Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

Corpus overview

— and description —

July 2016

Nils Norman Schiborr
University of Bamberg

cite this document as:

Schiborr, Nils N. 2016.
Multi-CAST corpus overview and description.
In Haig, Geoffrey & Schnell, Stefan (eds.),
Multi-CAST (Multilingual Corpus of Annotated Spoken Texts).
(handle) (date accessed.)

file v1.4

Contents

1 Introduction

Multi-CAST (Multilingual Corpus of Annotated Spoken Texts) is a collection of annotated non-elicited, spoken texts from diverse languages, most of them monologic narratives. Multi-CAST was compiled and annotated under the supervision of Geoffrey Haig and Stefan Schnell, with technical implementation and archiving undertaken by the Language Archive Cologne (LAC) at the University of Cologne.

The recorded texts in the Multi-CAST corpora adhere to common principles of annotation. For each text in each corpus, a sound file, translation, as well as morphological glossing and syntactic annotations are provided, along with background information on the recordings and additional sources. The annotations are available as EAF-files, an XML-based file format produced by the annotation software ELAN.[1]

The corpus overview and description serves to document the contents of the collection, its structure, and the decisions that went into its design. As Multi-CAST is amenable to the inclusion of further data sets, this description is intended to evolve alongside it. The most recent version of this document can be found on the archive homepage,[2] and a timeline of additions and changes to the corpus is provided at the end of this file.

1.1 Citing Multi-CAST

The entire collection should be cited as follows:

> Haig, Geoffrey & Schnell, Stefan (eds.). 2016.
> *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts).*
> (https://lac.uni-koeln.de/multicast/) (date accessed.)

The individual corpora in Multi-CAST should be cited as contributions to the whole collection, analogously to the following:

> Mosel, Ulrike & Schnell, Stefan. 2016. Teop.
> In Haig, Geoffrey & Schnell, Stefan (eds.),
> *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts).*
> (https://lac.uni-koeln.de/en/multicast-teop/) (date accessed.)

Please refer to individual files in the corpora by their permanent archive handles ('permalinks') as given on the webpages of the archive.

[1] https://tla.mpi.nl/tools/tla-tools/elan/
[2] https://lac.uni-koeln.de/multicast/
[3] http://creativecommons.org/licenses/by-nc-sa/4.0/

1.2 Research context

Multi-CAST has been designed to address research questions from areas of study such as discourse structure and referentiality, including the notions of *preferred argument structure* (Du Bois 1987, 2003), *referential density* (Bickel 2003; Noonan 2003), and *accessibility theory* (Ariel 1990, 2004). For further discussion of the research background, together with exemplary applications, see the *Multi-CAST research context* (Haig & Schnell 2016b), available on the archive webpage.[4]

1.3 Acknowledgements

Data collection and annotation of part of the collection were graciously supported by the Australian Research Council as part of the DECRA project *Typology of language use* (Stefan Schnell, 2012–2015), hosted by La Trobe University, and by the VolkswagenStiftung-funded *Documentation of Endangered Languages* project (DOBES)[5] (Ulrike Mosel, 2000–2007; Stefan Schnell, 2006–2012). The Lehrstuhl für Allgemeine Sprachwissenschaft at the University of Bamberg contributed departmental funding and research infrastructure to the project.

The English texts were made available in cooperation with the University of Freiburg via the Freiburg English Dialect Corpus (FRED), supervised by Bernd Kortmann and Lieselotte Anderwald.

We would like to thank Felix Rau and the staff of the Language Archive Cologne (LAC) at the University of Cologne for maintaining Multi-CAST as part of the archive.

Editors	Geoffrey Haig (University of Bamberg) Stefan Schnell (University of Melbourne)
Assistant editor	Nils Norman Schiborr (Bamberg)
Contributors	Shirin Adibifar (Bamberg) / Persian Timothy Brickell (La Trobe University) / Tondano Harris Hadjidas (Bamberg) / Cypriot Greek Geoffrey Haig / Northern Kurdish Ulrike Mosel (University of Kiel) / Teop Nils Norman Schiborr / English Stefan Schnell / Vera'a, Teop Hanna Thiele (Bamberg/Freiburg) / Northern Kurdish Maria Vollmer (Bamberg) / Cypriot Greek

The contributors and editorial staff are indebted to their respective communities for their support and stimulating criticism.

[4] https://lac.uni-koeln.de/multicast/

[5] http://dobes.mpi.nl/

2 Corpus overview

Multi-CAST contains data from seven languages — Cypriot Greek, Northern Kurdish (also known as Kurmanji), English, Persian, Teop, Tondano, and Vera'a — and various narrative text types including folktales, oral history, and *Pear Film* retellings.[6] A summary of the Multi-CAST corpora is given in Table 1, and Figure 1 shows the area where each of the languages is spoken. Multi-CAST contains roughly 12,000 clause units and 65,000 words in 7 hours of recordings.

In the following, a brief overview of the corpora and their annotators is provided. Further details on the languages and individual recordings can be found on each corpus' metadata sheet.

- **Cypriot Greek** (*cypgreek*)
  - ISO 639-3: ell
  - affiliation: Indo-European, Greek, Attic
  - area spoken: Cyprus
  - variety rec'd: Yeri-pyroi
  - text type: traditional narratives
  - annotations: Vollmer & Hadjidas (2016)
- **English** (*english*)
  - ISO 639-3: eng
  - affiliation: Indo-European, Germanic, West
  - area spoken: United Kingdom
  - variety rec'd: South East England
  - text type: autobiographical narratives
  - annotations: Schiborr (2016)
- **Northern Kurdish** (*nkurd*)
  - ISO 639-3: kmr
  - affiliation: Indo-European, Iranian, Northwestern
  - area spoken: Turkey, East / Iraq, North / Iran, West
  - variety rec'd: Northern Kurmanji, Erzurum / Muş
  - text type: traditional narratives
  - annotations: Haig & Thiele (2016)
- **Persian** (*persian*)
  - ISO 639-3: pes
  - affiliation: Indo-European, Iranian, Southwestern
  - area spoken: Iran
  - varieties rec'd: Farsi, Tehran / Sari
  - text type: stimulus-based narratives
  - annotations: Adibifar (2016)

[6] See Mettouchi et al. (2015) for a similar approach in developing corpora from African languages for comparative purposes.

- **Teop** (*teop*)
  - ISO 639-3: tio
  - affiliation: Austronesian, Malayo-Polynesian, Oceanic, Nehan-Bougainville
  - area spoken: Papua New Guinea, Bougainville
  - variety rec'd: Teop island
  - text type: traditional narratives
  - annotations: Mosel & Schnell (2016)
- **Tondano** (*tondano*)
  - ISO 639-3: tdn
  - affiliation: Austronesian, Malayo-Polynesian, Philippine, Minahasan, North, Northeast
  - area spoken: Indonesia, North Sulawesi, Tondano town
  - variety rec'd: Toulour dialect
  - text types: autobiographical / stimulus-based narratives
  - annotations: Brickell (2016)
- **Vera'a** (*veraa*)
  - ISO 639-3: vra
  - affiliation: Austronesian, Malayo-Polynesian, Oceanic, Vanuatu
  - area spoken: Vanuatu, Banks Islands, Vanua Lava
  - variety rec'd: Vera'a village
  - text type: traditional narratives
  - annotations: Schnell (2016)

Lists of all corpus texts and speakers are respectively provided in Appendix A and Appendix B.

2.1 Corpus components

The core component of the Multi-CAST corpora are natural language texts ('original texts' in terms of Chapter 4 of Haig et al. 2011) that have been recorded in their respective cultural contexts (where possible), transcribed, and annotated across multiple levels. Transcriptions and annotations are provided as EAF files, an XML-based file format produced by the annotation software ELAN. The internal structure of the EAF files in Multi-CAST is discussed below in Section 3.1.

In addition to the audio recordings and accompanying annotations, each corpus contains a number of supplementary files. These serve both to document its structure and ensure the accountability of the annotators' decisions, and to allow users interested in the project a top-level look into the data. Table 2 lists the supplementary files included with each of the corpora. Some of these files are provided per corpus, some per text, some per text part. Their file names follow a consistent pattern: all begin with the name of the *corpus*, followed, where applicable, by the title of the *text*. Longer recordings have been split into multiple *parts*, each of which is labelled with an additional

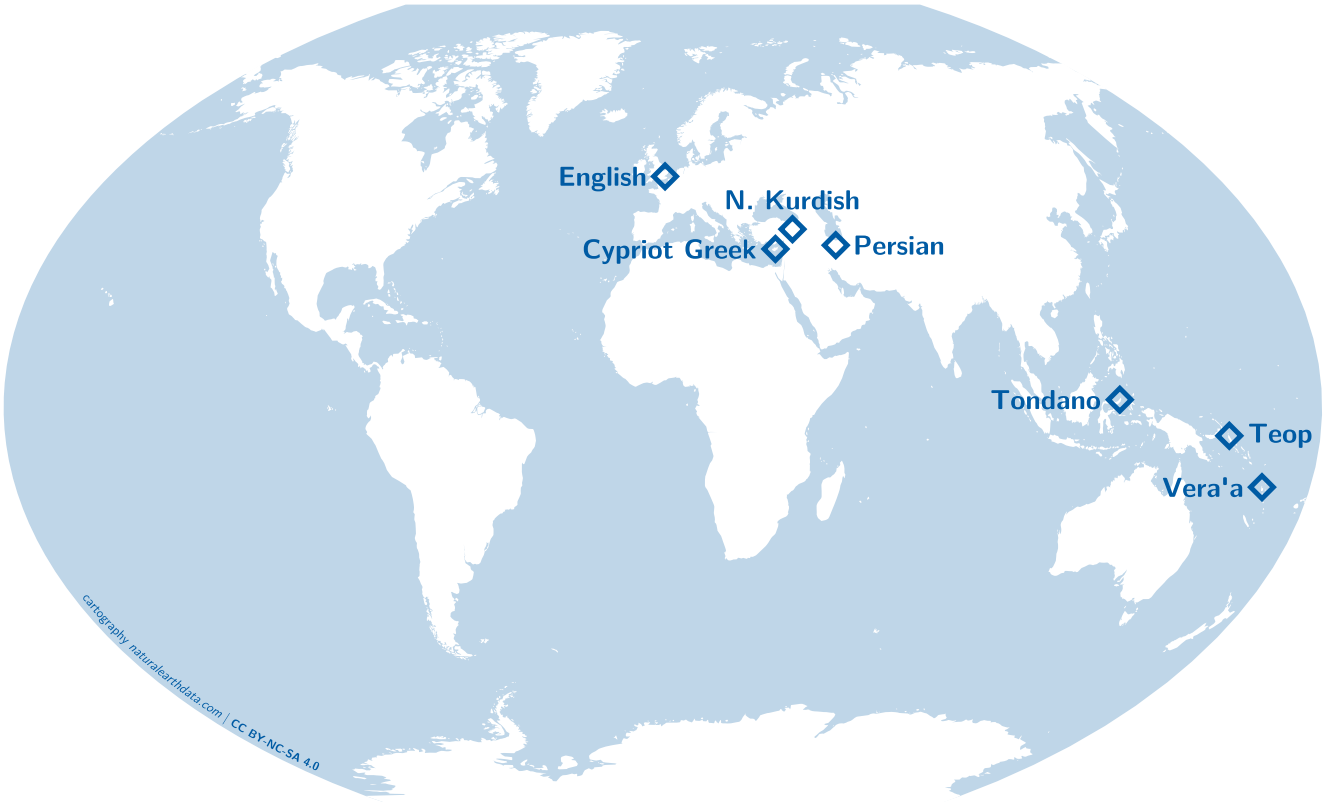

Figure 1. Multi-CAST language locator map.

language	corpus	text type	texts	length h:mm:ss	clause units
Cypriot Greek	*cypgreek*	traditional	3	—	1,071
English	*english*	autobiogr.	2	1:30:09	2,245
Northern Kurdish	*nkurd*	traditional	2	0:32:05	1,101
Persian	*persian*	stim.-based	29	0:52:32	1,417
Teop	*teop*	traditional	4	0:46:35	1,302
Tondano	*tondano*	auto./st.-b.	8	1:16:18	1,086
Vera'a	*veraa*	traditional	10	2:01:48	3,606
		collection totals	58	6:59:27	11,828

Table 1. The Multi-CAST corpora.

per corpus	format	file name
metadata sheet	PDF	*corpus_metadata.pdf*
annotation notes	PDF	*corpus_annotation-notes.pdf*
tabulated GRAID counts	PDF	*corpus_corpus-counts.pdf*
raw GRAID counts	CSV	*corpus_counts-raw.csv*
(related publications)	PDF	—

per text	format	file name
audio recording, *one per part*	WAV	*corpus_text.wav*
annotation file, *one per part*	EAF	*corpus_text.eaf*
transcription and translation	PDF	*corpus_text_transcription.pdf*
raw GRAID annotations	CSV	*corpus_text_graid-raw.csv*
raw GRAID counts	CSV	*corpus_text_counts-raw.csv*

Table 2. Text and supplementary files in the corpora.

alphabetical character *a*, *b*, *c*, … *z* (e.g. *english_london01a*, *english_london01b*, *english_london01c*, etc.), which carries forward into the labeling of each part's supplementary files.

Audio recordings. *One per text or text part.* The recorded texts, provided in the WAV file format. Longer recordings have been split into multiple parts.

Annotation files. *One per recorded text or text part.* The transcriptions and annotations, time-aligned with the audio recording. Uses the EAF file format, to be used in conjunction with the accompanying audio recording with the annotation software ELAN. The internal structure of the EAF files is described in detail in Section 3.1. A list of all annotated texts is provided in Appendix A.

Metadata sheets. *One per corpus.* Quick reference to the texts in each corpus, their associated files, the sociolinguistic background of the speakers, and the circumstances of the recordings. Also contains instructions on how to cite the corpus and its components.

Annotation notes. *One per corpus.* Descriptions of the pertinent analytical issues that surfaced during the annotation, and the annotators' decisions on how to handle them. The notes also contain a list of non-standard GRAID symbols employed in the annotation (see Section 3.2).

Tabulated GRAID counts. *One per corpus.* Tables with counts of the chief form-function GRAID glosses in each text and the entire corpus. Can be used as a quick reference on how common a particular GRAID symbol is in a particular text.

Transcriptions and translations. *One per recorded text.* The transcribed text in the object language, side-by-side with its English translation. Utterances are numbered continuously, aligning with the corresponding *utterance_id* in the annotated files. These files provide interested readers easy access to the content of the texts.

Raw GRAID annotations and text counts. *One per corpus, and one per recorded text.* CSV files (comma separated values) of (i) the GRAID annotations as they appear on the *graid* tier in the annotated EAF files (i.e. a vector of strings), and (ii) counts of all form-function GRAID symbols that occur in the annotations (i.e. a numerical table). These files are intended as quick start aids for users interested in exploring the data with text-mining or statistical analysis software.

3 Corpus design

Multi-CAST is a collection of annotated texts from narrative genres. It has been designed with the intent of facilitating cross-linguistic inquiries into discourse and related areas of research with a particular focus on quantitative approaches. To this end, great care was taken to make all Multi-CAST corpora adhere to the same uniform structure and design philosophy.

The texts in Multi-CAST are taken from a variety of text types, all of which comply with the following three key conditions: they are (i) overwhelmingly monologic, (ii) narrative in nature, and (iii) original (i.e. not translated). Texts with more than one speaker feature only negligible amounts of interlocution. The text types in Multi-CAST fall into one of three overarching groups: (i) traditional narratives, (ii) autobiographical narratives, and (iii) stimulus-based narratives:

Traditional narratives. Traditional stories and tolktales usually told to an audience of native speakers. Note that in typical language documentation settings, most narration events are in fact what Himmelmann (1998) calls 'staged communicative events', rather than truly incidental occurrences.

Autobiographical narratives. Oral history interviews on past events and the speaker's personal history, usually recorded in private settings. The texts selected for inclusion in Multi-CAST feature only minimal interviewer participation, allowing for long stretches of uninterrupted interviewee monologue.

Stimulus-based narratives. Narrative renditions of various short movie clips, among them the *Pear Film*, a six-minute short film without dialogue about a child stealing fruit (cf. Chafe 1980), as well as clips depicting everyday scenes from relevant cultural contexts.

The texts in Multi-CAST have been annotated across two levels, yielding a multi-tier structure that lends itself to a variety of complex research queries. In addition to basic morphological glossing, the texts feature annotations with the **GRAID annotation scheme** (Haig & Schnell 2014), which combine information on the form of a referring expression with information on its syntactic function and its semantics. The GRAID annotation scheme is described in Section 3.2.

The structure and internal organisation of the annotated texts is described in the next section.

3.1 EAF file structure

The annotated texts in Multi-CAST are provided as **EAF files**, and are intended to be used, alongside the accompanying audio recordings, with the **annotation software ELAN**.

ELAN (EUDICO Linguistic Annotator) is an open-source annotation tool developed at the Max Planck Institute for Psycholinguistics in Nijmegen.[7] It was written in the Java programming language, and is thus (technically) platform independent. ELAN stores annotation data in an XML-based file format with the extension **.eaf* (the EUDICO Annotation Format), and allows these files to be exported in a variety of other formats, including raw text, FLEx, and Toolbox files, thereby enabling some degree of interoperability between different annotation platforms. Please refer to the ELAN manual for details on how to operate the software.

Annotations in ELAN are organised across multiple **tiers**, which are hierarchically linked through different kinds of relationships (see below). The Multi-CAST EAF files contain a total of six core tiers, which include the audio-aligned transcription, various levels of annotation, and a free translation. Because each level of annotation is logically dependent on another, Multi-CAST uses a deep hierarchy, as is visualised in Figure 2.

ELAN employs so-called '**linguistic types**' to establish cardinal relationships between parent and daughter tiers. Multi-CAST annotations use three types of relation: (i) *time alignment* with the audio recording (called 'none' by default in ELAN), (ii) *symbolic subdivision* (for one-to-many relations), and (iii) *symbolic association* (for one-to-one relations). The root *utterance_id* tier (see below) is the only time-aligned tier, and the *grammatical_words* tier is the only one with a one-to-many relation (symbolic subdivision) to its parent; all other tiers have a one-to-one relation (symbolic association) to their respective superordinate.

In the following, the six basic tiers common to all Multi-CAST texts are described briefly; Table 3 provides a summary. The order in which the tiers are given here matches their sequence in the corpus files.

[7] `https://tla.mpi.nl/tools/tla-tools/elan/`

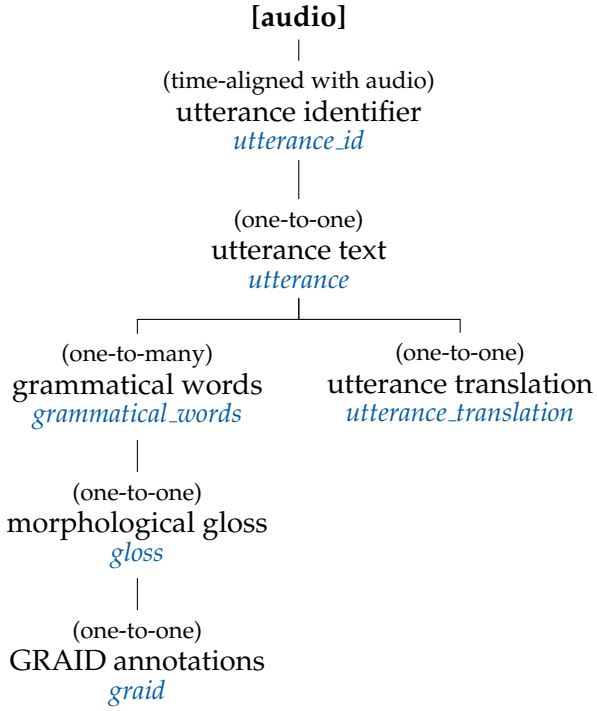

Figure 2. Multi-CAST EAF tier hierarchy.

tier	label	relational type
utterance identifier	*utterance_id*	*time alignment*
utterance text	*utterance*	*symbolic association* (1-to-1)
grammatical words	*grammatical_words*	*symbolic subdivision* (1-to-n)
morphological glosses	*gloss*	*symbolic association* (1-to-1)
GRAID annotations	*graid*	*symbolic association* (1-to-1)
utterance translation	*utterance_translation*	*symbolic association* (1-to-1)

Table 3. Multi-CAST EAF tier structure.

Crucially, GRAID includes annotations for unrealised elements (zeroes) and several meta-elements (clause boundary markers and operators) that are not strictly speaking contained in the original text. These annotations are represented on the *grammatical_words* tier and all of its dependents.

Utterance identifier. (*utterance_id*) The root tier to which all other tiers are subordinate. Time-aligned with the recording. Utterance identifiers are composed of the file name (in turn containing the corpus and text names) plus a numerical, three-digit counter. The identifier *veraa_jjq_075*, for example, is the label of the 75th sequential intonation unit ('utterance') of the *jjq* text from the *veraa* corpus. Utterance identifiers can be used to cite individual examples from Multi-CAST.

Utterance text. (*utterance*) The transcription of the recorded text. Daughter of *utterance_id* in a one-to-one relation. Utterance units generally contain at least one whole clause unit, generally more.

Grammatical words. (*grammatical_words*) The object language text, segmented into individual 'word' units. Daughter of *utterance* in a one-to-many relation. The segmentation into grammatical words forms the basis of all further annotations. At the annotator's discretion, clitics may be split off as separate annotation cells. Additionally, cells for zero elements and clause boundary markers are inserted on this and all dependent tiers. 'Word' in this context should be understood in terms of GRAID annotation units, see Section 3.2 below.

Morphological glosses. (*gloss*) Morphological glosses of the word units. Daughter of *grammatical_words* in a one-to-one relation. The morphological glossing provides for the identification of inflectional morphology, using conventionalised labels recommended by the *Leipzig Glossing Rules*.[8] Note that the level of morphological detail and the extent to which individual forms have been morphologically segmented varies from corpus to corpus.

GRAID annotations. (*graid*) Morphosyntactic annotations using the GRAID annotation scheme (Haig & Schnell 2014). Daughter of *glosses* in a one-to-one relation. GRAID glosses combine information on the form of a referential expression, in particular major clause constituents, with information on its syntactic function and its semantics. In addition to overt forms, GRAID also analyses non-expressed ('zero') arguments, for which the placeholder <0> is inserted on all levels subordinate to the *grammatical_words* tier. Clause boundaries are marked with the hash <#> and percentage <%> signs.

[8] `http://www.eva.mpg.de/lingua/resources/glossing-rules.php`

A brief rundown of the system is given in Section 3.2 below. Please refer to the *GRAID manual* (Haig & Schnell 2014), available from the archive webpage, for an extensive description.

Utterance translation. (*utterance_translation*) An English translation of the primary text. Daughter of *utterance* in a one-to-one relation.

At the annotator's discretion, texts may include other tiers in addition to this core set, for instance the primary text in a non-transliterated orthographic system, or comments on the annotation or translation. The name of these tiers is preceded by *add_*, yielding, for example, *add_comments* or *add_orthography*. Their position in the tier hierarchy is dependent on their content and function.

3.2 GRAID annotations

The GRAID (*Grammatical Relations and Animacy in Discourse*) annotation scheme (Haig & Schnell 2014) was specifically designed to facilitate quantitative cross-linguistic research of discourse structure and referentiality, of the type pioneered in Du Bois (1987), Du Bois et al. (2003), Bickel (2003), Stoll & Bickel (2009), and Noonan (2003), among others.

The chief tenet of GRAID is consistency: quantitative cross-linguistic investigations are only possible when annotations in different corpora use the same inventory of symbols in the same way. GRAID uses a small core set of approximately 30 symbols for glossing the form, grammatical relations, as well as animacy features of major clause constituents in a text, and provides simple conventions for combining them. GRAID symbols aim for a level of granularity similar in spirit to that of 'comparative concepts' in Haspelmath (2010). The system was built with flexibility in mind, allowing both for different levels of detail in the glossing of different items while maintaining cross-corpus comparability, and for the limited inclusion of additional symbols specific to a particular language or research approach.

The annotator's decisions on how to handle pertinent issues of annotation and a list of the non-standard GRAID symbols employed are provided in the annotation notes for each corpus.

GRAID glosses are aligned with single words, but target entire referential expressions and their functions (i.e. phrases). Glosses couple an annotation of *form* (e.g. <pro> for 'full pronoun'), which may include an *animacy feature* (e.g. <h> 'human'), with a *function* (e.g. <s> 'subject of an intransitive clause'). Animacy features are linked to form glosses via a full stop <.> and functions via a colon <:>, yielding, in this example, the first constituent of (1):

(1) ## *he* *was* *leaving* *then*
pro.h:s aux v:pred other
(= full pronoun, human, in S function)

As the main target of GRAID annotations is the realisation of referential expressions, the glossing of predicates is comparatively coarse-grained.

Unexpressed arguments, when licenced by a predicate expression, are noted in GRAID annotations via the gloss <0> (i.e. the digit zero). To the extent to which it is possible, unexpressed arguments are aligned with whatever position their overt 'counterpart' would slot into in a pragmatically neutral clause. Note that zero glosses receive their own annotation cells in the EAF texts, and are therefore represented on all annotation tiers:

(2)

##	*he*	*went*	*into*	*town*	##	*and*	*0*	*bought*	*a*	*donkey*
##	pro.h:s	v:pred	adp	np:g	##	other	0.h:a	v:pred	ln	np:p

(= unexpressed argument, human, in A function)

See Section 2.2 of the *GRAID manual* (Haig & Schnell 2014: 8) for further notes on the annotation of zero arguments.

The basic unit of glossing is the clause, defined as all constituents associated with a particular clause. As seen in the above examples, GRAID signals the left-hand boundary of syntactically independent main clauses with double hashes <##> and that of syntactically dependent clauses with a single hash <#>. The right-hand boundary of centre-embedded clauses is marked with a percentage sign <%>.

Users of Multi-CAST are advised to refer to the *GRAID manual* (Haig & Schnell 2014), currently in its version 7.0, for an in-depth description of the GRAID annotation system, its motivations, and a full inventory of its symbols. The *manual* is available, alongside other descriptive material and related publications, on the archive webpage.[9]

[9] `https://lac.uni-koeln.de/multicast/`

References

Adibifar, Shirin. 2016. Persian. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts)*. (`https://lac.uni-koeln.de/multicast-persian/`) (accessed 2016-07-01.) [p. 4]

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge. [p. 3]

Ariel, Mira. 2004. Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse Processes* 37(2). 91–116. [p. 3]

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736. [pp. 3, 12]

Brickell, Timothy C. 2016. Tondano. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts)*. (`https://lac.uni-koeln.de/multicast-tondano/`) (accessed 2016-07-01.) [p. 5]

Chafe, Wallace (ed.). 1980. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex. [p. 8]

Du Bois, John. 1987. The discourse basis of ergativity. *Language* 63(4). 805–855. [pp. 3, 12]

Du Bois, John. 2003. Argument structure: Grammar in use. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure*, 11–60. Amsterdam: John Benjamins. [p. 3]

Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.). 2003. *Preferred argument structure: Grammar as architecture for function*. Amsterdam: John Benjamins. [p. 12]

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (version 7.0)*. (`https://lac.uni-koeln.de/en/multicast/`) (accessed 2015-12-30.) [pp. 9, 11, 12, 13]

Haig, Geoffrey & Schnell, Stefan (eds.). 2016a. *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts)*. (`https://lac.uni-koeln.de/multicast/`) (accessed 2016-02-08.)

Haig, Geoffrey & Schnell, Stefan. 2016b. Multi-CAST research context. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts)*. [p. 3]

Haig, Geoffrey & Schnell, Stefan & Wegener, Claudia. 2011. Comparing corpora from endangered languages: Explorations in language typology based on original texts. In Haig, Geoffrey & Nau, Nicole & Schnell, Stefan & Wegener, Claudia (eds.), *Documenting endangered languages*, 55–86. Berlin: Mouton de Gruyter. [p. 5]

Haig, Geoffrey & Thiele, Hanna. 2016. Northern Kurdish. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST (Multilingual Corpus of Annotated Spoken*

Texts). (https://lac.uni-koeln.de/multicast-northern-kurdish/) (accessed 2016-02-22.) [p. 4]

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(4). 663–687. [p. 12]

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(2). 161–195. [p. 8]

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. [p. 4]

Mosel, Ulrike & Schnell, Stefan. 2016. Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts)*. (https://lac.uni-koeln.de/multicast-teop/) (accessed 2016-02-22.) [p. 5]

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.). (Unpublished manuscript.) [pp. 3, 12]

Schiborr, Nils N. 2016. English. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts)*. (https://lac.uni-koeln.de/multicast-english/) (accessed 2016-06-31.) [p. 4]

Schnell, Stefan. 2016. Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts)*. (https://lac.uni-koeln.de/multicast-veraa/) (accessed 2016-02-22.) [p. 5]

Stoll, Sabine & Bickel, Balthasar. 2009. How deep are differences in referential density? In Guo, Jiansheng & Lieven, Elena & Budwig, Nancy & Ervin-Tripp, Susan & Nakamura, Keiko & Özçaliskan, Seyda (eds.), *Crosslinguistic approaches to the psychology of language*, 543–555. London: Psychology Press. [p. 12]

Vollmer, Maria C. & Hadjidas, Harris. 2016. Cypriot Greek. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts)*. (https://lac.uni-koeln.de/multicast-cypriot-greek/) (accessed 2016-02-22.) [p. 4]

Appendices

A List of corpus files

file name	speaker	date rec'd	text type	length h:mm:ss	clause units
cypgreek_jitros	CG01	1960	traditional	*—	272
cypgreek_minaes	CG01	1960	traditional	*—	359
cypgreek_psarin	CG01	1964	traditional	*—	440
english_kent01	EN01	1975	autobiogr.	27:28	622
english_kent02a	EN01	1975	autobiogr.	30:00	764
english_kent02b	EN01	1975	autobiogr.	32:41	859
nkurd_muserz01	NK01	2000	traditional	19:53	619
nkurd_muserz02	NK02	2002	traditional	12:12	482
persian_g1-f-01	PS01	2015	stim.-based	1:34	47
persian_g1-f-02	PS02	2015	stim.-based	2:10	54
persian_g1-f-05	PS05	2015	stim.-based	2:15	60
persian_g1-f-07	PS07	2015	stim.-based	1:05	38
persian_g1-f-08	PS08	2015	stim.-based	1:40	25
persian_g1-f-09	PS09	2015	stim.-based	4:29	100
persian_g1-f-10	PS10	2015	stim.-based	3:19	83
persian_g1-f-11	PS11	2015	stim.-based	1:42	60
persian_g1-f-12	PS12	2015	stim.-based	1:45	49
persian_g1-f-14	PS14	2015	stim.-based	3:03	99
persian_g1-m-03	PS03	2015	stim.-based	0:45	17
persian_g1-m-04	PS04	2015	stim.-based	2:03	61
persian_g1-m-06	PS06	2015	stim.-based	0:51	22
persian_g1-m-13	PS13	2015	stim.-based	2:50	69
persian_g2-f-01	PS15	2015	stim.-based	2:23	58
persian_g2-f-02	PS16	2015	stim.-based	1:27	44
persian_g2-f-03	PS17	2015	stim.-based	1:37	40
persian_g2-f-04	PS18	2015	stim.-based	1:03	25
persian_g2-f-05	PS19	2015	stim.-based	1:52	26
persian_g2-f-06	PS20	2015	stim.-based	1:27	56
persian_g2-f-07	PS21	2015	stim.-based	1:41	51
persian_g2-m-08	PS22	2015	stim.-based	1:44	49
persian_g2-m-09	PS23	2015	stim.-based	1:20	42
persian_g2-m-10	PS24	2015	stim.-based	1:17	41
persian_g2-m-11	PS25	2015	stim.-based	1:01	25
persian_g2-m-12	PS26	2015	stim.-based	1:08	40
persian_g2-m-13	PS27	2015	stim.-based	1:23	52
persian_g2-m-14	PS28	2015	stim.-based	1:03	36
persian_g2-m-15	PS29	2015	stim.-based	2:35	48
teop_iar	TP01	2003	traditional	14:34	348
teop_mat	TP02	2004	traditional	6:54	210
teop_sii	TP03	2004	traditional	19:21	587

table continued on next page

➠ table continued from previous page

file name	speaker	date rec'd	text type	length h:mm:ss	clause units
teop_viv	TP04	2004	traditional	5:46	157
tondano_gulamera	TD01	2011	stim.-based	10:15	129
tondano_holiday	TD01	2011	autobiogr.	5:16	89
tondano_kiniar01	TD02	2013	stim.-based	8:50	143
tondano_kiniar02	TD03	2013	stim.-based	12:36	193
tondano_kiniar03	TD03	2013	stim.-based	8:46	99
tondano_mapalus	TD04	2011	autobiogr.	6:51	150
tondano_water	TD05	2011	stim.-based	5:04	80
tondano_watulaney	TD06	2011	autobiogr.	18:20	203
veraa_anv	VR01	2007	traditional	6:07	185
veraa_as1	VR02	2007	traditional	5:16	213
veraa_gabg	VR03	2007	traditional	8:41	174
veraa_gaqg	VR04	2007	traditional	8:52	226
veraa_hhak	VR05	2007	traditional	12:39	431
veraa_isam	VR06	2007	traditional	7:21	238
veraa_iswm	VR07	2007	traditional	21:43	576
veraa_jjq	VR08	2007	traditional	30:19	879
veraa_mvbw	VR09	2007	traditional	10:07	305
veraa_palaa	VR10	2007	traditional	4:02	140
veraa_palab	VR10	2007	traditional	6:41	239

Table A. List of Multi-CAST corpus files.
*Note: no audio files are available for the Cypriot Greek data.

B List of corpus speakers

speaker	corpus	gender	age	date born	date rec'd	clause units
CG01	*cypgreek*	female	73 77	1887	1960 1964	1,071
EN01	*english*	male	85	1890	1975	2,245
NK01	*nkurd*	male	~50	~1950	2000	619
NK02	*nkurd*	female	~60	~1940	2002	482
PS01	*persian*	female	39	1976	2015	47
PS02	*persian*	female	29	1986	2015	54
PS03	*persian*	male	22	1993	2015	17
PS04	*persian*	male	25	1990	2015	61
PS05	*persian*	female	26	1989	2015	60
PS06	*persian*	male	32	1983	2015	22
PS07	*persian*	female	25	1990	2015	38
PS08	*persian*	female	25	1990	2015	25
PS09	*persian*	female	25	1990	2015	100
PS10	*persian*	female	31	1984	2015	83
PS11	*persian*	female	33	1982	2015	60
PS12	*persian*	female	33	1982	2015	49
PS13	*persian*	male	35	1980	2015	69
PS14	*persian*	female	29	1986	2015	99
PS15	*persian*	female	20	1995	2015	58
PS16	*persian*	female	20	1995	2015	44
PS17	*persian*	female	20	1995	2015	40
PS18	*persian*	female	20	1995	2015	25
PS19	*persian*	female	21	1995	2015	26
PS20	*persian*	female	38	1977	2015	56
PS21	*persian*	female	33	1982	2015	51
PS22	*persian*	male	20	1995	2015	49
PS23	*persian*	male	22	1993	2015	42
PS24	*persian*	male	20	1995	2015	41
PS25	*persian*	male	25	1990	2015	25
PS26	*persian*	male	20	1995	2015	40
PS27	*persian*	male	20	1995	2015	52
PS28	*persian*	male	20	1995	2015	36
PS29	*persian*	male	27	1988	2015	48
TP01	*teop*	female	~70	~1930	2003	348
TP02	*teop*	female	~30	~1970	2004	210
TP03	*teop*	female	~60	~1940	2004	587
TP04	*teop*	female	~30	~1970	2004	157
TD01	*tondano*	female	+50	~1960	2011	218
TD02	*tondano*	male	~40	~1970	2013	143
TD03	*tondano*	male	+50	~1960	2013	292
TD04	*tondano*	female	+50	~1960	2011	150

table continued on next page ➠

➠ *table continued from previous page*

speaker	corpus	gender	age	date born	date rec'd	clause units
TD05	*tondano*	female	~40	~1970	2011	80
TD06	*tondano*	female	~40	~1970	2011	203
VR01	*veraa*	female	~20	~1985	2007	185
VR02	*veraa*	male	~40	~1965	2007	213
VR03	*veraa*	male	~40	~1965	2007	174
VR04	*veraa*	male	~40	~1965	2007	226
VR05	*veraa*	male	~20	~1985	2007	431
VR06	*veraa*	male	~60	~1950	2007	238
VR07	*veraa*	male	~60	~1950	2007	576
VR08	*veraa*	male	~60	~1950	2007	879
VR09	*veraa*	male	~30	~1975	2007	305
VR10	*veraa*	female	~40	~1965	2007	379

Table B. List of Multi-CAST speakers.

Corpus timeline

date	changes	affected texts
2016-06	added 37 corpus texts	*persian_** *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_** *gulamera*, *holiday*, *kiniar01*, *kiniar02*, *kiniar03*, *mapalus*, *water*, *watulaney*
2015-05	added 21 corpus texts	*cypgreek_** *jitros*, *minaes*, *psarin*; *english_** *kent01*, *kent02a–b*; *nkurd_** *muserz01*, *muserz02*; *teop_** *iar*, *mat*, *sii*, *viv*; *veraa_** *anv*, *as1*, *gabg*, *gaqg*, *hhak*, *isam*, *iswm*, *jjq*, *mvbw*, *palaa–b*