of GRAID. A second round of annotation, adhering to the guidelines of the GRAID 7.0, was completed by Maria Vollmer under supervision of Geoffrey Haig.

3.3 English

Nils Norman Schiborr

glottocode	`sout3282`
affiliation	Indo-European, Germanic, West
area spoken	United Kingdom
varieties rec'd	Southeast and South England
text types	autobiographical narratives
sources	Huddleston & Pullum 2002
identifier	`english`
availability	since May 2015, version `1505`
GRAID	7.0 (≥ `1505`)
RefIND	✓ (≥ `1908`)
ISNRef	✓ (≥ `1908`)
citation	Schiborr, Nils N. 2015. Multi-CAST English. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (`multicast.aspra.uni-bamberg.de/#english`)

The Multi-CAST English corpus contains autobiographical narratives taken from the Freiburg English Dialect Corpus (FRED, English Dialects Research Group 2005),[6] which has been compiled under the supervision of Bernd Kortmann and Lieselotte Anderwald at the University of Freiburg from texts recorded during the 1970s and 80s as part of various oral history projects. Session name correspondences between Multi-CAST and FRED are as follows:

- *devon01* *DEV_002*
- *kent01* *KEN_002*
- *kent02* *KEN_002*
- *kent03* *KEN_004*
- *london01* *LND_006*, *LND_007*

The texts annotated for Multi-CAST were recorded with older working-class speakers from southern and southeastern England. They depict everyday scenes and personal experiences from the speakers' lives: recurring topics include agriculture, animal husbandry, shipwrighting, work in the London docks, and the two World Wars.

6 Note that the audio recordings in this corpus are in the public domain, and thus do not fall under the *Creative Commons* licence applied to the annotations and the rest of Multi-CAST.

3.4 Mandarin

Maria Vollmer

glottocode	mand1415
affiliation	Sino-Tibetan, Sinitic
area spoken	People's Republic of China
varieties rec'd	Pǔtōnghuà, Xī'ān
text types	traditional narratives
identifier	mandarin
availability	since January 2020, version 2001
GRAID	7.0 (≥ 2001)
RefIND	✓ (≥ 2001)
ISNRef	✓ (≥ 2001)
citation	Vollmer, Maria C. 2020. Multi-CAST Mandarin. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (multicast.aspra.uni-bamberg.de/#mandarin)

The Multi-CAST Mandarin corpus consists of traditional narratives from three native speakers of Modern Standard Mandarin (MSM, officially referred to as *Pǔtōnghuà*, 'common speech'). 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 Xī'ān in Northwest China; two of the speakers are originally from Northeast China (Dōngběi), the third hails from Xī'ān.

The recordings were made by Maria Vollmer during an exchange semester in 2015 and 2016, transcribed by Liu Ruoyu in 2016 and 2017 under the supervision of Maria Vollmer, and subsequently translated, glossed, and annotated with GRAID between 2016 and 2019 by Maria Vollmer. Annotations with RefLex and ISNRef were added by Maria Vollmer and Adrian Kuqi in 2019. Further stories have been recorded and transcribed and are planned to be added to the corpus in the future.

3.5 Nafsan

Nick Thieberger, Timothy Brickell

glottocode	sout2856
affiliation	Austronesian, Malayo-Polynesian, Oceanic, Vanuatu, Central
area spoken	Vanuatu, Central Vanuatu, Efate
varieties rec'd	Efate
text types	traditional narratives
sources	Thieberger 2006
identifier	nafsan
availability	since August 2019, version 1908
GRAID	7.0 (≥ 1908)
RefIND	✓ (≥ 1908)
ISNRef	✓ (≥ 1908)
citation	Thieberger, Nick & Brickell, Timothy. 2019. Multi-CAST Nafsan. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (multicast.aspra.uni-bamberg.de/#nafsan)

The Nafsan language, also known as South Efate, is a Southern Oceanic language spoken on the island of Efate in central Vanuatu. As of 2005, there are approximately 6 000 speakers of Nafsan living in coastal villages from Pango to Eton. A description of the language can be found in Thieberger (2006).

The Multi-CAST Nafsan corpus constitutes a subset of the material collected by Nick Thieberger for his PhD research over three periods of fieldwork in the villages of Eratap and Erakor in South Efate between 1995 and 2000, and during subsequent trips. The entirety of the data has been archived in PARADISEC,[7] and can also be accessed via ANNIS.[8] See further Thieberger (2004). Session name correspondences with Multi-CAST are as follows:

- *kori* *094* 'A devil at Nguna'
- *lelep* *092* 'Tabu stories'
- *lisau* *077* 'Lisau'
- *litog* *075* 'Litong'
- *maal* *024* 'The hawk and the owl'
- *nmatu* *013* 'The pig wife'
- *ntwam* *019* 'The devil pig'
- *taapes* *078* 'The chicken and the swamphen'
- *tafra* *023* 'A story of a whale'

The texts were glossed with GRAID by Nick Thieberger and Timothy Brickell, and subsequently annotated with RefIND by Adrian Kuqi under supervision of Stefan Schnell.

3.6 Northern Kurdish

Geoffrey Haig, Maria Vollmer, Hanna Thiele

glottocode	`nort2641`
affiliation	Indo-European, Iranian, Northwestern
area spoken	eastern Turkey; northern Iraq; western Iran
varieties rec'd	Northern Kurmanji, Erzurum and Muş
text types	traditional narratives
sources	Haig 2018; Haig & Öpengin 2018; Öpengin & Haig 2014
identifier	`nkurd`
availability	since May 2015, version `1505`
GRAID	7.0 (≥ `1505`)
RefIND	(✓) (≥ `1907`)
ISNRef	(✓) (≥ `1907`)
citation	Haig, Geoffrey & Vollmer, Maria & Thiele, Hanna. 2019. Multi-CAST Northern Kurdish. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (`multicast.aspra.uni-bamberg.de/#nkurd`)

Northern Kurdish, also known as Kurmanjî, is a Northwest Iranian language spoken in eastern Turkey, Iraq, Syria, and parts of western Iran. The three texts recorded here are traditional narratives, from a female and a male speaker who grew up near the townships of Erzurum and Muş in eastern Turkey, respectively.

7 `catalog.paradisec.org.au/collections/NT1`
8 `gerlingo.com/language_detail.php?langID=6`

The texts were recorded in Germany in the 1990s and early 2000s, and subsequently transcribed, translated, and annotated by Geoffrey Haig, Abdullah Incekan, Hanna Thiele, and Maria Vollmer. A description of the language can be found in Haig (2018).

Background to the recordings

muserz01, muserz03 These two texts were recorded by Geoffrey Haig with a speaker called Miheme (NK01), who grew up in a village near Muş. The speaker had left Turkey approximately ten years previously and had since settled in Germany. The recordings were made in Miheme's allotment garden in Kiel, North Germany, in the company of his wife and another friend of the family. Geoffrey Haig made a long series of recordings with Miheme, most of which have been transcribed and translated by Geoffrey Haig with the assistance of native speakers.

The stories are Miheme's renderings of traditional Kurdish folkloric texts. Although not a trained storyteller, Miheme relished the opportunity to tell these stories, most of which he was recalling from childhood memories. He had no qualms about embellishing them in various ways when his memory failed him. His Kurdish is quite strongly influenced by Turkish, his main language of communication over the past two decades, but he is undoubtedly a fluent speaker of Kurmanji.

muserz02 This text was recorded by Abdullah Incekan in 2002 in Essen, Germany; the speaker is his grandmother Güllü Tunç (NK02), who was visiting Germany at the time. The atmosphere was relaxed; a number of family members including small children were present during the recordings. The speaker is a monolingual Kurmanji speaker who has spent her lifetime in a village of the region Tekman, south of Erzurum. The text was transcribed by Abdullah Incekan and Geoffrey Haig, and translated by Geoffrey Haig.

As regards content, this text is undubitably related to the well-known fairy tale Cinderella, and contains key motifs such as the evil stepmother, the slipper, the prince and so on, but the latter part of the story seems to stem from a different source, and at times the narrative lacks coherence.

3.7 Persian

Shirin Adibifar

glottocode	`tehr1242`
affiliation	Indo-European, Iranian, Southwestern
area spoken	Iran
varieties rec'd	Farsi, Tehran and Sari
text types	stimulus-based narratives
identifier	`persian`
availability	since June 2016, version `1606`
GRAID	7.0 (≥ `1606`)
RefIND	✘
ISNRef	✘
citation	Adibifar, Shirin. 2016. Multi-CAST Persian. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (`multicast.aspra.uni-bamberg.de/#persian`)

Persian is an Iranian language with official variants spoken in Iran, Afghanistan, and parts of Tajikistan; the variety spoken in Iran is also referred to as Farsi.

The texts in this corpus are narrative retellings of the Pear film (Chafe 1980), a roughly five minute-long short film about a boy stealing the fruit a man had been picking. The recordings were made by Shirin Adibifar in Tehran and locations in province of Mazandaran in 2015. In total, there are 29 recordings, each from a different native speaker of Persian, 17 of which are female and 12 male; the median age is 25, with a range of 20 to 39. All speakers have received at least some measure of university-level education.

Each text was produced in an interview-like setting, in which the corpus compiler (Adibifar) showed the speaker a 6-minute video (the *Pear Story*, cf. Chafe 1980) on a Dell color laptop computer with a 14-inch screen. At the end of the film, the interviewees were asked to recount the events of the film in their own words. The instructions were given by the researcher in their native language, Persian, and each participant received the same set of instructions. The interval between speakers watching of the movie and retelling its contents was less than five minutes. The participants were also asked to provide basic information regarding their age, gender, level of education, places of socialization, languages of communication (inside and outside of domestic settings), the language(s) of their parents, as well as their contact information in case of further questions.

The first half of the recordings (with *g1* in the text name) took place in a relaxed domestic setting in the interviewer's hometown in the province of Mazanderan in northern Iran, and in three cases, in the speakers' apartments in Tübingen, Germany. The remainder (with *g2* in the text name) were conducted with students from the Islamic Azad University in Tehran and Behšahr University in Mazanderan Province in seminar rooms of the two universities.

3.8 Sanzhi Dargwa

Diana Forker, Nils Norman Schiborr

glottocode	sanz1248
affiliation	Nakh-Daghestanian (Caucasian), Dargwa, Southern Dargwa
area spoken	Druzhba town, central Daghestan, Russia
varieties rec'd	Sanzhi
text types	traditional narratives, autobiographical narratives
sources	Forker Under revision
identifier	sanzhi
availability	since May 2019, version 1905
GRAID	7.0 (≥ 1905)
RefIND	✓ (≥ 1905)
ISNRef	✓ (≥ 1905)
citation	Forker, Diana & Schiborr, Nils N. 2019. Multi-CAST Sanzhi Dargwa. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (multicast.aspra.uni-bamberg.de/#sanzhi)

Sanzhi Dargwa is a Nakh-Daghestanian (Caucasian) language from the Dargwa subbranch. From 1968 onwards, over a relatively short span of time, all Sanzhi speakers left their village of Sanzhi in the mountains of central Daghestan, Russia, to move to linguistically and ethnically heterogeneous settlements in the lowlands, mostly to the town of Druzhba. Today Sanzhi is spoken by approximately 250 speakers, and heavily endangered.

The eight texts in this corpus comprise a small subset of the material that was recorded, transcribed, translated, and glossed by Diana Forker and other researchers with the assistance of

Gadzhimurad Gadzhimuradov, a native speaker, as part of a DOBES language documentation project (2012–2019). The entirety of the data has been archived at the Language Archive of the MPI, and is available on request.[9] A subcorpus of around ten hours has been fully glossed and translated into Russian and English, and is freely accessible online.[10] Session name correspondences between the Language Archive and Multi-CAST are as follows:

- *asabali* *Sanzhi_04_08_2013_DF_003*
- *bazhuk* *Sanzhi_03_08_2013_DF_001*
- *dragon* *Sanzhi_03_08_2013_DF_002*
- *kurban* *Sanzhi_26_07_2011_RM_005*
- *mill* *Sanzhi_30_08_2013_HM_001*
- *patima* *Sanzhi_19_03_2013_DF_001*
- *ramazan* *Sanzhi_08_08_2012_RMDF_004*
- *tape* *Sanzhi_26_07_2011_RM_010*

A grammar of Sanzhi Dargwa based on the collected material (Forker Under revision) is currently under revision. The texts chosen for Multi-CAST are a mixture of spontaneously produced autobiographical and traditional narratives. They were annotated for Multi-CAST by Nils Schiborr.

Background to the recordings

asabali Speaker SD01. Recorded by Diana Forker in August 2013 in Druzhba, Daghestan, Russia. The autobiographical retelling of the speaker's years as a young man, working first as a guard in the army, then as a miner, and later as a bus driver for a local factory.

bazhuk Speaker SD02. Recorded by Diana Forker in August 2013 in Druzhba, Daghestan, Russia. A traditional narrative in which a young shepherd is abducted by a witch after eating from her apple trees. He manages to hide himself and kill her in her own cooking pot.

dragon Speaker SD02. Recorded by Diana Forker in August 2013 in Druzhba, Daghestan, Russia. A traditional narrative in which a precocious young girl with a healthy appetite devours everyone in her village. Her brother, away on work, refuses to believe the rumours about her, and returns to the village only to be chased up a tree by his ravenous little sister, who has turned into a giant, fire-spewing monster. He calls the village dogs on her, and she is torn to shreds.

kurban Speaker SD03. Recorded by Rasul Mutalov in July 2011 in Druzhba, Daghestan, Russia. An autobiographical narrative in which the speaker recounts the story of him and a friend playing a trick on the speaker's cousin, who desperately desires to become the head of a village – despite being highly unqualified – and will go to great lengths to get the appointment.

mill Speaker SD01. Recorded by Gadzhimurad Gadzhimuradov in August 2013 in Druzhba, Daghestan, Russia. Two traditional narratives, the first of which explains the significance of a particular mountain peak to the village of Sanzhi, the second of which humorously relates the story of the Sanzhi people's early troubles with watermills. Only the second, longer narrative has been annotated with RefIND.

patima Speaker SD02. Recorded by Diana Forker in March 2013 in Druzhba, Daghestan, Russia. A traditional narrative in which a girl, Patima, goes into the forest to gather nuts for her sisters,

9 hdl.handle.net/1839/00-0000-0000-0018-A4D4-6
10 web-corpora.net/SanzhiDargwaCorpus/search/

only to find on her return that they have been eaten by a wolf. With the help of a sympathetic fox, she manages to kill the wolf and rescue her siblings from its gut.

ramazan Speaker SD04. Recorded by Diana Forker and Rasul Mutalov in August 2012 in Druzhba, Daghestan, Russia. The autobiographical recollections of the speaker in his three decades of work as a long-distance lorry driver. He heaps much praise on the Baltic countries, but has less favourable things to say about other places.

tape Speaker SD03. Recorded by Rasul Mutalov in July 2011 in Druzhba, Daghestan, Russia. An autobiographical narrative in which the speaker and a friend visit a shop that sells household goods. There they manage to get into an argument about which of them talks too much, inebriated or sober.

3.9 Teop

Ulrike Mosel, Stefan Schnell

glottocode	teop1238
affiliation	Austronesian, Malayo-Polynesian, Oceanic, Nehan-Bougainville
area spoken	Papua New Guinea, Bougainville
varieties rec'd	Teop island
text types	traditional narratives
sources	Mosel & Thiesen 2007
identifier	teop
availability	since May 2015, version 1505
GRAID	7.0 (≥ 1505)
RefIND	✓ (≥ 1905)
ISNRef	✓ (≥ 1905)
citation	Mosel, Ulrike & Schnell, Stefan. 2015. Multi-CAST Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (multicast.aspra.uni-bamberg.de/#teop)

Teop is an Oceanic language belonging to the Nehan-North Bougainville network of the North-West Solomonic group of the Meso-Melanesian cluster (Ross 1988). Accurate figures for the number of speakers are difficult to ascertain; figures from the last decade range from 5 000 to 10 000. The island of Bougainville was torn by a civil war which lasted from 1989 to 1998 and resulted in an estimated 18 000 to 20 000 casualities, with a devastating effect on the speech population. Factors like marriage outside of the Teop speech community, the pressure of neighbouring languages, the growing influence of Tok Pisin as a lingua franca, and the use of English in education all contribute to making Teop a highly endangered language.

The Teop texts were recorded by Ulrike Mosel and Enoch Horai Magum during the course of a DOBES language documentation project (principal investigator: Ulrike Mosel) funded by the Volkswagen-Stiftung (grant no. II 77 973). A sketch grammar of Teop (Mosel & Thiesen 2007) and additional materials are available on the DOBES website.[11] Another source of information on the meaning and construction of functional and content words is *A multifunctional Teop-English dictionary* (MTED, Mosel 2019).[12]

11 dobes.mpi.nl/projects/teop/
12 dictionaria.clld.org/contributions/teop#twords

The texts were annotated with GRAID by Ulrike Mosel and Stefan Schnell. Referent indexing with RefIND was added in 2019 in a joint effort by Ulrike Mosel, Stefan Schnell, and Maria Vollmer.

3.10 Tondano

Timothy Brickell

glottocode	tond1251
affiliation	Austronesian, Malayo-Polynesian, Philippine, Minahasan, North, Northwest
area spoken	Indonesia, North Sulawesi, Tondano town
varieties rec'd	Toulour dialect
text types	autobiographical narratives, stimulus-based narratives
sources	Sneddon 1975; Brickell 2015
identifier	nkurd
availability	since June 2016, version 1606
GRAID	7.0 (≥ 1606)
RefIND	✘
ISNRef	✘
citation	Brickell, Timothy. 2016. Multi-CAST Tondano. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (multicast.aspra.uni-bamberg.de/#tondano)

The Toulour dialect of Tondano is an Austronesian language spoken in and around the town of Tondano and the lake of the same name, and also in several villages to the east of this area. Tondano is located in the Minahasa regency on the northern tip of the island of Sulawesi, Indonesia. Current speaker numbers are difficult to ascertain, however earlier estimations of 70 000 (Sneddon 1975: 1) and 91 000 (Wurm & Hattori 1981) are now almost certainly incorrect. All Minahasan languages are endangered and have been shifting to the most commonly used language of wider communication, Manado Malay, since the early 20th century (Wolff 2010: 299). Anecdotal evidence and the personal experience of the researcher result in an upper range figure of 30 000 fluent speakers as being considered more accurate.

Tondano is not dominant in any domains of use, and is rarely used in everyday communication such as in workplaces, markets, or in the home. The last domain in which Tondano use remained strong was traditional agricultural work. However, with almost all remaining fluent Tondano speakers now aged 50 years and above, this situation is changing as speakers cease working in the fields. In contemporary society the language has little more than a token role in certain cultural settings such as church services, weddings, or occasionally speech contests in which people read from pre-prepared texts.

The only previous research on this language by a western academic was undertaken in 1975, the result of which was a phonology and sketch grammar (Sneddon 1975) in the framework of Tagmemic grammar theory (as per Longacre 1960; Pike 1964). The sole contemporary linguistic research on is the PhD dissertation of Brickell (2015). The data for this grammatical description come from various recording sessions which took place in North Sulawesi between 2011 and 2013 during three separate fieldtrips. These audio and video recording sessions all occurred at houses in Tondano township or at various locations closer to the lake. All the data were transcribed and translated in situ together with language consultants from within the Tondano speech community. There are approximately seven hours of recordings in total. All recordings are either

monologues or dialogues which were "staged" in the sense of Himmelmann (1998: 185), in that they took place predominantly for the purpose of the collection of primary linguistic data.

The data comprise a number of different recording genres. The first are instances where speakers narrated village and family history, or a specific culturally relevant story or event. The second are procedural narratives where speakers described how to carry out traditional indigenous activities (e.g. cooking, or making handicrafts, or collecting flora and fauna) as they performed them. Finally, some narratives were elicited with the aid of visual stimuli (video recordings) whereby speakers watched and narrated as other community members performed these tasks. A number of dialogic texts were also recorded, but are not included in Multi-CAST.

Despite the staged nature of these communicative events, the recordings in the Tondano corpus are probably as natural as it is possible to be. Moreover, all data were recorded within the culture specific context of the indigenous Tondano speech community. All speakers who were recorded for the corpus gave informed consent for this data to be archived and accessed for further viewing and/or use. The research undertaken by Brickell in North Sulawesi was subject to the *La Trobe University Human Research Ethics guidelines*.[13] These guidelines are required to comply with the 2007 *Australian National Statement on Ethical Conduct in Human Research*.[14]

Background to the recordings

gulamera, kiniar02 These recording were taken in November 2011 and May 2013 at the houses of two speakers (TD01 and TD03) in the Rinegetan and Kiniar suburbs of Tondano town. The speakers narrate while watching an elicitation video which depicts a group of people collecting palm sugar sap from the sugar palm (*arenga pinnata*) tree. The sap is then heated before being poured into coconut shells to be sold as palm sugar when it has cooled.

holiday In this recording the speaker (TD01) describes the experience of travelling to Australia and staying with her granddaughters in Sydney and Brisbane. She describes the things she did and places she saw while there. This narration was recorded in the Rinegetan suburb of Tondano town in September 2011.

kiniar01, kiniar03 In these recordings the speakers (TD02 and TD03) narrate an elicitation video in which people buy fruit bats (commonly *pteropus alecto* or *chironax melancephalus*) from a marketplace. The process of preparing, cooking, and eating bat curry is then described. The recordings took place at two houses in the Kiniar neighbourhood of Tondano town in May 2013.

mapalus The speaker (TD04) in this recording session talks about an aspect of Minahasan culture known as *mapalus,* which is the term for how community members traditionally work together for mutual assistance. She also speaks about her experience during a well known historical event called the Permesta rebellion in which some Minahasans fought against the Brawijaya regiment of the Indonesian National Army. This narration was recorded at a house in the Rinegetan suburb of Tondano town in September 2011.

water This recording session took place in the Rinegetan suburb of Tondano town in August 2011. The speaker (TD05), her mother, and her mother's friend were all recorded on this day. The speaker is narrating an elicitation video which depicts the collecting, cooking, and eating of sago

13 latrobe.edu.au/__data/assets/pdf_file/0008/259217/Human-research-ethics-guidelines-may-2015.pdf.

14 nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018.

grubs (the larvae of the *thynchophorus ferrungineus* beetle) from a sugar palm (*arenga pinnata*) tree.

watulaney This narration was recorded in the lounge room of a house in Tataaran, a suburb just outside of Tondano town in September 2011. The speaker (TD06) is discussing her family history and the history of her village of Watulaney, which is located approximately 30 kilometres to the east of Tondano.

3.11 Tulil

Chenxi Meng

glottocode	`taul1251`
affiliation	Papuan, Taulil-Butam
area spoken	East New Britain, Papua New Guinea
varieties rec'd	Tulil
text types	traditional narratives, autobiographical narratives
sources	Meng 2018
identifier	`tulil`
availability	since July 2019, version `1907`
GRAID	7.0 (≥ `1907`)
RefIND	✓ (≥ `1907`)
ISNRef	✓ (≥ `1907`)
citation	Meng, Chenxi. 2019. Multi-CAST Tulil. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (`multicast.aspra.uni-bamberg.de/#tulil`)

Tulil, also known as Taulil, is a Papuan language spoken in the East New Britain Province of Papua New Guinea. In 2000, Tulil was spoken by approximately 2 000 people spread out over four villages. The Tulil people are referred to by their neighbours as "Taulil", while "Tulil" is the name they call themselves. The Tulil share their villages with the Butam people, whose language (Butam) is to be considered extinct after the last speaker died in 1938 (Laufer 1959). According to the oral history of the Tulil people, they along with the Butam migrated from the island of New Ireland Island to their current home on New Britain at some point in the past, before the arrival of the Tolai people in the area.

The six texts in this corpus comprise a subset of a larger collection of material that was recorded and transcribed by Chenxi Meng during two field trips to East New Britain in 2012 and 2015 for her PhD project, which has resulted in a comprehensive grammar of Tulil (Meng 2018). The entirety of the data has been deposited in PARADISEC;[15] session name correspondences with Multi-CAST are as follows:

- *all1* *AL_L1*
- *alrm* *AL_RM*
- *jkpp* *JK_PP*
- *lnsl* *LN_SL*
- *lrdw* *LR_DW*
- *sves* *SV_ES*

15 `catalog.paradisec.org.au/collections/CM2`

The texts selected for Multi-CAST include both traditional and autobiographical narratives. Annotations with RefIND were added by Maria Vollmer.

3.12 Vera'a

Stefan Schnell

glottocode	vera1241
affiliation	Austronesian, Malayo-Polynesian, Oceanic, Vanuatu
area spoken	Vanuatu, Banks Islands, Vanua Lava
varieties rec'd	Vera'a village
text types	traditional narratives
sources	Schnell 2010, 2011, 2016
identifier	veraa
availability	since May 2015, version 1505
GRAID	7.0 (≥ 1505)
RefIND	✓ (≥ 1905)
ISNRef	✓ (≥ 1905)
citation	Schnell, Stefan. 2015. Multi-CAST Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (multicast.aspra.uni-bamberg.de/#veraa)

The Vera'a language has around 450 speakers, 250 of which live in the village of Vera'a on Vanua Lava, Vanuatu, and the other 200 being scattered along the coast line reaching from Vera'a up to the northern shore. The language was initially researched by Catriona Hyslop in the late 1990s and Alexander Francois in the early 2000s (various publications by Francois deal with Vera'a). Since 2007, the language has been extensively documented by Stefan Schnell, first as part of a documentation project funded within the VolkswagenStiftung's DOBES language documentation programme, and since 2012 as part of Schnell's ASC-funded project on argument realization in discourse of diverse languages.

To date, several hours of video and audio recordings of speech events have been collected by Schnell and other reseachers.[16] A large proportion of the data has been transcribed, handwritten transcriptions being undertaken by native speakers, and later entered into ELAN by Stefan Schnell, together with a translation into English. Some of the recorded narratives have later been edited by a speaker of Vera'a, Makson Vores, and published as a book (Vores et al. 2012).

All speech events recorded took place in the Vera'a community on Vanua Lava. Most of these comprise "staged events" in the sense of Himmelmann (1998), that is they took place mainly for the sake of being recorded as part of the documentation project. Recordings of public events are also included in the Vera'a documentation, and these do not constitute staged events in the strict sense, though speakers were at all times informed about their being recorded. Nevertheless, the Vera'a corpus can be regarded as comprising a large set of fairly natural speech data recorded within the indigenous cultural setting of the speech community. While most recordings were made by Stefan Schnell, Makson Vores collected several narratives in 2012 and 2013.

In addition to narratives and public events, procedural and descriptive texts were recorded. The latter comprise descriptions of plant and fish species. Both of these types of descriptions were recorded in dedicated sessions focussing on ethno-biologial aspects of the Vera'a language.

16 dobes.mpi.nl/projects/vures_veraa/

In both sessions, speakers were asked to describe the respective plant or fish species, with a part of the plant or a picture card of the fish in front of them or in their hands.

The Vera'a subcorpus of Multi-CAST constitutes a relatively small portion of the entire Vera'a corpus. In addition to narratives, some of the plant and fish descriptions have been GRAID'ed and will be added to the corpus in the future, as will edited narratives in an effort to enable research into medium-related variation in argument expression.

Annotations with RefIND were added to the texts in 2019 by Stefan Schnell and Maria Vollmer.

4 Structure

The following sections provide a more general overview of the collection's design philosophy and composition, starting with the supplementary material that documents the annotations (Section 4.1), then moving on to considerations of text types (Section 4.2) and the issue of replicability (Section 4.3).

4.1 Documentation

The following documents are provided in PDF file format for every corpus:

- *annotation notes*
  Descriptions of the pertinent analytical issues that surfaced during the annotation of each corpus, and the annotators' decisions on how to address them. Attached to each corpus' annotation notes are a lists of the abbreviated morphological glosses used, as well as of the additional, corpus-specific GRAID symbols introduced by the annotators.
- *lists of referents*
  Lists of all referents in texts that have been annotated with referent indices with the RefIND scheme (see Sections 2.2 and 3). Alongside the referent indices, the lists of referents contain a label and a short description for each referent, and indicate the ontological class of the referent and its relations to other referents in the text. The lists are also available respectively as tab-separated values files, see Section 5.3. They can further be accessed in R via the *multicastR* package and the `mc_referents` function; refer to Section 6 for more information.
- *corpus counts*
  Tables containing frequency counts of all combinations of selected GRAID form and function symbols in each text. The corpus counts do not provide exact summaries of the annotations; instead, they are intended to provide a cursory overview of the relative proportions of certain types of referring expression in certain grammatical roles. Only a small number of basic GRAID categories are counted; complex GRAID symbols representing more fine-grained distinctions are subsumed under the more general category.
- *translated texts*
  The transcribed object language text side-by-side its English translation, as a parallel text. Utterances are numbered sequentially, corresponding to their respective utterance identifiers. These documents allow unobstructed access to the primary texts, separate from the morphological glosses and annotations.

Lastly, metadata on the texts and speakers in the Multi-CAST collection can be found in these places and formats:

- *Appendix*
  in Appendices A and B of this document;
- *TSV metadata*
  collected into a tab-separated values (TSV) table, see Section 5.2.1, and in R via the `mc_metadata` function from *multicastR*; as well as
- *IMDI metadata*
  in the standardized IMDI for media metadata from the Multi-CAST webpage, see Section 5.2.2.

4.2 Genres and text types

All texts in Multi-CAST are spoken narratives which are overwhelmingly monologic, non-elicited, unrehearsed, and "original" in the sense that they are not translated (see Haig et al. 2011). Each text belongs to one of three broadly defined narrative text types:

Traditional narratives
: Traditional stories, folktales, and fairy tales, usually told to an audience of native speakers.

Autobiographical narratives
: Narrative accounts of the speaker's personal history or other past events and memories, usually recorded in private settings. In many cases, these texts were originally collected as part of oral history projects.

Stimulus-based narratives
: Retellings of short video recordings, for instance the *Pear film*, a six-minute film without dialogue about children stealing fruit (cf. Chafe 1980), or various clips depicting scenes from relevant cultural contexts.

Note that in typical language documentation settings, most narrative events are in fact what Himmelmann (1998) calls "staged communicative events" rather than truly incidental occurrences, and would not have taken place if not by request of an observer.

4.3 Versioning

The Multi-CAST collection continues to develop as new material is added and the annotations of older texts are revised. In order to ensure stable points of reference for the reproduction of published results, successive releases of the corpus data are assigned version numbers composed of the year and month they were published (e.g. `2001` for the January 2020 release). A comprehensive list of additions and changes to the collection introduced with each version can be found in Appendix C.

To ensure the replicability of research results, analysts are advised to note which version(s) of the Multi-CAST data has been used in the analysis, be that in the running text, in a footnote, or as part of the citation for the collection. To date, the Multi-CAST has received the following releases:

2001 the latest version, published in January 2020, adding Mandarin (Vollmer 2020);

1908 published in August 2019, adding Arta (Kimoto 2019) and Nafsan (Thieberger & Brickell 2019);

1907 published in July 2019, adding Tulil (Meng 2019);

1905 published in May 2019, adding Sanzhi Dargwa (Forker & Schiborr 2019) and the first annotations with RefIND and ISNRef (see Section 2.2);

1606 published in June 2016, adding Persian (Adibifar 2016) and Tondano (Brickell 2016);

1505 the original version of Multi-CAST, published in May 2015, beginning with Cypriot Greek (Hadjidas & Vollmer 2015), English (Schiborr 2015), Northern Kurdish (Haig et al. 2019), Teop (Mosel & Schnell 2015), and Vera'a (Schnell 2015).

4.4 Licensing

In the spirit of open science, all material in the Multi-CAST collection, including the recordings, transcriptions, annotations, and documentation, is published under the *Creative Commons Attribution 4.0 International Licence* (CC BY 4.0).[17,18] The licence allows full and unrestricted access to Multi-CAST for any purpose related to research, art, journalism, or any other endeavour, under the condition that proper credit is given to the editors of the collection and its contributors. This must include a short note about the licensing terms and a link to the Multi-CAST website (`multicast.aspra.uni-bamberg.de/`).

Prior to May 2019, the Multi-CAST licence included the *ShareAlike* (SA) and *NonCommerical* (NC) conditions, which we have since decided to remove from the terms.

4.5 Contributors

The editorial team behind the Multi-CAST project consists of Geoffrey Haig, Stefan Schnell, Nils Schiborr, and Maria Vollmer, all at the Department of General Linguistics, University of Bamberg, Germany. In addition, the following researchers were involved in the collection, transcription, translation, and annotation of the various Multi-CAST corpora:

Shirin Adibifar	Yukinori Kimoto	Rasul Mutalov
Timothy Brickell	Adrian Kuqi	Nicholas Peterson
Diana Forker	Jenny Herzky	Liu Ruoyu
Gadzhimurad Gadzhimuradov	Enoch Horai Magum	Nick Thieberger
Harris Hadjidas	Chenxi Meng	Hanna Thiele
Abdullah Incekan	Ulrike Mosel	Makson Vores

We are all indebted to our respective research communities for their support and stimulating criticism.

4.6 Acknowledgements

The collection and annotation of the data in Multi-CAST have graciously received support from the following institutions and organizations:

2017–2020 the German Research Foundation (DFG) via the project *Does morphosyntactic alignment shape discourse?* — principal investigators: Geoffrey Haig and Stefan Schnell (DFG project no. 323627599);[19]

2018–2020 the Centre of Excellence for the Dynamics of Language (CoEDL) as part of CoEDL's corpus development project, funded by the Australian Research Council (ARC) and headed by Nick Thieberger at The University of Melbourne, for annotation work in collaboration with the aforementioned DFG project;

17 The text of the licence can be found online at `creativecommons.org/licenses/by/4.0`.
18 The sole exception are the audio recordings in the English corpus, which belong to the public domain.
19 `gepris.dfg.de/gepris/projekt/323627599`

2012–2019 the VolkswagenStiftung as part of the *Documentation of endangered languages* (DOBES) project for the documentation of Shiri and Sanzhi — PI: Diana Forker;[20]

2012–2015 the Australian Research Council (ARC) as part of the DECRA project *Typology of language use*, hosted by La Trobe University, Melbourne — PI: Stefan Schnell (ARC grant no. DE120102017);

2006–2012 as part of the DOBES project for the documentation of Vera'a and Vurës — Stefan Schnell (PI: Catriona Malau);[21]

2000–2007 as part of the DOBES project for the documentation of Teop — PI: Ulrike Mosel (grant no. II 77 973).[22]

In addition, the Department of General Linguistics at the University of Bamberg contributed departmental funding and research infrastructure to the Multi-CAST project.

A number of texts in the collection are made available in cooperation with the following researchers and institutions:

- the texts in the Cypriot Greek corpus were taken from a book of traditional Cypriot tales (Giangoullis 2009), which were kindly made available for inclusion in Multi-CAST by the author of the book, Konstantinos Giangoullis;
- the English corpus consists of a subset of the *Freiburg English Dialect Corpus* (FRED, English Dialects Research Group 2005), compiled under the supervision of Bernd Kortmann and Lieselotte Anderwald at the University of Freiburg, Germany.

4.7 Guidelines for contributors

The shared utility of Multi-CAST grows with increasing typological representativity of the language sample it contains. We therefore encourage scholars to contribute additional data sets to Multi-CAST, which can be incorporated into the collection as stand-alone resources, citable with their names as the authors and annotators.

If you wish to contribute data, here are some points to consider:

Open access corpus data
: Your data should be free of copyright and other restrictions on availability or usage. Multi-CAST is committed to open science, and hence makes all of its data freely available under a *Creative Commons* licence (CC BY 4.0, see Section 4.4). All data sets are citable online resources, with your name(s) as author(s).

Monologues
: Texts should be (predominantly) monologic. Coping with multi-person discourse raises additional issues of annotation and analysis, which we have chosen not to tackle in this collection.

Media-linked time-aligned annotations
: Transcribed texts are ideally accompanied by a sound file in an uncompressed WAV file format, morphologically glossed, and translated into English. The annotations are time-aligned with the audio recordings.

Minimum size of 1 000 clauses
: All corpora in Multi-CAST minimally contain 1 000 clause units.

20 dobes.mpi.nl/projects/shiri_sanzhi/
21 dobes.mpi.nl/projects/vures_veraa/
22 dobes.mpi.nl/projects/teop/

If you have a data set that complies with these conditions and you are interested in contributing it to Multi-CAST, please contact Geoffrey Haig and Stefan Schnell in order to coordinate the next steps.

Technically speaking, this involves transferring your data into the EAF file format of the annotation software ELAN, for which purpose we will provide you with a Multi-CAST ELAN template (see also Section 5.1.1), and annotating your texts with GRAID (Section 2.1). The latter involves some quite tricky analytical decisions, and we strongly recommend that potential contributors liaise with us before undertaking this task. The actual labour input required will vary from language to language, but we will certainly assist you and be able to give you a realistic assessment of what may be necessary.

5 Data formats

The data in Multi-CAST can be accessed in a variety of file formats, each suited to different workflows. This section outlines the contents and internal structure of the files and file formats included in Multi-CAST, specifically of the annotated texts (Section 5.1) and the metadata (Section 5.2).

5.1 Annotation formats

The core component of the Multi-CAST corpora are natural language texts which have been recorded, where possible, in their respective cultural contexts, and subsequently transcribed, translated into English, and annotated across multiple levels. The texts themselves are provided as WAV audio files, while the annotation values – including the object language texts, the translations, the morphological glosses, and various levels of annotations – are available in the following formats:

- as EAF files, a file format used with the free annotation software ELAN, see Section 5.1.1;
- as human-readable XML files, which can be queried with XPath and transformed into other formats via XSLT, see Section 5.1.2, and
- as flat tab-separated values (TSV) files, see Section 5.1.3.

The three formats differ in structure, but are equivalent in terms of their content, with the exception of the TSV files, which lack the unsegmented object language texts and translations. The EAF files are directly linked to the audio files and allow parallel playback of individual segments in ELAN; the XML files carry the same timestamps.

Because some of the recording sessions in Multi-CAST are quite long – the longest spanning multiple hours – they had to be split into multiple EAF files to accommodate technical limitations with ELAN. The XML and TSV files follow suit for the sake of consistency. All parts of a text should be considered in unison during analysis, as they constitute a single, connected discourse. Each part of a split text is marked with a lowercase letter (*a*–*z*) in its file name: the two parts of the *kent02* text from the English corpus are labelled *mc_english_kent02_a* and *mc_english_kent02_b*, for instance. This label is applied only to the file names, however, and not to the utterance identifiers (see below), which instead are numbered continuously across all parts of a text. Supplementary materials (specifically, the corpus counts and lists of referents) likewise do not make reference to parts. Texts consisting of a single part do not carry the additional label.

5.1.1 EAF

EAF is the file format used by ELAN, an annotation tool developed at the Max Planck Institute for Psycholinguistics in Nijmegen.[23] ELAN is free and open software released under the GNU General Public Licence (v3). As it is written in Java, it can be used on just about any platform.

Annotations in ELAN are organized across multiple levels; ELAN uses the term "tier" for a distinct level in the interlinear structure of an annotated text. Tiers directly reference the audio recording or stand in a hierarchical relationship with other tiers. The EAF files in Multi-CAST have one time-aligned tier (containing the utterance identifiers, `utterance_id`), which splits the recording into segments of various length. All other tiers are descendants of this root tier.

The relationship (or "tier type", as ELAN calls it) between cells on child and parent tiers is defined in terms of its cardinality: in the Multi-CAST EAF files, a cell on a parent tier either has (A) exactly one child (i.e. a one-to-one relation, labelled "symbolic association"), or (B) one or more children (i.e. a one-to-many relation, "symbolic subdivision"). The sole instance of a one-to-many relation is that between the `utterance` tier, containing the object language text, and its child, the `grammatical_words` tier, as each utterance is split into multiple word tokens. All other tiers have a one-to-one relation with their respective parent, extending it in a meaningful way: each object language word on the `grammatical_words` tier, for instance, is linked to a corresponding cluster of morphological glosses on the `gloss` tier.

The resulting hierarchy of tiers is comparatively deep, which facilitates the connection of information on one tier with information on others, as the association of a cell with all of its ancestors and descendants is clearly derivable in all cases. For a practical demonstration of working with Multi-CAST using complex multi-level queries, refer to the *Multi-CAST case studies*, available from the Multi-CAST webpage, and for a general reference to designing complex queries with regular expressions in ELAN, see Mosel (2015).

The Multi-CAST EAF files contain six core tiers (eight with RefIND and ISNRef, see Sections 2.2 and 2.3). The tiers are described in the following, and the diagram in Figure 2 illustrates their relationship in graphical form.

- `utterance_id` — *segmentation unit identifier*
  A unique identifier applied to each object language segment. The identifier is composed of the internal name of the corpus, the name of the text, and a four-digit number, e.g. *teop_iar_0210*. Time aligned with the recording (*time alignment*); this is the root tier to which all other tiers are subordinate.
- `utterance` — *object language text*
  A segment of the object language text. Child of the `utterance_id` in a one-to-one relation (*symbolic association*).
- `grammatical_words` — *grammatical words*
  The object language text divided into "word" tokens. The segmentation into grammatical words forms the basis for all annotation tiers. Grammatical words in this context should be understood in terms of GRAID annotation units (see Section 2.1): as such, cells for abstract elements such as zero anaphors and clause boundaries are inserted on this tier, and certain types of bound forms (e.g. pronominal clitics) may be split off into separate cells where relevant to the annotations. Child of the `utterance` tier in a one-to-many relation (*symbolic subdivision*).

23 Download and manual available at `tla.mpi.nl/tools/tla-tools/elan/`

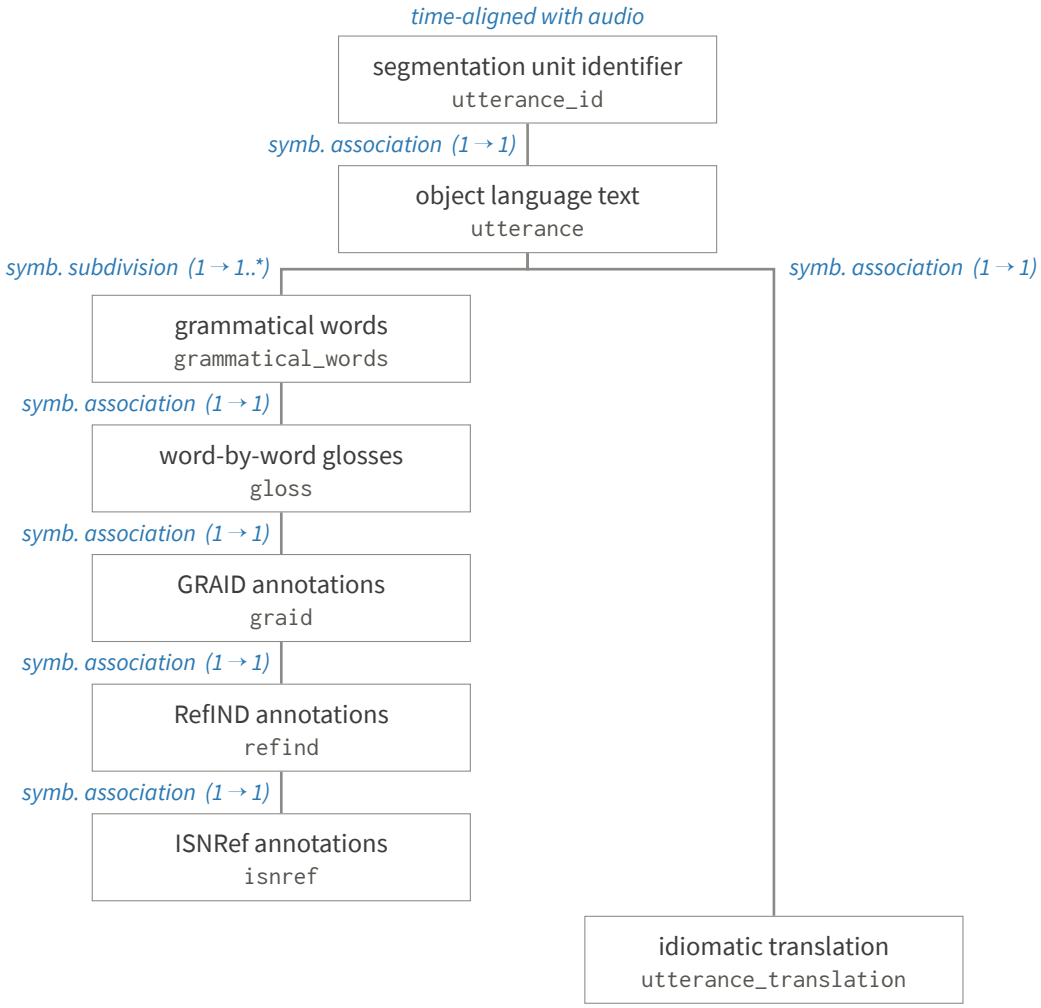

Figure 2 The hierarchical structure and cardinal relations of the tiers in the EAF files. Texts that have yet to receive annotations with the RefIND and ISNRef schemes lack the associated tiers. Optional tiers for comments or additional orthographies are not shown.

- `gloss` — *word-by-glosses*
  Morphological glossing with conventionalized labels as per the *Leipzig Glossing Rules*. The extent to which individual forms have been morphologically segmented and the level of detail of the glossing are at the discretion of the annotator, and so may vary between corpora. A list of the abbreviated morphological labels used is included in each corpus' *annotation notes*. Child of the `grammatical_words` tier in a one-to-one relation (*symbolic association*).
- `graid` — *GRAID annotations*
  Syntactic annotations with the GRAID scheme (*Grammatical relations and animacy in discourse,* Haig & Schnell 2014). GRAID combines information on the form of referential expressions, in particular of major clause constituents, with information on their grammatical relations and semantics. As mentioned above, GRAID also notes referential elements that are not overtly expressed (zero) and includes markers for left and right clause boundaries. A brief outline of the GRAID scheme can be found in Sec-

tion 2.1; an in-depth description is provided in the *GRAID Manual* Haig & Schnell (2014), available from the Multi-CAST website. Child of the *gloss* tier in a one-to-one relation (*symbolic association*).

- `refind` — *RefIND annotations*
  Referent identification with the RefIND scheme (*Referent indexing in natural-language discourse,* Schiborr et al. 2018). RefIND assigns a unique numerical identifier to every occurrence of a specific referent in a text, allowing it to be tracked over the course of a discourse. Refer to Section 2.2 for a summary of the principles of RefIND, and to the *RefIND annotation guidelines* (Schiborr et al. 2018) for a comprehensive description. As RefIND can be considered an extension of the GRAID annotations, this tier is a child of the *graid* tier in a one-to-one relation (*symbolic association*). Note that not all of the corpora in Multi-CAST have been outfitted with referent indices; the EAF files of texts that lack annotations with RefIND do not have a `refind` tier.
- `isnref` — *ISNRef annotations*
  ISNRef (*Information status of new referents,* Schiborr et al. 2018: 15) is a simplified version of Riester & Baumann's (2017) RefLex annotation scheme; see Section 2.3. ISNRef captures information on the information status of referents at the point of their formal introduction into the discourse. The simplified scheme distinguishes between references to discourse-new referents, bridging anaphora, and references to unused (i.e. known but unevoked) entities. The ISNRef annotations extend RefIND annotations, so only texts annotated with the latter have annotations with the former. As such, this tier is a child of the `refind` tier in a one-to-one relation (*symbolic association*).
- `utterance_translation` — *idiomatic translation*
  An idiomatic English translation of the object language utterance. This tier is a child of the `utterance` tier in a one-to-one relation (*symbolic association*).

Beside these core tiers, EAF file may include further optional tiers, such as:

- `add_comments` — *annotators' comments*
  Optional tier. Comments on the glossing or annotations, the cultural context of the text, the recording situation, and so on. If present, a child of the *utterance* tier in a one-to-one relation (*symbolic association*).
- `add_orthography` — *additional orthographies*
  Optional tier. The object language text in another orthographical system; in Mandarin or Japanese, for instance, this tier contains the text in its original orthography (hanzi, or kanji and kana) while the `utterance` tier is a transliteration of the text (pinyin, or romaji). If present, a child of the *utterance* tier in a one-to-one relation (*symbolic association*).

Annotators may choose to include optional tiers beyond these two, for example for the purposes of specific research questions. In all cases, the names of extra tiers are prefixed with the label "`add_`". Extra tiers may be children of any of the core tiers, but may not interfere with the basic structure in any way.

Lastly, the EAF files encode limited metadata, specifically the title of the text, the identity of the speaker, and the names of the annotators. For more comprehensive metadata, refer to Appendices A and B.


<multicast version>
<corpus c_name> (1→1..*)
<text t_name annotators speaker> (1→1..*)
<file f_name audio updated> (1→1..*)
<unit uid start_time end_time> (1→1..*)
<utterance_id> (1→1)
<utterance> (1→1)
<add_orthography> (1→0..1)
<utterance_translation> (1→1)
<add_comments> (1→0..1)
<annotations> (1→1)
 (1→1..*)
<gword> (1→1)
<gloss> (1→1)
<graid> (1→1)
<refind> (1→0..1)
<isnref> (1→0..1)

Figure 3 The structure of the Multi-CAST XML files. Attributes are given in blue. Cardinalities: `(0..1)` 'zero or one', `(1)` 'exactly one', `(1..*)` 'one or more'.

5.1.2 XML

While the EAF file format used by the annotation software ELAN itself derives from XML, it does not take advantage of the strengths of XML: clear hierarchization and human readability. The XML files included in Multi-CAST are generated automatically from the EAF files, and as such contain the exact same data, but restructure them in a more sensible manner. This allows them to be more easily queried with XPath (the XML Path Language) and transformed into other XML structures via XSL (the Extensible Stylesheet Language). The `<text>` node contains in its attributes the same limited metadata as the EAF files; the root `<multicast>` node additionally encodes the four-digit corpus version number in its `version` attribute.

The structure of the Multi-CAST XML files is illustrated in Figure 3. The following is an outline of each node, its attributes, and its relation to other nodes.

- `<multicast>`
  The root node of the XML tree, encompassing the entire Multi-CAST collection. Has one attribute:
  - `version`
    The four-digit version number of the collection, see Section 4.3.
- `<corpus>`
  A corpus comprising multiple texts. Child of the `<multicast>` node in a one-to-many relation (1 → 1..*). Has one attribute:

    - `c_name`
      The internal name of the corpus, e.g. *sanzhi* for the Sanzhi Dargwa corpus.

- `<text>`
  A text, which contains one or more files. Child of a `<corpus>` node in a one-to-many relation (1 → 1..*). Has three attributes:
    - `t_name`
      The name of the text.
    - `annotators`
      The name(s) of the annotator(s), separated by commas.
    - `speaker`
      The speaker identifier, composed of a two-letter corpus label and a two-digit number.
- `<file>`
  A file belonging to a text. Most texts only have one file; some longer texts had to be split into multiple parts, and hence contain multiple files. Child of a `<text>` node in a one-to-many relation (1 → 1..*). Has three attributes:
    - `f_name`
      The full name of the file sans file extension, e.g. *mc_veraa_pala_a*.
    - `audio`
      The name of the corresponding WAV audio file. If no audio file is available, has a value of *NA*.
    - `updated`
      The date of the last change to the text, in YYYY-MM-DD format.
- `<unit>`
  A time-aligned utterance unit. A child of a `<file>` node in a one-to-many relation (1 → 1..*). Has three attributes:
    - `uid`
      The numerical part of the utterance identifier; see below. Unique within each text.
    - `start_time`
      The start time of the utterance unit, in milliseconds since the start of recording.
    - `end_time`
      The end time of the utterance unit, in milliseconds since the start of the recording.
- `<utterance_id>`
  A label uniquely identifying an utterance within the collection. Composed of the name of the corpus, the names of the text, and a four-digit number. Child of a `<unit>` node in a one-to-one relation (1 → 1).
- `<utterance>`
  A segment of the object language text. A child of a `<unit>` node in a one-to-one relation (1 → 1).
- `<add_orthography>`
  The object language text, optionally represented with a different orthographic system. Child of a `<unit>` node in a one-to-zero-or-one relation (1 → 0..1).

- <utterance_translation>
  An idiomatic English translation of the object language utterance. Child of a <unit> node in a one-to-one relation (1 → 1).
- <add_comment>
  An optional comment on the utterance or parts therein. Child of a <unit> node in a one-to-zero-or-one relation (1 → 0..1).
- <annotations>
  A wrapper element containing the word tokens, glosses, and annotation values for the utterance. Child of a <unit> node in a one-to-one relation (1 → 1).
-
  A segment within the utterance containing one "word" unit and its annotations. Its children may also contain abstractions such as zero anaphora and clause boundary markers. Child of a node in a one-to-many relation (1 → 1..*).
- <gword>
  A grammatical word in the object language. "Word" here should be understood in terms of a GRAID annotation unit. Child of a node in a one-to-one relation (1 → 1).
- <gloss>
  The morphological glossing for the grammatical word, as per the *Leipzig Glossing Rules*. Child of a node in a one-to-one relation (1 → 1).
- <graid>
  An annotation unit using the GRAID scheme (*Grammatical relations and animacy in discourse,* Haig & Schnell 2014); see Section 2.1. Child of a node in a one-to-one relation (1 → 1).
- <refind>
  A referent index, unique within each text, as defined by the RefIND scheme (*Referent indexing in natural-language discourse,* Schiborr et al. 2018); see Section 2.2. Child of a node in a one-to-zero-or-one relation (1 → 0..1).
- <isnref>
  The information status of the referent at the point of its introduction into discourse. Based on a simplified version of Riester & Baumann's (2017) RefLex annotation scheme; see Section 2.3. Only present alongside a <refind> node. Child of a node in a one-to-zero-or-one relation (1 → 0..1).

5.1.3 TSV

Like the XML files, the tab-separated values (TSV) files are generated automatically from the EAF, and so contain the same data albeit with a different structure, with the sole exception that the TSV files lack the unsegmented object language utterance, the English translation, and any extra information such as metadata, timestamps, and comments. The reason for their exclusion is that these data are of comparatively little use for the types of quantitative analyses for which the TSV format is intended; the metadata is available as a separate TSV table, and so can be easily joined to the annotation values when needed (see Section 5.2.1 below).

The TSV tables contain the following eight columns:

- corpus
  The internal names of the corpora, e.g. *nkurd* for the Northern Kurdish corpus.

- `text`
  The names of the texts.
- `uid`
  The numerical parts of the utterance identifiers, e.g. *0240*. Unique within each text.
- `gword`
  The grammatical words in the object language.
- `gloss`
  Morphological glosses for the grammatical words, as per the *Leipzig Glossing Rules*.
- `graid`
  Annotations with the GRAID scheme (*Grammatical relations and animacy in discourse,* Haig & Schnell 2014), see Section 2.1. For ease of analysis, the GRAID words have been pre-emptively split into their components parts and placed in the `gform`, `ganim`, and `gfunc` columns.
- `gform`
  The GRAID symbols for the form of an expression (e.g. ⟨`np`⟩, ⟨`pro`⟩), as well as other symbols (e.g. ⟨`adp`⟩, ⟨`ln`⟩) and clause boundary markers.
- `ganim`
  The GRAID symbols for the person and animacy of a referring expression (e.g. ⟨`1`⟩, ⟨`2`⟩, ⟨`h`⟩).
- `gfunc`
  The GRAID symbols for the function of an expression (e.g. ⟨`a`⟩, ⟨`pred`⟩, ⟨`poss`⟩).
- `refind`
  Referent indexing with the RefIND scheme (*Referent indexing in natural-language discourse,* Schiborr et al. 2018), see Section 2.2. Unlike in the EAF and XML files, the `refind` (and `isnref`) columns are present even if the text in question has not been annotated with RefIND and ISNRef. In such cases the two columns are simply left empty.
- `isnref`
  Annotations on the information status of referents at the point of their formal introduction into the discourse, using a simplified version of Riester & Baumann's (2017) RefLex annotation scheme, see Section 2.3.

The annotation values in this TSV format can be easily accessed in R via the `multicast` function from the *multicastR* package, as described in Section 6. The metadata can be accessed via the `mc_metadata` function.

5.2 Metadata formats

5.2.1 TSV

The metadata on the texts and speakers is listed in this document in Appendices A and B. For the purpose of combining the metadata with the annotation values during analysis, the same information is also available as a TSV table with the following eight columns:

- `corpus`
  The internal name of the corpus, e.g. *cypgreek* for the Cypriot Greek corpus.
- `text`
  The name of the text.

- `type`
  The text type of the text, one of *TN* 'traditional narrative', *AN* 'autobiographical narrative', or *SN* 'stimulus-based narrative'. See Section 4.2.
- `recorded`
  The year (YYYY) the text was recorded.
- `speaker`
  The identifier for the speaker, unique within Multi-CAST, composed of a two-letter corpus label and a two-digit number, e.g. *EN01* for the first speaker from the English corpus.
- `gender`
  The gender of the speaker.
- `age`
  The age of the speaker at the time of recording. Approximate values are prefixed with a *c*, e.g. *c50*.
- `born`
  The speaker's year of birth. Approximate values are prefixed with a *c*, e.g. *c1970*.

This table can be accessed directly in R via the *multicastR* package and its `mc_metadata` function. See Section 6 for details.

5.2.2 IMDI

A revision of older, existing IMDI metadata is currently underway. They will be described here and made available from the Multi-CAST website as soon as the revision is completed.

5.3 Lists of referents as TSV

The *lists of referents* containing additional information on the referents charted by the RefIND scheme (Schiborr et al. 2018, Section 2.2) are available both in PDF (see Section 4.1) and TSV file formats. The latter of has the following eight columns:

- `corpus`
  The internal name of the corpora.
- `text`
  The name of the text.
- `refind`
  The four-digit referent index, unique to each referent in a text.
- `label`
  The label used for the referent.
- `description`
  A short description of the referent.
- `class`
  The semantic class of the referent; one of *hum* 'human', *anm* 'non-human animate', *inm* 'inanimate', *bdp* 'body part', *mss* 'mass', *loc* 'location', *tme* 'time', or *abs* 'abstract'. Only a single label is assigned to a referent, even where a group contains entities belonging to multiple classes. In such cases humans outweigh other animates, animates outweigh inanimates, and inanimates outweigh everything else in no particular order.

- `relations`
  The relations of the referent to other referents; including < 'set member of (partial co-reference)', < 'includes (split antecedence)', and *M* 'part-whole'; referents with the same relation are delimited by commata, and different types of relations by semicola, e.g. *> 0001, 0002; M 0003*.
- `notes`
  Annotators' notes on the referent and its properties.

The `mc_referents` function from the *multicastR* package allows the lists of referents for all corpora and texts to accessed directly in R. See Section 6 for details.

6 The *multicastR* package

multicastR (Schiborr 2018) is a companion package to the Multi-CAST collection for the statistical programming language and environment R.[24] R is free and open software released under the GNU General Public Licence (v2), runs on a multitude of platforms, and is nearly infinitely extensible through the use of packages.

The *multicastR* package offers a quick and convenient way of accessing the Multi-CAST annotation data and metadata in R. The functions it provides download the data directly from the servers of the University of Bamberg and output them as R tables. The installation and use of the package is described in the following sections. Gracious thanks go to Jenny Herzky, Nick Peterson, and Maria Vollmer for helping with testing *multicastR*.

6.1 Installation and use

The `multicastR` package can be freely installed from CRAN (the Comprehensive R Archive Network)[25] via the command

```
# install multicastR
install.packages("multicastR")
```

The files (tarballs) for a manual installation can be also found on the Multi-CAST website.[26] The `multicastR` package depends on a number of other packages (`data.table`, `curl`, `stringi`) which will be downloaded automatically when installing from CRAN.

Once installed, the `multicastR` package can be attached to the R workspace via

```
# load multicastR
library(multicastR)
```

just like any other package, and subsequently used. In lieu of the following sections, the package documentation can be accessed in R with `?multicastR`, `?multicast`, `?mc_index`, and so on.

6.2 List of functions

24 cran.r-project.org/
25 cran.r-project.org/package=multicastR
26 multicast.aspra.uni-bamberg.de/data/mcr/pkg/

6.2.1 multicast

The centrepiece `multicast` function downloads a table with the Multi-CAST annotation values from the servers of the University of Bamberg, and presents it as a `data.table`; it therefore requires an active internet connection to work. The table accessed by this function has the same shape and contains the same information as the TSV tables described in Section 5.1.3 above.

```
# access the annotation values
multicast()
```

The `multicast` function has an *optional* argument, `vkey`, which takes a four-digit version number (either as an integer or a string, e.g. `1908` or `"1908"`) to download a specific version of the Multi-CAST data. This allows past published analyses to be reproduced easily with the help of *multicastR*. A list of the accepted version numbers can be accessed in R via the `mc_index` function, as described below, as well as found in Section 4.3 and Appendix C. If no version number is supplied to the `multicast` function, it defaults to downloading the most recent version of the corpus data.

```
# access a specific version of the data
multicast(1908)
```

6.2.2 mc_index

`mc_index` downloads a table with version information for the tables accessed by the `multicast` function, and presents it as a `data.table`. This function serves as a signpost for reproducing published research results based on Multi-CAST data. The table has four columns:

1. `version` lists the four-digit version numbers,
2. `date` the publication date in YYYY-MM-DD format,
3. `size` the file sizes of the tables with annotation values, and
4. `corpora` is a comma-separated list of the corpora included in each version.

```
# access the version index
mc_index()
```

6.2.3 mc_metadata

`mc_metadata` downloads a table with the metadata for all Multi-CAST corpora and presents it as a `data.table`. The table is organized by text, and contains the columns and information listed in Section 5.2.1 above. Like `multicast`, this function may take an optional argument `vkey` for selecting specific versions of the metadata.

```
# access the metadata
mc_metadata()
```

```
# access a specific version of the metadata
mc_metadata(1908)
```

The metadata table can be joined to a table with annotation values (e.g. from the `multicast` function) via

```
# join the metadata to the annotation values
merge(mc,
```

```
      mc_metadata(),
      by = c("corpus", "text"))
```

where `mc` is a table with annotation values.

6.2.4 `mc_referents`

`mc_referents` downloads a table with the lists of referents (see Section 4.1) for all Multi-CAST with RefIND annotations (Section 2.2) and presents it as a `data.table`. The table contains the columns and information listed in Section 5.3 above. Like `multicast`, this function may take an optional argument `vkey` for selecting specific versions of the list of referents.

```
# access the lists of referents
mc_referents()

# access a specific version of the lists of referents
mc_referents(1908)
```

The list of referents can be joined to a table with annotation values (e.g. from the `multicast` function) via

```
# join the list of referents to the annotation values
merge(mc,
      mc_referents(),
      by = c("corpus", "text", "refind"),
      all.x = TRUE)
```

where `mc` is a table with annotation values.

6.2.5 `mc_clauses`

`mc_clauses` generates a table with

1. `nClause` the number of valid clause units (i.e. excluding ⟨#nc⟩),
2. `nAll` the total number of clause units, valid or otherwise,
3. `nNC` the number of segments not considered (i.e. ⟨#nc⟩), and
4. `pNC` the percentage of ⟨#nc⟩ segments of the total

in each corpus, or, if the argument `bytext` is set to `TRUE`, in each text. To work, the function requires a table with annotation values, such as the ones accessed by the `multicast` function, containing minimally the `corpus`, `text`, and `graid` columns.

```
# count the number of clauses in each corpus
mc_clauses(multicast())

# count the number of clauses in each text
mc_clauses(multicast(),
           bytext = TRUE)
```

The table is printed to the console by default, but can also be assigned to an object, for example via `clauses <- mc_clauses(mc)`.

Bibliography

The Multi-CAST collection

Haig, Geoffrey & Schnell, Stefan (eds.). 2015. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/) (Accessed 2019-03-08).

Adibifar, Shirin. 2016. Multi-CAST Persian. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#persian) (Accessed 2019-03-08).

Brickell, Timothy C. 2016. Multi-CAST Tondano. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#tondano) (Accessed 2019-03-08).

Forker, Diana & Schiborr, Nils N. 2019. Multi-CAST Sanzhi Dargwa. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#sanzhi) (Accessed 2019-04-05).

Hadjidas, Harris & Vollmer, Maria C. 2015. Multi-CAST Cypriot Greek. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#cypgreek) (Accessed 2019-03-08).

Haig, Geoffrey & Vollmer, Maria C. & Thiele, Hanna. 2019. Multi-CAST Northern Kurdish. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#nkurd) (Accessed 2019-07-05).

Kimoto, Yukinori. 2019. Multi-CAST Arta. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#arta) (Accessed 2019-07-05).

Meng, Chenxi. 2019. Multi-CAST Tulil. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#tulil) (Accessed 2019-07-05).

Mosel, Ulrike & Schnell, Stefan. 2015. Multi-CAST Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#teop) (Accessed 2019-03-08).

Schiborr, Nils N. 2015. Multi-CAST English. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#english) (Accessed 2016-02-28).

Schnell, Stefan. 2015. Multi-CAST Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#veraa) (Accessed 2019-03-08).

Thieberger, Nick & Brickell, Timothy. 2019. Multi-CAST Nafsan. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#nafsan) (Accessed 2019-07-31).

Vollmer, Maria. 2020. Multi-CAST Mandarin. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#mandarin) (Accessed 2020-01-03).

Haig, Geoffrey & Schnell, Stefan. 2016a. Multi-CAST research context. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/) (Accessed 2019-03-08).

Schiborr, Nils N. 2018. multicastR: A companion to the Multi-CAST collection. R package version 1.3.0. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://cran.r-project.org/package=multicastR) (Accessed 2019-08-30).

References

Andrews, Avery. 2007. The major functions of the noun phrase. In Shopen, Timothy (ed.), *Language typology and syntactic description*, vol. 1: Clause structure, 132–223. Cambridge: Cambridge University Press.

Ariel, Mira. 1988. Referring and accessibility. *Journal of Linguistics* 24(1). 65–87.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

Ariel, Mira. 2004. Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse Processes* 37(2). 91–116.

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736.

Brickell, Timothy C. 2015. *A grammar of Tondano*. Melbourne: La Trobe University Ph.D. dissertation.

Chafe, Wallace (ed.). 1980. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

Dowle, Matt & Srinivasan, Arun. 2019. *data.table: Extension of 'data.frame'. R package version 1.12.2.* (`http://CRAN.R-project.org/package=data.table`) (Accessed 2019-08-23).

Du Bois, John. 1987. Absolutive zero: Paradigm adaptivity in Sacapultec Maya. *Lingua* 71(2). 203–222.

Du Bois, John. 2003. Argument structure: Grammar in use. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure: Grammar as architecture for function*, 11–60. Amsterdam: John Benjamins.

Du Bois, John. 2017. Ergativity in discourse and grammar. In Coon, Jessica & Massam, Diane & Travis, Lisa D. (eds.), *The Oxford handbook of ergativity*, 23–57. Oxford: Oxford University Press.

English Dialects Research Group. 2005. *Freiburg English Dialect Corpus (FRED)*. (`http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/`) (Accessed 2016-02-03).

Forker, Diana. Under revision. *A grammar of Sanzhi Dargwa*. Berlin: Language Science Press.

Giangoullis, Konstantinos G. 2009. *Kypriaka paradosiaka paramytha: Ek stomatos Elenis Mich, Satsia, Apo to Geri-Pyroi (1887–1982)* (Viviothiki Kyprion Laikon Poiiton 71). Leukosia: Theopress Publications.

Haig, Geoffrey. 2018. Northern Kurdish (Kurmanjî). In Haig, Geoffrey & Khan, Geoffrey (eds.), *The languages and linguistics of Western Asia: An areal perspective*, 106–158. Berlin: Mouton de Gruyter.

Haig, Geoffrey & Nau, Nicole & Schnell, Stefan & Wegener, Claudia (eds.). 2011. *Documenting endangered languages: Achievements and perspectives*. Berlin: Mouton de Gruyter.

Haig, Geoffrey & Öpengin, Ergin. 2018. Kurmanji in Turkey: Structure, varieties, and status. In Bulut, Christiane (ed.), *Linguistic minorities in Turkey and Turkic-speaking minorities of the peripheries*, vol. 111. Wiesbaden: Harrassowitz.

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (version 7.0)*. (`https://multicast.aspra.uni-bamberg.de/#annotations`) (Accessed 2019-03-08).

Haig, Geoffrey & Schnell, Stefan. 2016b. The discourse basis of ergativity revisited. *Language* 92(3). 591–618. *(*`https://doi.org/10.1353/lan.2016.0049`*)*.

Halliday, M. A. K. & Hasan, Ruqaiya. 1976. *Cohesion in English*. London: Longman.

Hammarström, Harald & Forkel, Robert & Haspelmath, Martin. 2019. *Glottolog 4.0*. Jena: Max Planck Institute for the Science of Human History. (`http://glottolog.org`) (Accessed 2019-08-04).

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(2). 161–195.

Huddleston, Rodney D. & Pullum, Geoffrey K. (eds.). 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Kimoto, Yukinori. 2017. *A grammar of Arta: A Philippine Negrito language*. Kyoto: Kyoto University Ph.D. dissertation.

Laufer, Carl. 1959. P. Futschers Aufzeichnungen über die Butam-Sprache (Neubritannien). *Anthropos* 54(1/2). 183–212.

Longacre, Robert. 1960. String constituent analysis. *Language* 36(1). 63–88.

Meng, Chenxi. 2018. *A grammar of Tulil*. Melbourne: La Trobe University Ph.D. dissertation.

Mettouchi, Amina & Martine, Vanhove & Caubet, Dominique (eds.). 2015. *Corpus-based studies of lesser-described languages: The CorpAfroAs corpus of spoken AfroAsiatic languages* (Studies in Corpus Linguistics 68). Amsterdam: John Benjamins.

Mosel, Ulrike. 2015. *Searches in ELAN with regular expressions*. (https://tla.mpi.nl/wp-content/uploads/2011/12/Searches_in_ELAN_with_regular_expressions.pdf) (Accessed 2019-08-19).
Mosel, Ulrike. 2019. A multifunctional Teop-English dictionary. *Dictionaria* 4(1-6488). *(*https://doi.org/10.5281/zenodo.3257580*)*. (https://dictionaria.clld.org/contributions/teop) (Accessed 2019-08-19).
Mosel, Ulrike & Thiesen, Yvonne. 2007. *The Teop sketch grammar*. Kiel: University of Kiel. (https://hdl.handle.net/1839/00-0000-0000-0008-24F6-3@view) (Accessed 2016-05-14).
Noonan, Michael. 2003. *A crosslinguistic investigation of referential density*. Milwaukee: University of Wisconsin-Milwaukee. (http://crossasia-repository.ub.uni-heidelberg.de/190/) (Accessed 2016-02-08).
Öpengin, Ergin & Haig, Geoffrey. 2014. Regional variation in Kurmanji: A preliminary classification of dialects. *Kurdish Studies* 2(2). 143–176.
Pike, Kenneth. 1964. Discourse analysis and tagmeme matrices. *Oceanic Linguistics* 3(1). 5–26.
Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Cole, Peter (ed.), *Radical pragmatics*, 223–255. New York: Academic Press.
R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. (https://www.R-project.org) (Accessed 2019-08-23).
Riester, Arndt & Baumann, Stefan. 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) (Accessed 2018-03-01).
Ross, Malcolm D. 1988. *Proto Oceanic and the Austronesian languages of western Melanesia*. Canberra: Pacific Linguistics.
Schiborr, Nils N. & Schnell, Stefan & Thiele, Hanna. 2018. *RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines (v1.1)*. Bamberg / Melbourne: University of Bamberg. (https://multicast.aspra.uni-bamberg.de/#annotations) (Accessed 2019-03-08).
Schnell, Stefan. 2010. *Animacy and referentiality in Vera'a*. Kiel: Kiel University Ph.D. dissertation.
Schnell, Stefan. 2011. *A grammar of Vera'a*. Kiel: University of Kiel Ph.D. dissertation. (https://www.academia.edu/2317752/Schnell_00002011_A_grammar_of_Veraa_an_Oceanic_language_of_North_Vanuatu) (Accessed 2016-02-22).
Schnell, Stefan. 2016. Multi-CAST Vera'a annotation notes. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://multicast.aspra.uni-bamberg.de/#veraa) (Accessed 2019-03-08).
Sneddon, James N. 1975. *Tondano phonology and grammar*. Canberra: Pacific Linguistics.
Thieberger, Nick. 2004. Documentation in practice: Developing a linked media corpus of South Efate. In Austin, Peter (ed.), *Language documentation and description*, 169–178. London: Hans Rausing Endangered Languages Project, SOAS.
Thieberger, Nick. 2006. *A grammar of South Efate: An Oceanic language of Vanuatu*. Honolulu: University of Hawaii Press. (http://hdl.handle.net/11343/31242) (Accessed 2019-07-31).
Vores, Makson & Schnell, Stefan & Atkins, George B. & Malau, Armstrong & Malau, Catriona. 2012. *Stories from the village of Vera'a (Vanua Lava, TorBa)*. Port Vila: Spite Design.
Wolff, John. 2010. *Proto-Austronesian phonology*. Ithaca, NY: Cornell Southeast Asia Program Publications.
Wurm, Stephen A. & Hattori, Shirō. 1981. *Language atlas of the Pacific area* (Pacific Linguistics Series C, 66 and 67). Canberra: Australian National University.

Appendices

A List of texts (2001)

corpus	*text name*	*speaker*	*year rec'd*	*text type*	*RefIND ISNRef*	*length in hh:mm:ss*	*clause units*
arta	*alisiya*	AR01	2018	AN	no	00:02:54	55
arta	*arsenyo*	AR02	2018	AN	no	00:07:03	65
arta	*child*	AR03	2018	AN	no	00:14:11	163
arta	*delia*	AR03	2018	AN	no	00:10:24	120
arta	*disubu*	AR03	2018	AN	no	00:05:05	59
arta	*hapon*	AR02	2018	AN	no	00:11:27	132
arta	*husband*	AR01	2018	AN	no	00:04:10	47
arta	*marry*	AR02	2018	AN	no	00:03:30	45
arta	*swateng*	AR03	2018	TN	no	00:10:36	190
arta	*typhoon*	AR01	2018	AN	no	00:05:54	72
arta	*udulan*	AR03	2018	TN	no	00:06:13	82
cypgreek	*jitros*	CG01	1960	TN	yes	—	272
cypgreek	*minaes*	CG01	1960	TN	yes	—	359
cypgreek	*psarin*	CG01	1964	TN	yes	—	440
english	*devon01*	EN02	1980	AN	yes	00:31:40	590
english	*kent01*	EN01	1975	AN	yes	00:27:28	622
english	*kent02*	EN01	1975	AN	yes	*(a)* 00:30:00	*(a)* 765
						(b) 00:32:41	*(b)* 872
english	*kent03*	EN03	1976	AN	yes	*(a)* 00:25:07	*(a)* 675
						(b) 00:29:59	*(b)* 660
english	*london01*	EN04	1985	AN	no	*(a)* 00:29:16	*(a)* 733
						(b) 00:29:40	*(b)* 732
mandarin	*hml*	MD01	2015	TN	yes	00:10:25	301
mandarin	*jgz*	MD02	2015	TN	yes	00:21:14	711
mandarin	*lzh*	MD03	2015	TN	yes	00:08:13	182
nafsan	*kori*	NF01	1998	TN	yes	00:06:49	284
nafsan	*lelep*	NF01	1998	TN	yes	00:03:28	129
nafsan	*lisau*	NF02	1998	TN	yes	00:02:44	58
nafsan	*litog*	NF02	1998	TN	yes	00:03:32	86
nafsan	*maal*	NF03	1997	TN	yes	00:03:09	52
nafsan	*nmatu*	NF03	1996	TN	yes	00:04:15	88
nafsan	*ntwam*	NF04	1996	TN	yes	00:07:41	186
nafsan	*taapes*	NF02	1998	TN	yes	00:02:09	67
nafsan	*tafra*	NF03	1997	TN	yes	00:04:20	64
nkurd	*muserz01*	NK01	2000	TN	yes	00:19:53	627
nkurd	*muserz02*	NK02	2002	TN	no	00:12:12	482
nkurd	*muserz03*	NK01	2000	TN	yes	00:19:51	732
persian	*g1-f-01*	PS01	2015	SN	no	00:01:34	47
persian	*g1-f-02*	PS02	2015	SN	no	00:02:10	54
persian	*g1-f-05*	PS05	2015	SN	no	00:02:15	60
persian	*g1-f-07*	PS07	2015	SN	no	00:01:05	38
persian	*g1-f-08*	PS08	2015	SN	no	00:01:40	25
persian	*g1-f-09*	PS09	2015	SN	no	00:04:29	100
persian	*g1-f-10*	PS10	2015	SN	no	00:03:19	83

continue →

corpus	text name	speaker	year rec'd	text type	RefIND ISNRef	length in hh:mm:ss	clause units
persian	g1-f-11	PS11	2015	SN	no	00:01:42	60
persian	g1-f-12	PS12	2015	SN	no	00:01:45	49
persian	g1-f-14	PS14	2015	SN	no	00:03:03	99
persian	g1-m-03	PS03	2015	SN	no	00:00:45	17
persian	g1-m-04	PS04	2015	SN	no	00:02:03	61
persian	g1-m-06	PS06	2015	SN	no	00:00:51	22
persian	g1-m-13	PS13	2015	SN	no	00:02:50	69
persian	g2-f-01	PS15	2015	SN	no	00:02:23	58
persian	g2-f-02	PS16	2015	SN	no	00:01:27	44
persian	g2-f-03	PS17	2015	SN	no	00:01:37	40
persian	g2-f-04	PS18	2015	SN	no	00:01:03	25
persian	g2-f-05	PS19	2015	SN	no	00:01:52	26
persian	g2-f-06	PS20	2015	SN	no	00:01:27	56
persian	g2-f-07	PS21	2015	SN	no	00:01:41	52
persian	g2-m-08	PS22	2015	SN	no	00:01:44	49
persian	g2-m-09	PS23	2015	SN	no	00:01:20	42
persian	g2-m-10	PS24	2015	SN	no	00:01:17	41
persian	g2-m-11	PS25	2015	SN	no	00:01:01	25
persian	g2-m-12	PS26	2015	SN	no	00:01:08	40
persian	g2-m-13	PS27	2015	SN	no	00:01:23	52
persian	g2-m-14	PS28	2015	SN	no	00:01:03	36
persian	g2-m-15	PS29	2015	SN	no	00:02:35	48
sanzhi	asabali	SD01	2012	AN	yes	00:06:22	142
sanzhi	bazhuk	SD02	2013	TN	yes	00:04:18	99
sanzhi	dragon	SD02	2013	TN	yes	00:05:35	121
sanzhi	kurban	SD03	2011	AN	yes	00:04:25	164
sanzhi	mill	SD01	2013	TN	yes	00:05:10	130
sanzhi	patima	SD02	2013	TN	yes	00:05:13	133
sanzhi	ramazan	SD04	2012	AN	yes	00:07:16	209
sanzhi	tape	SD03	2011	AN	yes	00:01:47	68
teop	iar	TP01	2003	TN	yes	00:14:34	348
teop	mat	TP02	2004	TN	yes	00:06:54	207
teop	sii	TP03	2004	TN	yes	00:19:21	589
teop	viv	TP04	2004	TN	yes	00:05:46	158
tondano	gulamera	TD01	2011	SN	no	00:10:15	129
tondano	holiday	TD01	2011	AN	no	00:05:16	89
tondano	kiniar01	TD02	2013	SN	no	00:08:50	142
tondano	kiniar02	TD03	2013	SN	no	00:12:36	193
tondano	kiniar03	TD03	2013	SN	no	00:08:46	99
tondano	mapalus	TD04	2011	AN	no	00:06:51	150
tondano	water	TD05	2011	SN	no	00:05:04	80
tondano	watulaney	TD06	2011	AN	no	00:18:20	203
tulil	all1	TL01	2012	TN	yes	00:05:20	93
tulil	alrm	TL01	2014	AN	yes	00:23:02	407
tulil	jkpp	TL02	2014	AN	yes	00:25:28	414
tulil	lnsl	TL03	2014	TN	yes	00:06:27	92
tulil	lrdw	TL04	2007	TN	yes	00:08:24	157
tulil	sves	TL05	2014	TN	yes	00:05:08	101

continue →

corpus	*text name*	*speaker*	*year rec'd*	*text type*	*RefIND ISNRef*	*length in hh:mm:ss*	*clause units*
veraa	*anv*	VR01	2007	TN	yes	00:06:07	182
veraa	*as1*	VR02	2007	TN	yes	00:05:16	213
veraa	*gabg*	VR03	2007	TN	yes	00:08:41	174
veraa	*gaqg*	VR04	2007	TN	yes	00:08:52	226
veraa	*hhak*	VR05	2007	TN	yes	00:12:39	432
veraa	*isam*	VR06	2007	TN	yes	00:07:21	238
veraa	*iswm*	VR07	2007	TN	yes	00:21:43	576
veraa	*jjq*	VR08	2007	TN	yes	00:30:19	880
veraa	*mvbw*	VR09	2007	TN	yes	00:10:07	307
veraa	*pala*	VR10	2007	TN	yes	*(a)* 00:04:02 *(b)* 00:06:41	*(a)* 141 *(b)* 239
collection totals					99 texts	14:18:19	21 540

Table A.1 List of texts in the Multi-CAST collection as of January 2020, version `2001`.
TN = traditional narratives, AN = autobiographical narratives, SN = stimulus-based narratives.

B List of speakers (2001)

speaker	*corpus*	*text(s)*	*gender*	*age(s)*	*year born*	*year(s) rec'd*	*clause units*
AR01	*arta*	*alisiya, husband, typhoon*	female	c60	c1960	2018	174
AR02	*arta*	*arsenyo, hapon, marry*	male	c63	c1955	2018	242
AR03	*arta*	*child, delia, disubu, swateng, udulan*	male	c65	c1955	2018	614
CG01	*cypgreek*	*jitros, minaes, psarin*	female	73, 77	1887	1960, 1964	1071
EN01	*english*	*kent01, kent02*	male	85	1890	1975	2259
EN02	*english*	*devon01*	male	c80	c1900	1980	590
EN03	*english*	*kent03*	male	87	1889	1976	1335
EN04	*english*	*london01*	male	61	1924	1985	1465
MD01	*mandarin*	*hml*	male	23	1992	2015	301
MD02	*mandarin*	*jgz*	male	23	1992	2015	711
MD03	*mandarin*	*lzh*	male	22	1993	2015	182
NF01	*nafsan*	*kori, lelep*	male	65	1933	1998	413
NF02	*nafsan*	*lisau, litog, taapes*	female	67	1931	1998	211
NF03	*nafsan*	*maal, nmatu, tafra*	male	85	1912	1996, 1997	202
NF04	*nafsan*	*ntwam*	male	45	1951	1996	186
NK01	*nkurd*	*muserz01, muserz03*	male	c50	c1950	2000	1359
NK02	*nkurd*	*muserz02*	female	c60	c1940	2002	482
PS01	*persian*	*g1-f-01*	female	39	1976	2015	47
PS02	*persian*	*g1-f-02*	female	29	1986	2015	54
PS03	*persian*	*g1-m-03*	male	22	1993	2015	17
PS04	*persian*	*g1-m-04*	male	25	1990	2015	61
PS05	*persian*	*g1-f-05*	female	26	1989	2015	60
PS06	*persian*	*g1-m-06*	male	32	1983	2015	22
PS07	*persian*	*g1-f-07*	female	25	1990	2015	38
PS08	*persian*	*g1-f-08*	female	25	1990	2015	25
PS09	*persian*	*g1-f-09*	female	25	1990	2015	100
PS10	*persian*	*g1-f-10*	female	31	1984	2015	83
PS11	*persian*	*g1-f-11*	female	33	1982	2015	60
PS12	*persian*	*g1-f-12*	female	33	1982	2015	49

continue →

speaker	*corpus*	*text(s)*	*gender*	*age(s)*	*year born*	*year(s) rec'd*	*clause units*
PS13	*persian*	*g1-m-13*	male	35	1980	2015	69
PS14	*persian*	*g1-f-14*	female	29	1986	2015	99
PS15	*persian*	*g2-f-01*	female	20	1995	2015	58
PS16	*persian*	*g2-f-02*	female	20	1995	2015	44
PS17	*persian*	*g2-f-03*	female	20	1995	2015	40
PS18	*persian*	*g2-f-04*	female	20	1995	2015	25
PS19	*persian*	*g2-f-05*	female	20	1995	2015	26
PS20	*persian*	*g2-f-06*	female	38	1977	2015	56
PS21	*persian*	*g2-f-07*	female	33	1982	2015	52
PS22	*persian*	*g2-m-08*	male	20	1995	2015	49
PS23	*persian*	*g2-m-09*	male	22	1993	2015	42
PS24	*persian*	*g2-m-10*	male	20	1995	2015	41
PS25	*persian*	*g2-m-11*	male	25	1990	2015	25
PS26	*persian*	*g2-m-12*	male	20	1995	2015	40
PS27	*persian*	*g2-m-13*	male	20	1995	2015	52
PS28	*persian*	*g2-m-14*	male	20	1995	2015	36
PS29	*persian*	*g2-m-15*	male	27	1988	2015	48
SD01	*sanzhi*	*asabali, mill*	male	76, 77	1935	2012, 2013	272
SD02	*sanzhi*	*bazhuk, dragon, patima*	male	51	1963	2013	353
SD03	*sanzhi*	*kurban, tape*	male	60	1951	2011	232
SD04	*sanzhi*	*ramazan*	male	58	1954	2012	209
TP01	*teop*	*iar*	female	c70	c1930	2003	348
TP02	*teop*	*mat*	female	c30	c1970	2004	207
TP03	*teop*	*sii*	female	c60	c1940	2004	589
TP04	*teop*	*viv*	female	c30	c1970	2004	158
TD01	*tondano*	*gulamera, holiday*	female	c50	c1960	2011	218
TD02	*tondano*	*kiniar01*	male	c40	c1970	2013	142
TD03	*tondano*	*kiniar02, kiniar03*	male	c50	c1960	2013	292
TD04	*tondano*	*mapalus*	female	c50	c1960	2011	150
TD05	*tondano*	*water*	female	c40	c1970	2011	80
TD06	*tondano*	*watulaney*	female	c40	c1970	2011	203
TL01	*tulil*	*all1, alrm*	male	53, 55,	1959	2012, 2014	500
TL02	*tulil*	*jkpp*	male	74	1940	2014	414
TL03	*tulil*	*lnsl*	male	c55	c1960	2014	92
TL04	*tulil*	*lrdw*	male	77	1930	2007	157
TL05	*tulil*	*sves*	female	c80	c1930	2014	101
VR01	*veraa*	*anv*	female	c20	c1985	2007	182
VR02	*veraa*	*as1*	male	c40	c1965	2007	213
VR03	*veraa*	*gabg*	male	c40	c1965	2007	174
VR04	*veraa*	*gaqg*	male	c40	c1965	2007	226

continue →

speaker	*corpus*	*text(s)*	*gender*	*age(s)*	*year born*	*year(s) rec'd*	*clause units*
VR05	*veraa*	*hhak*	male	c20	c1985	2007	432
VR06	*veraa*	*isam*	male	c60	c1950	2007	238
VR07	*veraa*	*iswm*	male	c60	c1950	2007	576
VR08	*veraa*	*jjq*	male	c60	c1950	2007	880
VR09	*veraa*	*mvbw*	male	c30	c1975	2007	307
VR10	*veraa*	*pala*	female	c40	c1965	2007	380

Table B.1 List of speakers in the Multi-CAST collection as of January 2020, version 2001.

C Changelog

The following is a timeline of the additions and alternations to the Multi-CAST collection and its annotations. As a rule, all data in the collection is updated at once and published as a self-contained version. Each successive version is associated with a unique four-digit version number (e.g. "1505"). These identifiers can be used with the *multicastR* package to access specific earlier states of the annotations; see Section 6 for details.

Release version 2001 (12 January 2020)

- added 1 new corpus:
  - Mandarin [mandarin] (Vollmer 2020)
- minor improvements to the glosses and annotations in the Arta, English, Nafsan, Sanzhi Dargwa, and Tulil corpora

Mandarin [mandarin]

- added new corpus with 3 texts: *hml, jgz, lzh*
- added RefIND and ISNRef annotations to all texts

Release version 1908 (30 August 2019)

- added 2 new corpora:
  - Arta [arta] (Kimoto 2019)
  - Nafsan [nafsan] (Thieberger & Brickell 2019)
- relabelled the annotations with (a drastically simplified version of) the "RefLex" scheme to "ISNRef" for clarity; accordingly, the corresponding EAF tier, XML node, and TSV column have been renamed as well

Arta [arta]

- added new corpus with 11 texts: *alisiya*, *arsenyo*, *child*, *delia*, *disubu*, *hapon*, *husband*, *marry*, *swateng*, *typhoon*, *udulan*

English [english]

- fully revised annotations of 2 texts: *kent01*, *kent02*
- added 3 new texts: *devon01*, *kent03*, *london01*
- added RefIND and ISNRef annotations to 4 texts: *devon01*, *kent01*, *kent02*, *kent03*

Nafsan [nafsan]

- added new corpus with 9 texts: *kori*, *lelep*, *lisau*, *litog*, *maal*, *nmatu*, *ntwam*, *taapes*, *tafra*
- added RefIND and ISNRef annotations to all texts

Northern Kurdish [nkurd]

- minor improvements to the glosses and annotations

Sanzhi Dargwa [sanzhi]

- minor improvements to the glosses and annotations

Teop [teop]

- added GRAID symbols:
  - ⟨conj_other⟩ 'conjunction'
  - ⟨pn_np⟩ 'proper name'
- minor improvements to the glosses and annotations

Release version 1907 (30 July 2019)

- added 1 new corpora:
  - Tulil [tulil] (Meng 2019)

Northern Kurdish [nkurd]

- updated the citation for the corpus, adding Maria Vollmer as co-author
- fully revised annotations of 2 texts: *muserz01*, *muserz02*
- added 1 new text: *muserz03*
- added RefIND and ISNRef annotations to 2 texts: *muserz01*, *muserz03*

Sanzhi Dargwa [sanzhi]

- updated GRAID symbols:
  - ⟨rc_f0⟩ → ⟨rel_f0⟩ 'gapped argument of a relative clause'
- minor improvements to the glosses and annotations

Tulil [tulil]

- added new corpus with 6 texts: *all1, alrm, jkpp, lnsl, lrdw, sves*
- added RefIND and ISNRef annotations to all texts

Vera'a [veraa]

- minor improvements to the glosses and annotations

Release version 1905 (4 May 2019)

- added 1 new corpus:
  - Sanzhi Dargwa [sanzhi] (Forker & Schiborr 2019)
- added annotations with the RefIND (Schiborr et al. 2018) and ISNRef (based on Riester & Baumann 2017) schemes to a number of corpora
- new EAF annotation tiers:
  - `refind` (a 1-to-1 child of the `graid` tier)
  - `isnref` (a 1-to-1 child of the `refind` tier)
- documentation: added *Lists of referents* (as PDF and TSV) for all corpora with RefIND annotations
- texts split into multiple parts are now numbered continuously across all parts on the `utterance_id` annotation tier of the EAF files
- two new file formats for the annotations: XML and TSV
- the GRAID symbols ⟨-⟩ 'bound form' and ⟨=⟩ 'cliticized form' now always attach to the form gloss (i.e. at the left edge of the GRAID gloss), irrespective of the direction of attachment

Cypriot Greek [cypgreek]

- added RefIND and ISNRef annotations to all texts
- updated GRAID symbols for consistency:
  - ⟨other.cop⟩ → ⟨cop_other⟩ 'copula, other'
  - ⟨predex⟩ → ⟨other:predex⟩ 'predicate of an existential construction'
  - ⟨:dtp⟩ → ⟨:dt_p⟩ 'dislocated topic, P role'
  - ⟨:dtobl⟩ → ⟨:dt_obl⟩ 'dislocated topic, oblique'
  - ⟨aux_l⟩ → ⟨lv_aux⟩ 'auxiliary'
  - ⟨l_aux⟩ → ⟨lv_aux⟩ 'auxiliary'
  - ⟨other.lv_aux⟩ → ⟨lv_aux_other⟩ 'auxiliary, other'
  - ⟨#*_neg⟩ → ⟨#*.neg⟩ 'negated clause tag'
- minor improvements to the glosses and annotations

English [english]

- updated GRAID symbols for consistency:
  - ⟨inter_pro⟩ → ⟨intrg_pro⟩ 'interrogative pronoun'
  - ⟨indef_pro⟩ → ⟨indef_other⟩ 'indefinite pronoun'
  - ⟨refl_pro⟩ → ⟨refl⟩ 'reflexive pronoun'
- minor improvements to the glosses and annotations

Northern Kurdish [nkurd]

- updated GRAID symbols for consistency:
  - ⟨inter_pro⟩ → ⟨intrg_pro⟩ 'interrogative pronoun'
  - ⟨refl_pro⟩ → ⟨refl⟩ 'reflexive pronoun'
  - ⟨*:poss⟩ → ⟨rn_*:poss⟩ 'NP-internal possessives'
  - ⟨excl⟩ → ⟨excl_other⟩ 'exclamation'
- minor improvements to the glosses and annotations

Persian [persian]

- updated GRAID symbols for consistency:
  - ⟨ind_pro⟩ → ⟨indef_other⟩ 'indefinite pronoun'
  - ⟨*:poss⟩ → ⟨rn_*:poss⟩ 'NP-internal possessives'
  - ⟨acc_rn⟩ → ⟨rn_acc⟩ 'postpositional object particle'
  - ⟨lvc⟩ → ⟨other:lvc⟩ 'non-verbal complement of a complex predicate'
- minor improvements to the glosses and annotations

Sanzhi Dargwa [sanzhi]

- added new corpus with 8 texts: *asabali, bazhuk, dragon, kurban, mill, patima, ramazan, tape*
- added RefIND and ISNRef annotations to all texts

Teop [teop]

- added RefIND and ISNRef annotations to all texts
- person/animacy and function symbols on cross-indices (⟨rv-pro⟩ and ⟨rv-pl_pro⟩) are now delimited by an underscore ⟨_⟩, e.g. ⟨rv-pro_h_s⟩
- updated GRAID symbols for consistency:
  - ⟨rn_#rc⟩ → ⟨#rc_rn⟩ 'NP-internal relative clause'
  - ⟨int_np⟩ → ⟨intrg_other⟩ 'interrogative pronoun'
  - ⟨int_other⟩ → ⟨intrg_other⟩ 'interrogative pronoun'
  - ⟨rv_n⟩ → ⟨rv_np⟩ 'NP inside the verbal complex'
- minor improvements to the glosses and annotations

Tondano [tondano]

- minor improvements to the glosses and annotations

Vera'a [veraa]

- added RefIND and ISNRef annotations to all texts
- person/animacy and function symbols on cross-indices (⟨lv-pro⟩) are now delimited by an underscore ⟨_⟩, e.g. ⟨lv-pro_h_s⟩
- updated GRAID symbols for consistency:
  - ⟨rn_#rc⟩ → ⟨#rc_rn⟩ 'NP-internal relative clause'
  - ⟨rv_pro*:p⟩ → ⟨pro*:p⟩ 'object pronouns'
  - ⟨dl_*⟩ → ø 'dual/paucal form'
  - ⟨tl_*⟩ → ø 'trial form'
  - ⟨pl_*⟩ → ø 'plural form'
- minor improvements to the glosses and annotations

Release version 1606 (1 June 2016)

- added 2 new corpora:
  - Persian [persian] (Adibifar 2016)
  - Tondano [tondano] (Brickell 2016)

Persian [persian]

- added new corpus with 29 texts: *g1-f-01, g1-f-02, g1-f-05, g1-f-07, g1-f-08, g1-f-09, g1-f-10, g1-f-11, g1-f-12, g1-f-14, g1-m-03, g1-m-04, g1-m-06, g1-m-13, g2-f-01, g2-f-02, g2-f-03, g2-f-04, g2-f-05, g2-f-06, g2-f-07, g2-m-08, g2-m-09, g2-m-10, g2-m-11, g2-m-12, g2-m-13, g2-m-14, g2-m-15*

Tondano [tondano]

- added new corpus with 8 texts: *gulamera, holiday, kiniar01, kiniar02, kiniar03, mapalus, water, watulaney*

Release version 1505 (1 May 2015)

- added 5 new corpora:
  - Cypriot Greek [cypgreek] (Hadjidas & Vollmer 2015)
  - English [english] (Schiborr 2015)
  - Northern Kurdish [nkurd] (Haig & Thiele 2015)
  - Teop [teop] (Mosel & Schnell 2015)
  - Vera'a [veraa] (Schnell 2015)

Cypriot Greek [cypgreek]

- added new corpus with 3 texts: *jitros, minaes, psarin*

English [english]

- added new corpus with 3 texts: *kent01, kent02a, kent02b*

Northern Kurdish [nkurd]

- added new corpus with 2 texts: *muserz01, muserz02*

Teop [teop]

- added new corpus with 4 texts: *iar, mat, sii, viv*

Vera'a [veraa]

- added new corpus with 11 texts: *anv, as1, gabg, gaqg, hhak, isam, iswm, jjq, mvbw, palaa, palab*

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/