Documentation of the MAGRAM Grammaticalization Database

This document (i) describes the history and development of the database (introduction to the project, its methodology, contributors, workflow), (ii) elaborates on its design (information on what can be found and what cannot be found, where it can be found, and how it is coded) and (iii) some further general information (credits, how to cite, license, contact).

We have conceived the database structure similar to the ValPal database.

I History and development of the project

1 Research agenda

MAGRAM is the acronym of the MAinz GRAMmaticalization project, a project funded by the German Research Foundation (DFG; under BI 591/12–1) at the Johannes-Gutenberg Universität Mainz under the lead of Prof. Dr. Walter Bisang and apl. Prof. Dr. Andrej Malchukov from January 2016 – March 2020.

MAGRAM started out from the observation that grammaticalization is not necessarily cross-linguistically homogeneous (cf. Bisang 2011). For that reason it tried to see to what extent there could actually be areal and/or cross-phylogenetic variation along the following two hypotheses:

(i) Grammaticalization paths of the type [SOURCE → TARGET] show areal/genetic variation in terms of the sources that undergo grammaticalization as well as in terms of targets linked to a specific concept.

(ii) There are cross-linguistic differences in the degree of covariation of meaning and form, in the sense that semantic changes do not necessarily entail changes in form-related parameters.

In addition to the present database, the major outcome of the project is the publication of the two-volume Comparative Handbook of Grammaticalization Scenarios (Bisang, Walter & Malchukov, Andrej L. (eds.), 2020. *Grammaticalization Scenarios: Cross-linguistic Variation and Universal Tendencies*, Volume 1: *Grammaticalization Scenarios from Europe and Asia*, Volume 2: *Grammaticalizations Scenarios from Africa, the Americas, and the Pacific*. Berlin: Mouton De Gruyter. 2020), which included the position paper by the MAGRAM team (Bisang et al. 2020a), the Questionnaire (Bisang et al. 2020b), as well as 25 in-depth studies of grammaticalization scenarios in individual languages, families or areas based on the questionnaire. The questionnaire (Bisang et al. 2020b) is presented as part of the database manual below.

2 Methodology

In this section we provide a brief explanation (described in more detail in the questionnaire) on how grammaticalization paths have been recorded for the database.

2.1 Basics

Each grammaticalization path is described by its source and its target. Both are described by their form and function/meaning, as is also customary in schematic representations of grammaticalization clines. A famous example would be:

(1) *willan* 'want' > *will* future tense (cf. Kuteva et al. 2019: 453, among others)

In this project, we have not systematically recorded all intermediate steps, but tried to record the earliest, most lexical meaning and the most grammaticalized function and the respective forms they take on.

2.2 Data

We extracted grammaticalization paths on the one hand from the chapters of the above-mentioned handbook, and on the other hand from 30-source-lists (see subset) contributed by the same authors (cf. workflow for more detail). The data were extracted in close contact with the contributors.

2.3 Parameters

Our selection of parameters for measuring grammaticalization is based on Lehmann's (1995) framework with its six parameters belonging to the three domains of weight, cohesion and variability with their paradigmatic and syntagmatic levels. We modified this framework in the following way:

Table 1: Modified list of Lehmann's (1995) parameters.

	Paradigmatic	Syntagmatic
Weight	1. Semantic Integrity (SI) 2. Phonetic Reduction (PR) (= Phonetic Integrity)	~~Structural Scope~~
Cohesion	3. Paradigmaticity (PM)	4. Bondedness (BD)
Variability	5. Paradigmatic Variability (PV)	6. Syntagmatic Variability (SV)

Paradigmatic Weight was split into the two parameters of Semantic Integrity (SI) and Phonetic Reduction (PR) for being able to check for correlations between function and form. We did not include Structural Scope because of its controversial theoretical and empirical status. The remaining four parameters were adopted.

We added the parameters of Decategorization (DC; extent to which the categorial properties of the source concept are maintained) and Allomorphy (AM) on the basis of further literature (cf. questionnaire).

2.4 Evaluation

For each parameter, we defined values 1, 2, 3 and 4 , from lowest to highest degree of grammaticalization.

Each target in a source-target combination receives a value for every of the 8 parameters. The following scale from the questionnaire (Bisang et al. 2020b: 93) shows the four values for the parameter of Bondedness (BD) as an example:

1 The linguistic sign is a free morpheme or is the lexical root of a word.
2 The linguistic sign is a clitic (its use is not limited to a single word class).

3 The linguistic sign is an agglutinative affix (affixed to individual words which are members of the same word class).
4 The linguistic sign is part of a *porte-manteau* morpheme or is a suprasegmental (e.g., tonal marker) or a process morpheme (*Ablaut*, ...), or a zero morpheme.

The source values were not assigned any values because in quite a number of cases it was not possible to reconstruct the details of the source properties. Since it was still possible to evaluate if there was a change in parameter value from source to target, we evaluated for every source-target combination whether or not there was a change in each parameter.

Thus, we collected two types of data for each path of grammaticalization:

- Change data: change (1) / no change (0) in value from source to target (per parameter)
- Value data: target value of 1, 2, 3 or 4 (per parameter).

The values for individual parameters and more remarks on the topic can be found in the Questionnaire .

3 Team & contributors

3.1 Team

The team was led by Walter Bisang and Andrej Malchukov. The core team of the MAGRAM project involved in data annotation and evaluation: Linlin Sun (LS), Iris Rieder (IR), Eduard Schroeder (ES) and Marvin Martiny (MM). Statistical evaluation for publications was performed by Svenja Luell (SL), and at later stages, by Laura Becker (LB). Since 2020, database management and curation has been performed by Marvin Martiny.

3.2 Contributors

The data of the Mainz Grammaticalization project is based on the contributions of 29 experts in the respective languages, among them Christian Lehmann as one of the founding fathers of grammaticalization research. The contributors to our database are listed in the alphabetical order of their languages/families of expertise:

Beja (Cushitic, Afroasiatic): Martine Vanhove

Chinese (Sinitic, Sino-Tibetan): Linlin Sun and Walter Bisang

Creoles and Pidgins: Susanne Michaelis and Martin Haspelmath

Emai (Edoid, Niger-Congo): Ronald P. Schaefer and Francis O. Egbokhare

German (Indo-European): Luise Kempf and Damaris Nübling

Hoocak (Core Siouan): Johannes Helmbrecht

Indo-Aryan (Indo-European): Annie Montaut

Iranian (Indo-European): Agnes Korn

Iroquoian: Marianne Mithun

Japhug (Rgyalrong, Sino-Tibetan): Guillaume Jacques

Khmer (Austroasiatic): Walter Bisang

Korean: Seongha Rhee

Lezgic (Northeast Caucasian): Timur Maisak

Malayo-Polynesian (Austronesian) & Mori (Yareban): Nikolaus P. Himmelmann

Manding (Mande, Niger-Congo): Denis Creissels

Mian (Papua New Guinea): Sebastian Fedden

Nyulnyulan (Non-Pamanyungan, Australian): William B. McGregor

Quechua and Aymara: Willem F. H. Adelaar

Romance (Indo-European): Michela Cennamo

Slavic (Indo-European): Björn Wiemer

Southern Uto-Aztecan: Zarina Estrada-Fernández

Thai: Walter Bisang

Tswana (Bantu, Niger-Congo): Denis Creissels

Tungusic (Manchu-Tungusic, Transeurasian): Andrej Malchukov

Uralic: Juha Janhunen

Yeniseian: Edward Vajda

Yucatecan (Mayan): Christian Lehmann

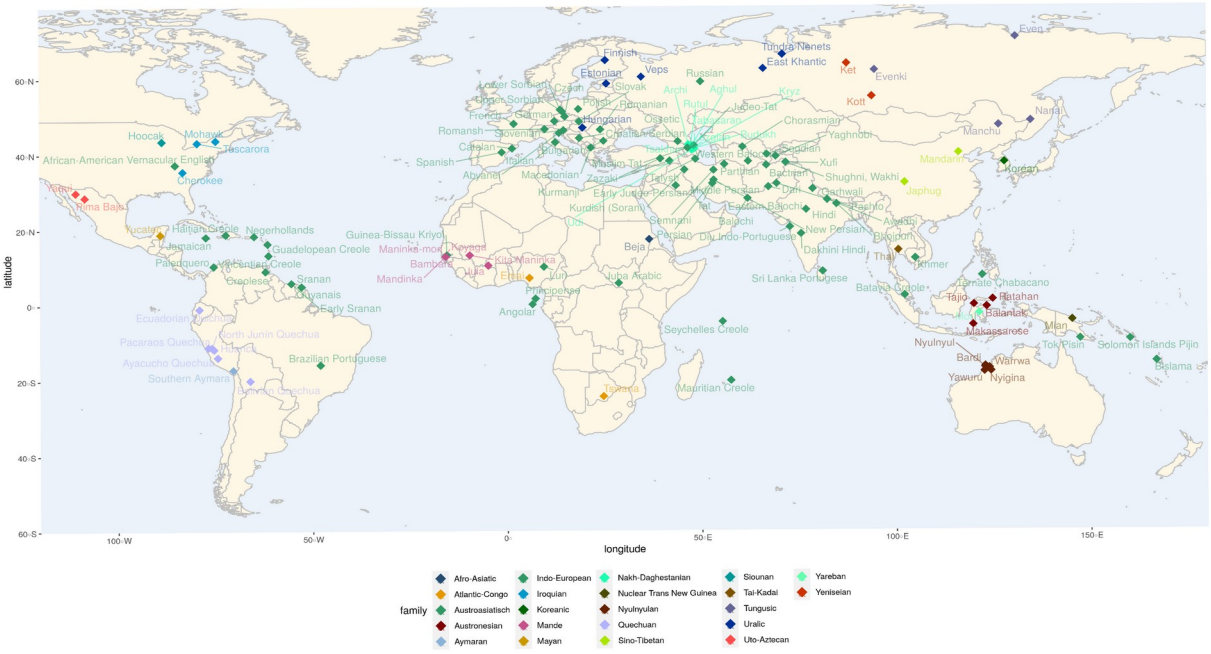

Figure 1: Languages in the dataset with their geographical location and language family (adopted from Bisang et al. forthcoming).

4 Workflow

This is a schematic overview of the steps which led to the creation of the current database.

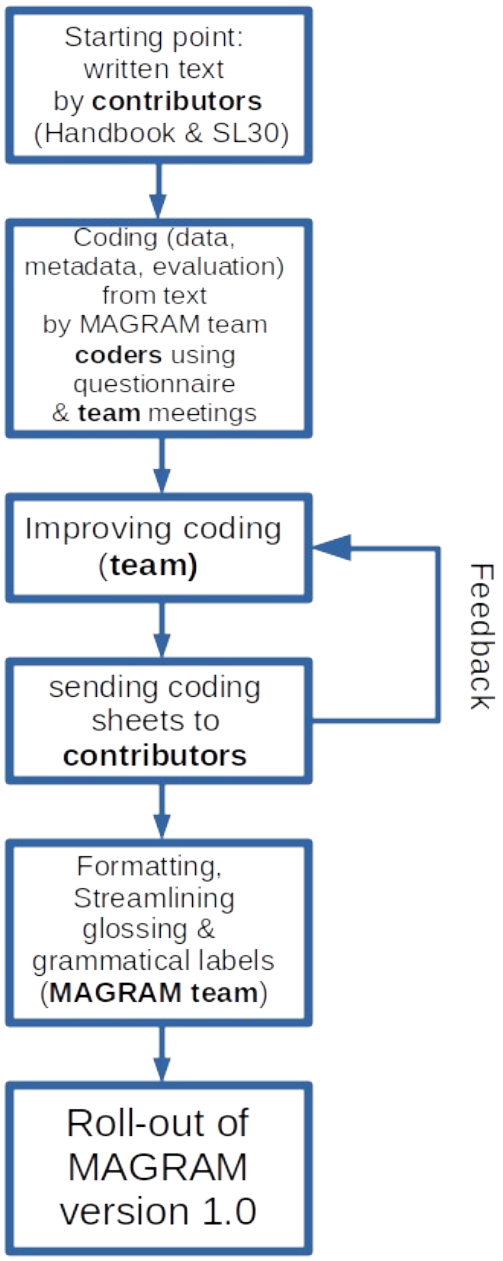

Figure 2: Schematic overview of the project's workflow.

II Database structure

2.1 Design principles – overview

There are basically three types of data within the database, subdivided into a total of six subtypes. They are here listed from left to right in the data sheet:

1 Metadata

1.a ***path specific metadata*** (Code (= path ID), Subset, Data source and Author(s))

1.b ***language specific metadata*** (Macro-area (Dryer), Macro-area (Hammarström/Donohue), Family, Glottocode, Genus, Language, Level of reconstruction)

Code	Subset	Data source	Author(s)
009001	other	Handbook	Malchukov
009002	other	Handbook	Malchukov
009003	30 Sources	Handbook	Malchukov
009004	30 Sources	Handbook	Malchukov

Figure 3: process specific metadata.

Macro-area (Dryer)	Macro-area (Hammarström /Donohue)	Family	Glottocode	Genus	Language	Level of reconstruction (if applicable)
Eurasia	Eurasia	Tungusic	even1260	Tungusic	Even	
Eurasia	Eurasia	Tungusic	even1260	Tungusic	Even	
Eurasia	Eurasia	Tungusic	even1260	Tungusic	Even	
Eurasia	Eurasia	Tungusic	even1260	Tungusic	Even	

Figure 4: language specific metadata.

2 Path description

2.a ***path description proper*** (Source form, Source meaning, Source meaning (simplified), Target form, Target meaning, Target meaning (simplified))

2.b ***example*** (Example (material), Example (glossing), Example (translation), Example (reference))

2.c ***comments***

Source form	Source meaning	Source meaning (simplified)	Target form	Target meaning	Target meaning (simplified)
doo	'inside'	'inside'	*-du*	DATIVE-	DATIVE

				LOCATIVE	
bi	personal PRONOUN (here: 1SG)	PERSONAL PRON	*-w,-u,-bu*	POSSESSIVE AGREEMENT (here: 1SG)	AGREEMENT
bu-	'give'	'give'	*-w-, -m-*	PASSIVE voice	PASSIVE

Figure 5: path description proper.

Example (material)	Example (glossing)	Example (translation)	Example (reference)	Comments
d'uu doo-la-n	house inside-LOC-3SG.POS	'in the house'	Malchukov (personal communication)	
d'uu-w	house-1SG.POS	'my house'	Malchukov (personal communication)	
maa-w-ra-n	kill-PASS-AOR-3SG	'(he) was killed'	Malchukov (personal communication)	
it-ne-n	see-VENT-AOR.3SG	'he went to see'	Malchukov 2020: 411	

Figure 6: exemplification & comments.

3 **Path evaluation** (parameter values: target level value, change value; parameters defined below)

Semantic integrity	Phonetic reduction	Paradig-maticity	Bondedness	Paradigmatic variability	Syntagmatic variability	Decategoriz ation	Allomorphy
4	2	4	3	4	4	4	2
4	3	4	3	2	4	4	3
3	3	3	3	2	4	4	2
3	2	3	3	1	4	4	2

Figure 7: path evaluation - values.

Semantic integrity	Phonetic reduction	Paradig-maticity	Bondedness	Paradigmatic variability	Syntagmatic variability	Decategoriz ation	Allomorphy
1	0	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	0	1	1	1

Figure 8: path evaluation - change.

2.2 Design principles – details

2.2.1 Code

The code uniquely identifies the respective grammaticalization path. It codes the language group (not genealogically, but in terms of data collection), as well as the number of the path within this set. Of course, each ID should code for exactly one path, and it will stay the same in case the database is expanded or changed.

2.2.2 Subset

At the moment, there are two subsets, which will be coded: *SL30* vs. *other sources*.

SL30 ('source list 30') means that the source concept belongs to a group of 30 (mostly lexical) source concepts, which are defined in the ***questionnaire*** (Bisang et al. 2020b) .

These data were gathered by asking the contributors to look for instances of grammaticalization with these source concepts in their respective languages.

Table 2: 30 source concepts (Bisang et al. 2020b: 97-101; with reference to Heine and Kuteva 2002).

No	*Source Concept*
1.	ARRIVE
2.	BACK (body part)
3.	BODY
4.	CHILD
5.	COME
6.	COPULA
7.	DEMONSTRATIVE
8.	DO
9.	FALL
10.	FINISH
11.	FOLLOW
12.	GET
13.	GIVE
14.	GO
15.	HAND (body part)
16.	HEAD (body part)
17.	HERE
18.	LEAVE
19.	LIVE
20.	LOVE
21.	MAN
22.	ONE
23.	SAY
24.	SEE
25.	SIDE
26.	SIT
27.	TAKE

28.	THING
29.	WANT
30.	WOMAN

Thus, by searching only for SL30 items there is a possibility to browse a comparative list (even though by the nature of grammaticalization, it may not be exhaustive: the absence of a certain source in the database for a given language does not warrant its non-existence).

2.2.3 Data source
This indicates, whether or not the path is explicitly described in the Handbook (Bisang & Malchukov 2020, eds.).

In case the path can be found in the Handbook, some issues explained in the respective chapter might not be commented on explicitly in the database.

2.2.4 Author(s)
In this column, the researcher(s) who provided the information on the path are mentioned with their last name. Their full name is given in the list of contributors**.**

2.2.5 Macro-area (Dryer)
The division into 6 Macro-Areas (Africa, Eurasia [excluding southeast Asia], Southeast Asia & Oceania, Australia-New Guinea, North America and South America) as proposed by Dryer (1992). This distribution is widespread, and e.g. was used in the original version of WALS (cf. Hammarström & Donohue 2014: 3).

Dryer (1992: 84-85) mentions that the choice of areas and their boundaries was somewhat arbitrary, and that he does not claim those areas to be linguistic areas. His goal was to have areas that appear roughly comparable in genetic and typological diversity.

Note that there is some 'leakage' where areas flow into each other (Dryer 1992: 83):

- Africa includes the Semitic languages of southwest Asia
- Southeast Asia is defined by language families (Sino-Tibetan, Thai, and Mon-Khmer)
- Australia & New Guinea excludes Austronesian languages of New Guinea
- North America includes Mayan and Aztecan languages in Central America, whereas South America encompasses languages in Central America except Mayan and Aztecan languages.

2.2.6 Macro-area (Hammarström/Donohue)
The second geographical pattern used for areal classification is the one presented by Hammarström & Donohue (2014):

Africa
Australia
Eurasia
Multinesia
North America
South America

Hammarström & Donohue (2014) define macro-areas as areas of roughly continental size. The same distribution is used in Glottolog (4.5, Hammarström, Forkel & Haspelmath 2022; only with the label 'Papunesia' instead of 'Multinesia'). The authors state (Glottolog 4.5, Hammarström, Forkel & Haspelmath 2022; https://glottolog.org/meta/glossary#macroarea, accessed Jan 10th 2022, 11:40CET):

"The division of the inhabited landmass into the macro-areas defined here is optimal in the following sense. It is the division

1 into 6 areas,
2 for which there are at least 250 languages in each area, such that
3 the distance between the component parts inside each area is minimized, and
4 the length of intersections between pairs of macro-areas is minimized."

5

2.2.7 Family
Here we give the top-level linguistic family, following Glottolog. Thus, it is based on the same comparative evidence and the same methodology applies.

2.2.8 Glottocode
The ***glottocode*** of the path's languoid, a unique and stable identifier, under which it can be found in Glottolog.

We always chose the most basic (dialect, language) variety with a glottocode, to be as specific as possible. Note: it is always the latest (~modern) variety in case there are glottocodes for historical varieties of the langoid.

2.2.10 Genus
Our definition of genus gives only information about data structure, since it is defined as the taxonomically lowest ranking group of languages which still entails all the related languages (per chapter).

An example would be the dataset of Romance: It includes paths from Spanish, Italian, French, Catalan, and one path each from Romanian, Romansh and Brazilian Portugues. That includes languoids from the Western and Eastern branches of Romance, but none from the Southern branch. We still use the label Romance, since there is no taxonomical level that combines Western and Eastern without including Southern Romance (cf. Agard 1984). If the path from Romanian was not there, no path from Eastern Romance would be present; in that case, we would probably speak of Western Romance or Italo-Western Romance more specifically.

2.2.11 Language
The languoid of the path as indicated by the contributing researcher, under the name used in the respective literature.

2.2.12 Level of reconstruction (if applicable)
In case the source is not an attested one, but a reconstructed source, we indicated the level of reconstruction (conservatively estimated) in this column.

Note: In case of the Iranian language data, the language (stage/variety) of the source construction was given even if it was an attested language.

2.2.13 Source form

1 **Material only**: In those columns, only **language material** is displayed. That means that there is no functional description of anything, be it nucleus or periphery in this column. **Note**: The single exception may be the labels <N>, <V>, and <A> (for indicating the basic lexical parts of speech, noun, verb and adjective) in cases where there is no other possibility of representation.

2 **One form (citation form)**: Most of the time, **only one form is given**, even if the source is inflectable. Ideally this form is a stem (or a language-specific citation form). If only one form is grammaticalized, there is no special marking either. If a whole paradigm or group of morphemes is grammaticalized, which cannot be shown separately (e.g., an unsegmentable subjunctive), we mostly display only one form and make a note like "(here: 1SG.SUBJ)" in Source meaning.

3 **Allomorphs**: Allomorphs may be coded here, **separated by commas**. In case of a zero-alternation (round) brackets are used instead, as is common.

4 **Phonological status**: The phonological status (clitic, bound, free) is indicated by using the **equal sign** "=", **minus** "-" or **space** " " respectively.

5 **Coding Periphery vs**. **Nucleus**: Peripheral parts of constructions will be given in **brackets**, and can be differentiated from zero-alternations like "-z(a)" by the use of a **plus (+) in the brackets for unbound** peripheral items, and a **minus (-) or equal sign (=) for bound** peripheral items. To put it differently: if in round brackets there is neither a plus (+) nor a minus (-) it is a zero-alternation allomorphy. If there is, it is a

peripheral morpheme. An example for the use of both comes from the language Beja (Vanhove & MAGRAM editorial team 2022):

Table 3: Source form and Source meaning including uses of plus and minus.

Source form	**Source meaning**
(o:n +) mari(-i)	(proximal DEMONSTRATIVE +) 'direction' (-GEN)

6 **Phonetic representation:** The phonetic representation may occasionally be given **in square brackets** additionally.

7 **Reconstruction:** Reconstructed forms are marked by an **asterisk < * >**. If they are particularly uncertain, we mark that by **(?)**.

8 **Zero morphemes:** zero is coded by the character **< Ø >**.

9 **(Non-latin) orthographic representation**: in a few cases, the native orthographic representation in the respective writing system is given in **angle brackets** (**< >**).

10 **Syntactically bound but phonologically unbound morphs**: in a few cases (mostly from the Korean dataset), there are syntactically bound morphs, which are phonologically unbound. This is represented by a **dot (< . >)** instead of a space character between two morphs.

2.2.14 Source meaning

1 **Function only**: In contrast to Source form and Target form, here only functions are represented, not language material.

2 **One function**: Only **one** source **function** is given here. If there are more, which are important to mention, this may be done in the comment section.

3 **Corresponding entries**: For complex constructions, **every item here should correspond to an item in** Source form , respectively. Peripheral items are marked equally as in **Source form**.

4 **Primordial source**: For the description of the source concept, always the 'primordial', that is, earliest attested or reconstructed concept and its expression, have been recorded.

5 **Formatting:** This field was more of a free text field for contributor and coder, and therefore we find some variation in how some functions are expressed. However, we tried to stick to the basic formatting: (core) **grammatical** labels in CAPS, **lexical** meaning in 'quotation marks', everything else lower case. Also lower case for additional higher-level description (grammatical categories, word classes), as in "RESULTATIVE aspect".

6 **Coding periphery vs. nucleus**: Peripheral parts of constructions will be given in brackets, with a plus (+) in the brackets for unbound peripheral items, and a minus (-) or equal sign (=) for bound peripheral items, in correspondance with the respective Source. If there is something in (regular) brackets without any further specification, it is an alternative description of the morph in front of the brackets: 'from' (ABLATIVE preposition).

With paths, where more than one item could be called nucleus according to our procedure, no brackets are used for either item.

In a few cases, brackets are used for other purposes: Where necessary context is not explicit, but implicit (restricted choice in grammar and lexemes), there may be a note like "(in present tense)", or something like "(.PRS)".If it was necessary to give a specific form, but other forms would be possible as well, that was indicated by "here: …" in brackets.

2.2.15 Source meaning (simplified)

1 **Main function of nucleus only:** This column is very similar to Source Meaning, but they represent only the one main function/meaning of the nucleus, and should be formulated in a more generalized manner, as these columns (or a derivative of them) will be used to make the databank searchable (at least, that is how it looks like at the moment).

2 **Grammatical labels:** For non-lexical functions (= grammatical functions) we have formulated a closed vocabulary of grammatical labels, which are listed and defined in a separate document.

3 **Formatting: grammatical** labels are in CAPS, **lexical** meaning in ‘quotation marks’.

2.2.16 Target form

1 **Material only**: In those columns, only **language material** was entered. So, there should be no functional description of anything, be it nucleus or periphery.

2 **One form (citation form)**: Only one form of the material is displayed, even if it is inflectable.

3 **Allomorphs**: What may be coded here, are allomorphs (**separated by commas**); in case of a zero-alternation (round) brackets are used instead.

4 **Phonological status**: status (clitic, bound, free) is indicated by using the **equal** “=”, **hyphen** “-“ or **space** “ “ respectively.

5 **Coding Periphery vs**. **Nucleus**: Although this is usually more relevant for the source, peripheral parts of constructions may be given in **brackets** for the target as well.

6 **Phonetic representation:** phonetic representation may occasionally be given **in square brackets** additionally.

7 **Zero morphemes:** zero is coded by the character **< Ø >**.

8 **(Non-Latin) orthographic representation**: in a few cases, the native orthographic representation in the respective writing system is given in **angle brackets** (**< >**).

9 **Syntactically bound but phonologically unbound morphs**: in a few cases (mostly from the Korean dataset), there are syntactically bound morphs, which are phonologically unbound. This is represented by a **dot (< . >)** instead of a space character between two morphs.

2.2.17 Target meaning

1. **One function:** Only **one target function** is given here. If there are more, which are important to mention, or if it is uncertain, this can be done in the comment section.
2. **Most grammaticalized / polygrammaticalization**: Rather than creating several paths, in cases of clear **polygrammaticalization** (several targets connected to the same instance of grammaticalization, i.e. source construction) we display only one prominent meaning in Target Meaning. This function should be the **most grammaticalized**, i.e. integrated into the grammatical system of a language. If it cannot be decided, the most basic of the most grammaticalized functions will be coded. Other target functions were sometimes mentioned into the comments.
3. **Function only**: In contrast to Source form and Target form, here it should be only functions, not material.
4. **Corresponding entries**: For complex constructions, **every item here corresponds to an item in** Target form. Peripheral items are marked equally as in the respective Target.
5. This field was more of a free text field for contributor and coder, and therefore we find some variation in how some functions are expressed. However, we tried to stick to the basic formatting: (core) **grammatical** labels in CAPS, **lexical** meaning in 'quotation marks', everything else lower case. Also lower case for additional higher-level description (grammatical categories, word classes), as in "RESULTATIVE aspect".

2.2.18 Target meaning (simplified)

1. **Main function of nucleus only:** These columns are very similar to the above ones, but they represent only the one main function/meaning of the nucleus, and should be formulated in a more generalized manner, as these columns (or a derivative of them) will be used to make the databank searchable (at least, that is how it looks like at the moment).
2. **Grammatical labels**: For non-lexical functions (= grammatical functions) we have formulated a closed vocabulary of grammatical labels, which are listed and defined in a separate document.
3. **Formatting: grammatical** labels are in CAPS, **lexical** meaning in 'quotation marks'.

2.2.19 Example (material)

1. **Target function**: the examples exemplify the target function (grammaticalized function) of the item in question.
2. **Sentential**: Examples are mostly whole sentences, sometimes only phrases or words.
3. **Orthography**: Consistent with the/a standard orthography of the respective language.

2.2.20 Example (glossing)

1. **Leipzig Glossing Rules**: We asked our contributors to stick to the Leipzig Glossing Rules for the glossing, whereever possible. We have no comprehensive list of all glosses used, and can only refer to the contributors on matters concerning the glossing of examples.
2. The glosses may reflect the earlier or later stage of the construction in question; as said above, their function in use, however, will be the grammaticalized target function.

2.2.21 Example (translation)

1. **Format:** translations are formatted in such a way that they begin with a **capital letter** if they are in fact sentential and are surrounded by **simple quotation marks** in any case.
2. **Optional: literal**: In cases where the source meaning of the construction is still a possible reading, a literal translation may be given additionally if the author chooses to (in brackets with the remark 'lit.' as is customary).

2.2.22 Example (reference)

1. **Contributor reference**: If the author did not provide a citation for the example, we took them to be the warrant of this information (p.c.).
2. **Bibliography**: The full bibliography will be found in the cldf version.

2.2.23 Comments

1. **Comments of any kind**: Here any comments concerning the path could be entered by the contributor, be it on metadata, description or evaluation.

2 **Several comments**: If several comments accumulate in one cell, those should be separated by semicolons.

2.2.24 Parameters

This part concerns the 8 grammaticalization parameters. Their precise definitions can be found in the questionnaire.

But some additional remarks might be necessary:

- The baseline for the parameter **decategorization** are the categories expressed on lexical items (nouns, verbs, adjectives) in the specific language. Similarly, for **syntagmatic variability**, the baseline is the degree of positional freedom of the lexical item.
- **Semantic Integrity**: The distinction between 'referential' and 'relational' (cf. values 2 and 3) should be understood in the conventional sense that lexical categories have denotations (they independently convey the concept of a property, action or object), while grammatical categories do not.
- Inflectional semantic case will be treated as value 4 for the parameter of **Semantic Integrity**, because practically every semantic case also constitutes a minor pattern of syntactic case.
- **Phonetic reduction**: the nucleus approach singles out one morph of the source construction as the nucleus, which is measured. It is compared to the **corresponding** part of the target construction. If elements fuse in the course of the development (univerbation) or the like, unintuitive values for phonetic reduction can occur. This method has the advantage oftheoretical consistency insofar as it applies the same metrics to the same unit of analysis (the nucleus).

2.3 Bibliography

Agard, Frederick B. 1984. A Diachronic View. (A Course in Romance Linguistics, 2.) Georgetown University Press.

Bisang, Walter. 2011. Grammaticalization and typology. In Heiko Narrog & Bernd Heine (eds.), Handbook of grammaticalization, 105–117. Oxford: Oxford University Press.

Bisang, Walter; Becker, Laura; Malchukov, Andrej & Martiny, Marvin. submitted. Grammaticalization scenarios: constraining typological variation.

Bisang, Walter & Andrej Malchukov (Eds.). *Grammaticalization Scenarios. Cross-linguistic Variation and Universal Tendencies*, 2 Volumes. Berlin: De Gruyter Mouton.

Bisang, Walter; Malchukov, Andrej; Rieder, Iris & Linlin, Sun. 2020a. Position paper: Universal and areal patterns in grammaticalization. In Walter Bisang, Andrej Malchukov (Eds.): *Grammaticalization Scenarios. Cross-linguistic Variation and Universal Tendencies.* Volume 1: *Grammaticalization Scenarios from Europe and Asia*. Berlin: De Gruyter Mouton (Comparative Handbooks of Linguistics [CHL], 4.1), pp. 1–87.

Bisang, Walter; Malchukov, Andrej; Rieder, Iris & Linlin, Sun. 2020b. Measuring grammaticalization: A questionnaire. In Walter Bisang, Andrej Malchukov (Eds.): *Grammaticalization Scenarios. Cross-linguistic Variation and Universal Tendencies.* Volume 1: *Grammaticalization Scenarios from Europe and Asia*. Berlin: De Gruyter Mouton (Comparative Handbooks of Linguistics [CHL], 4.1), pp. 89–103.

Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. *Language* 68: 81-138.

Hammarström, Harald, Forkel, Robert & Haspelmath, Martin. 2022. Glottolog. https://glottolog.org. Version 4.5, accessed Jan 10th 2022.

Harald Hammarström and Mark Donohue. 2014. Some Principles on the use of Macro-Areas in Typological Comparison. In Harald Hammarström and Lev Michael (eds.), Quantitative Approaches to Areal Linguistic Typology, 167-187. Leiden: Brill.

Heine, Bernd & Kuteva, Tania. 2002. World Lexicon of Grammaticalization. Cambridge University Press.

Kuteva, Tania; Heine, Bernd; Hong, Bo; Long, Haipang; Narrog, Heiko & Rhee, Seongha. 2019. World Lexicon of Grammaticalization (2nd ed.). Cambridge: Cambridge University Press.

Lehmann, Christian. 1995 [1982]. Thoughts on grammaticalization. A programmatic sketch. Munich: Lincom Europa.

III Further: How to cite / Contact / License

3.1 License

3.2 Contact

You can contact us via email at wbisang@uni-mainz.de

3.3 How to cite MAGRAM

The MAGRAM online database is an edited database consisting of different languages which should be regarded as separate publications, like chapters of an edited volume. These datasets exemplified by Mayan should be cited as follows:

Christian Lehmann & MAGRAM editorial team. 2025. *Yucatec Maya*. In: Bisang, Walter & Malchukov, Andrej & Martiny, Marvin (eds.), MAGRAM: Mainz Grammaticalization database. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Available online at: https://crossgram.clld.org/contributions/magram

The complete database should be cited as:

Bisang, Walter, Malchukov, Andrej & Martiny, Marvin (eds.). 2025. *MAGRAM: Mainz Grammaticalization database*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Available online at: https://crossgram.clld.org/contributions/magram