Frajzyngier, Zygmunt & Amina Mettouchi (2015). Functional domains and cross-linguistic comparability. In: Mettouchi, Amina, Martine Vanhove and Dominique Caubet (eds.), Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages. Studies in Corpus Linguistics 68. John Benjamins: Amsterdam-Philadelphia. vi, 332 pp. + index (pp. 257-279).

Functional domains and cross-linguistic comparability*

Zygmunt Frajzyngier and Amina Mettouchi
University of Colorado, Boulder, EPHE and CNRS-LLACAN

This paper investigates a strategy other than the one currently implemented in the CorpAfroAs project, allowing cross-linguistic comparison among multiple-language corpora. It involves comparing a database of functional domains and subdomains across languages. The underlying principle is that cross-linguistic comparison should be conducted on the basis of meanings/functions actually encoded in the grammatical systems of individual languages rather than on the basis of aprioristic categories. Such a study yields reliable information regarding the differences and similarities between grammatical systems.

1. Introduction

This paper proposes a theory and methodology aimed at overcoming a long-standing difficulty in choosing the proper object in cross-linguistic comparison. The fundamental assumption is that languages may differ significantly in the functional domains and subdomains encoded in their grammatical systems and also in the specific meaning and functions belonging to those grammaticalized domains. The term 'grammaticalized' in the present study indicates the encoding of a given meaning through some formal means, e.g. inflectional means, auxiliaries, linear order, etc. It has long been assumed that even if languages encode the same meanings or functions, the formal means by which they are encoded may differ considerably, even among related languages. Once we can isolate specific meanings grammaticalized in a given language, we will have a better tool for the comparison of the formal means used. This paper adapts a previously developed theoretical approach to linguistic typology (Frajzyngier & Shay 2003; Frajzyngier

* Frajzyngier's work on this study was supported by the *Chaire de chercheur étranger des pays de la Loire*. This support is gratefully acknowledged. Frajzyngier's previous work on Mina and Hdi was supported by grants from National Science Foundation.

2010, 2011, 2013, submitted), and considers in detail one of the issues for which the CorpAfroAs project provides an important source of information. To a certain extent, this paper complements the contribution by Mettouchi, Savà and Tosco, in this volume. The domain that will illustrate the theory and methodology is the reference system; the languages examined are Kabyle (Berber) and Mina (Central Chadic).

2. Basic question

What should be the starting point in typological comparison? The proper object of linguistic typology depends very much on the linguist's aims and the aspects of the language targeted for study. Therefore, there is no single answer to this question. When it comes to the comparison of what is often referred to as syntactic structures, linguists are as divided with respect to typology as they are with respect to other major theoretical issues. Some linguists agree that the proper object of linguistic typology should be functions (Lazard 2004; Seiler 1995; Haspelmath 2007, 2010); Newmeyer 2007 maintains the notion that formal structures are the starting point.

Even for those who favor functions, exactly which functions should be chosen remains a controversial issue. For Frajzyngier and Shay (2003), only functions actually encoded in the language should be targeted. For Haspelmath (2010) following Lazard (2004), one should make an abstraction from the functions actually encoded and instead postulate some canonical functions, that they refer to as 'comparative concepts'.

In practice, the selection of functions is largely intuitive, informed by linguists' experience with a variety of languages as evidenced by many chapters in World Atlas of Language Structures (Haspelmath *et al.* 2005) and discussions in Dixon (2010) where a large inventory of functions is given, without explicit explanation about why certain categories were selected rather than others.

3. Theoretical assumptions of the present approach

Every grammatical system encodes a finite number of grammaticalized meanings. Those meanings are organized into a system of functional domains, such as Aspect, Reference, Noun modification, etc. The number of functional domains is finite for each language (Frajzyngier & Mycielski 1998).

At any given time, the number of grammaticalized meanings in a language is also finite, although languages may grammaticalize new meanings and lose others.

For the purpose of synchronic analysis, the constant diachronic change may be ignored. Each grammaticalized meaning may be realized by one or several constructions, and those constructions do not have to share formal characteristics. One formal means, e.g. reduplication, may encode functions belonging to different domains such as aspect (progressive, habitual or perfective depending on the language), quantification, or plurality. Portmanteau morphemes by their nature encode functions belonging to different domains.

Grammaticalized meanings are organized in a system of functional domains and subdomains. Members of the same subdomain are in complementary distribution: if one occurs, the other cannot. Members of different domains can co-occur with each other. The discovery of the functional domains, subdomains, and grammaticalized meanings constitutes the discovery of the semantic structure of the language (Frajzyngier, submitted).

Once one has discovered individual forms, including constructions, one is able to discover functional domains and subdomains through the study of the co-occurrence of various forms. This done, the individual grammaticalized meanings can be described through contrast with other grammaticalized meanings belonging to the same domain.

Given the common physiological, cognitive, and psychological characteristics of humans, and their common social needs, one can expect some degree of overlap in the meanings that have been grammaticalized in individual languages. One can also expect a number of differences across languages with respect to the domains that have been grammaticalized and with respect to individual grammaticalized meanings.

4. How to discover grammaticalized meaning

4.1 General methodology

In order to discover grammaticalized meanings one must first have a full list of the formal means of encoding available for individual languages. Then, one can proceed with the discovery of which formal means of encoding are in complementary and which are in contrastive distribution, very much in the manner of discovering underlying segments in phonological analysis. Forms that can co-occur encode meanings belonging to different domains, unless those forms are components of one means of encoding. Forms that cannot co-occur most probably encode meanings belonging to the same domain. Once again, the description of the grammaticalized meaning is accomplished through comparison with other meanings encoded in the given domain. Consider the distribution of definite and indefinite articles in English. The fact that *a* and *the* cannot simultaneously determine a noun indicates that they belong to the same domain. Extending the argument, the deictics 'this' and 'that' are also in complementary distribution with definite and indefinite articles. Hence they also belong to the same domain of reference.

Consider now the co-occurrence of articles and deictics/demonstratives with possessive pronouns. Determiners and possessive pronouns cannot occur in any sequence with a noun in English: *a my book, *the my book, *this my book, *that my book, *my a book, etc. The constraints on co-occurrence in a sequence may imply that possessive pronouns and determiners also belong to the same domain. But that is actually a false conclusion. In English there exists a construction that is specifically dedicated to accommodating the encoding of functions subsumed under the definite and indefinite articles and the determiners with nominal or pronominal possessors. The construction has the form: DET N of Possessor, followed by a significant pause. The possessor may be nominal or pronominal. If it is pronominal it is marked by a specific set of independent possessive pronouns, e.g. *mine*, *yours*, etc. (1) (all English examples from COCA):

(1) a. A friend of mine called me,
b. All right, Jay-Z, of course, he's probably a friend of yours.
c. This friend of his was seen taking $5,000 in cash away from the home as well as a Rolex.

So possessive pronouns in English can co-occur with determiners in the same construction, and therefore they belong to different domains. Note that this conclusion has been stated only with respect to English. The structure of other languages may be quite different.

4.2 Formal means of encoding

Languages differ significantly with respect to the types of formal means available and with respect to the composition of each type. We do not yet have a comprehensive list of formal means of encoding available across languages or for any individual language. The following list is illustrative rather than exhaustive.

1. Lexical categories and subcategories. Languages differ in the types of categories they exhibit. Some languages have adjectives; others do not. Some languages have adverbs; others do not.
2. Free grammatical morphemes (prepositions, complementizers, subordinators, etc.)
3. Auxiliary verbs and the use of nouns as a encoding means in the grammatical system (e.g. spatial specifiers in a number of Chadic languages).

4. Linear orders (which is slightly different from the notion of 'word order' as it is used in functional typology: linear orders include different types of encoding through position, relative order and variation on the default order, as described in Frajzyngier (2011a).
5. Tone, stress, various types of vowel harmony, and other phonological means including: rhythm, pause, intonation, creaky voice, intensity.
6. Inflectional means on all lexical categories. The inflectional means include affixations, phonological changes of underlying segments, gemination, reduplication, apophony.
7. Serial verb constructions, i.e. the use of verbs to encode grammatical meaning. This formal means is widely attested in languages of Africa, South-East Asia, and Australia.
8. Repetition of phrases comprising a noun and a verb, hence quite distinct from reduplication (Gidar, Frajzyngier 2008).

5. Examples of functional domains

The notion of functional domain has been informally recognized in many traditional grammars, when talking about the tense system, aspectual system, or relative clauses, but the composition of a functional domain is generally not explicitly justified. Thus, in many traditional and contemporary works there is a good deal of hesitation regarding assignment of a given form to a given domain. Typical examples involve tense and aspect, even in some languages with a long and rich tradition of research (consider the early statement by Chaim Rabin about Semitic languages quoted in Izre'el (2002) that some tenses behave like aspects and vice versa).

In the present approach, the notion of functional domain is crucial for determining grammaticalized meaning. The functional domain is a set of functions that all share one semantic characteristic, and the forms that realize them are in complementary distribution within a relevant constituent. Thus, if a language has grammaticalized the domain of tense, one should not have two tenses in the same clause (a tense can however be realized by a complex construction involving several morphemes such as inflexion on a auxiliary, plus a participial form). If a language has also grammaticalized the domain of aspect, a clause may contain a tense and an aspect, but not two aspects. It may happen that two forms individually encode different functions, and are combined to encode a third function.

The following are some of the functional domains grammaticalized in many language families:

- Semantic relations among sentences in discourse;
- Semantic relations among clauses in sentences;
- Mood;
- Aspect;
- Tense;
- A variety of relations between nouns or noun phrases;
- A variety of relations between noun phrases and the predicate;
- Reference system;
- Equational predications;
- Social relationships between the speaker and the addressee(s);
- Spatial relationship between the place of speech and an event;
- Information structure.

An utterance may consist of several grammaticalized meanings. Some meanings may be concatenated and others may be embedded in a hierarchical structure. It will indeed be rare for an utterance to consist of just one grammaticalized meaning. Even a single-word utterance may represent two or more grammaticalized meanings. For example, the imperatives of intransitive verbs consist of the imperative modality and the intransitive predication. Moreover, structures consisting of the same form and the same lexical categories, e.g. coordinating conjunctions of adjectives or verbs, may represent different grammatical meanings across languages (Frajzyngier 2012).

6. Similarities and differences among languages

6.1 Languages may encode different functional domains

An examination of individual languages indicates that they may vary significantly in the functional domains they have grammaticalized. Thus some languages have grammaticalized the domain of aspect, others have grammaticalized the domain of tense, and still others have grammaticalized both. Some languages (e.g. Hausa, West Chadic, Nigeria) have grammaticalized the domain which indicates location of the event with respect to the place of speech, distinguishing between ventive and andative, whereas other languages (e.g French) have not grammaticalized such a distinction.

6.2 Languages differ in the internal structure of the functional domain

Even if languages have grammaticalized the same functional domains, the internal structure of such domains may differ significantly across languages. This fact has been well documented for the domains of tense and aspect (see Comrie 1976; Dahl 1985; Cohen 1989; Bybee 1994).

6.3 Languages differ in their grammaticalized meanings

Even if languages have grammaticalized the same functional domains and the same subdomains, they may differ in their grammaticalized meanings. If two languages have a subdomain of future, but in one language this consists of two or three tenses, while in another it consists of only one, then the tenses in the subdomains of future in the two languages have different values. Let us consider a specific example. Although English and Hdi (Central Chadic, Nigeria) both have a tense whose time reference is past with respect to the time of speech, the English past tense does not specify what kind of past it is, while the Hdi past tense refers to a past that must be accompanied by a specific time reference (2).

(2) *ká-xə̀n mántsá, sí ndá gá ká ndá rvídə̀k*
COMP-3PL like that PAST ASSC where 2SG ASSC night
'and they said: where were you last night?'

If the time of the event is past but not specific, the form *sí* is not used and there is no marking of tense. The clause below was taken from a narrative referring to an event in the past, but in another context it could have meant a habitual or ongoing event (3):

(3) *lá-ghà pákáw ghúvì kà mná-n-tá krì*
go-D:PVG hyena SEQ tell-3-REF dog
'And Hyena said to Dog, go!' (Frajzyngier with Shay 2002)

Thus, the English past tense does not encode the same time-reference as the Hdi referential past.

7. Potential meanings within the domain of reference

The domain of reference enables the listener to identify the referent of a nominal, pronominal, verbal, or propositional form. The types of subdomains encoded are quite large, and as with all other subdomains, differences across languages are quite significant. The traditional distinction between reference to the environment of speech (deixis) and reference to the elements in discourse (anaphora) is not sufficient to account for the varieties of forms in various languages. Languages from different families make an additional distinction between the *de dicto* and *de re* domains (Frajzyngier 1991, 1997). The *de dicto* domain includes hypothetical situations and reported speech. The *de re* domain includes actual rather than hypothetical referents. In some languages, the *de dicto* domain encodes fewer distinctions than the *de re* domain. Thus, in Mupun (West Chadic, Nigeria) there is a distinction between the second person masculine and feminine in the *de re* domain, but there is only one gender, masculine, in the *de dicto* domain. Here is an example illustrating the *de dicto* domain, where the masculine pronouns refer to an obviously feminine referent (4):

(4) *gaskiya, get kadan ka kə ak ɓe*
truly (H.) past if (H.) 2M with pregnancy SEQ
ba də mo pə ɪal dik n-ka
NEG PAST 3PL PREP marry PREP-2m
'Truly, in the past if you (masc.) were pregnant they wouldn't marry you.'
(Frajzyngier 1993: 88)

7.1 Mere formal categories are not good indicators of function

The system of reference is one of the many systems where formal categorial similarity across languages is not a good predictor of function. Let us consider the category 'subject pronoun' whose existence can be established through standard distributional analysis, the analysis of morphological forms, linear orders and other discovery techniques.

The functions of subject pronouns in English are quite different from those of Polish (Frajzyngier 1997). In English, subject pronouns sharing the features of gender, person, and number with the preceding subject noun encode coreference with the preceding subject. Subject pronouns in Polish, on the other hand, always encode switch reference or contrastive focus with respect to the preceding subject, regardless of how the preceding subject is marked (Polish examples from Polish National Corpus, NKJP). In example (5), the third person plural subject pronoun *oni* is deployed even though the verb in the complement clause also encodes the third person plural masculine subject. The reason for the deployment of the pronoun is that the preceding subject is second person singular (5):

(5) *I pomyślałaś, że oni nie chcieliby,*
CONJ think:PRF:PAST:2M:F COMP 3PL:M NEG want:PAST:HYP:3PL:M
żebyś była takim tchórzem, który
COMP:HYP:2SG be:3F:PAST such:INSTR coward:INSTR REL:M

boi się pójść na studia.
fear REFL go on study:PL
'And you thought that they wouldn't want you to be such a coward not wanting to go onto higher studies'

Coreference in Polish is encoded by the marking of gender, person, and number on the verb. Consider the following example where the complementizer is not followed by a pronoun, and the verb encodes the first person singular subject just like the verb in the matrix clause (6):

(6) *Właściwie nie mogę powiedzieć, że ją znam.*
rightly NEG be able:1SG:PRES say COMP 3F know:1SG
'To tell you the truth, I cannot say that I know her.'

One of the important factors for predicting whether identical categories have the same function is the analysis of the functional domain to which the category belongs, and in particular, whether or not there are other forms belonging to the same domain. Thus, for the examples above, the mere fact that Polish has an agreement system in addition to subject pronouns and English does not should tell us that subject pronouns in English and Polish must differ in their function.

Consider now a different case. English has only one set of subject pronouns. The third person subject pronoun in the complement clause of the verb of saying encodes coreference with the subject pronoun of the matrix clause (7):

(7) After that, he said he wanted to do some different moves on me...

Mupun, on the other hand, has two sets of subject pronouns, each with its own functions, and neither of which has the same function as the subject pronouns in English. One of the third person subject pronouns in Mupun encodes coreference with the subject of the verb of saying in the matrix clause; the other encodes switch reference with respect to the subject of the matrix clause. The logophoric pronouns, to use Hagège's (1974) term for pronouns encoding coreference, are morphologically different from the pronouns of the matrix clause. Pronouns encoding switch reference are morphologically identical with the pronouns of the matrix clause (Frajzyngier 1985, 1993). Some of the potential meanings encoded in the systems of reference are listed in the next section (7.2), based partially on Frajzyngier (2011b).

The use of a determiner or a deictic with a noun does not necessarily mean that the form involved belongs to the domain of reference. Thus, in many Chadic languages, topicalization (i.e. establishing the topic of a paragraph or of a sentence) is frequently encoded by adding a determiner to a noun. In Mina, this is the standard means of topicalizing a noun, as in example (8), which is the first clause of a narrative establishing the topic, thus providing the necessary evidence that the determiner does not have a deictic or an anaphoric function:

(8) *hìd-yíì wà í tàtà màkáɗ*
man-PL DEM 3PL 3PL three
'There were three men.'

7.2 Selected subdomains and grammaticalized meanings within the domain of reference system

In this section we include the subdomains of deixis and anaphoric reference, as well as individual grammaticalized meanings, e.g. a marker indicating that the speaker should deduce the identity of a referent from previous discourse, when the referent has not yet been mentioned. The deduced reference marker instructs the listener to identify the referent through a process of deduction using knowledge from various sources, including the listener's cognitive system, the speech environment, and previous discourse. In Mina (Central Chadic) the deduced reference marker *tá* may be the only component of a noun phrase or it may be a determiner, modifying another noun or a quantifier (9a–b).

(9) a. *žíŋ ngùl-yíi pár sùlúɗ tàn*
then man-PL other two DED
í nd-áhà bàhá
3PL go-GO again
nd-á mábàr mbír bàhá kə́ mə̀l tàŋ
go-GO lion leap again INF seize DED
'Later, when the two men arrived, the lion jumped to catch them.'
b. *mbígìŋ wàcíŋ í ɗál ngàm mə̀ts*
mbiguin DEM 3PL do because sickness
kə̀ ɗál nə̀ hàyák í hóynə̀ tàŋ
INF do PREP village 3PL calm (F.) DED
'This mbiguin [a ritual], they do it because there is sickness in the village. They cure it.'

One piece of evidence for the proposed function of the marker *tá* is that its antecedent need not have been mentioned previously. In fact, the presence of *tá* explicitly tells the listener that the referent is not the noun marked by *tá* but some other referent associated with that noun. In the last line of the following fragment (10), *tá* follows the noun *báy* 'chief', which has already been mentioned several times. However, the form *tá* does not identify the chief himself but rather his court, an entity that has not been mentioned in the discourse at all (10):

(10) *báy zá ngwáy bàhámàn bákà bá*
chief COMP 'People' Bahaman today still
dzán-á nòk mí
find-GO 1PL what
'The chief said, "People, what else did Bahaman find for us today?"'

hí ndə̀ lùw-á-ŋ mə́ ndà-hà
2PL go say-GO-3SG DEB go-GO
'"Go tell him to come here."'

ndá yà í y-ù
go call 3PL call-3SG
'Someone went to call him.'

tíl á nd-á á r báy tàŋ
go 3SG go-GO PRED PREP chief DED
'He went to the chief's [court].' (Frajzyngier 2005: 335)

Such grammaticalized meanings can constitute parts of a subdomain together with other grammaticalized meanings or they may make up a subdomain of their own. Note here the importance of the notion 'grammaticalized'. In addition to the implication that there are specific formal means that encode a function, it also implies that the function must be encoded if the event referred to contains referents that meet the criteria for the given function. Here is the list of grammaticalized meanings within the domain of reference in some Chadic languages:

- Deixis referring to entities: proximate and remote, with speaker, listener, or neither as point of reference, e.g. Hausa (Jaggar 1994);
- Locative deixis as distinct from locative anaphora, e.g. Mupun (Frajzyngier 1993), Hdi (Frajzyngier with Shay 2002);
- Previous reference to entities, with the distinction between proximate and remote as distinct from deixis and not overlapping with the English definite, e.g. Hdi, Mupun;
- Locative anaphor, proximate and remote, as distinct from locative deixis, e.g. Mupun;
- Deduced reference as described above, e.g. Hdi;
- Instruction to the listener to identify the referent in any way they can, e.g. definite article in English;
- Unspecified, indefinite member of a set;
- Disjoint/conjoined reference (also called switch reference and same reference) in sequential and complement clauses;
- Logophoricity with subject and object in its scope (e.g. Mupun, Frajzyngier 1993), confined only to a limited number of matrix clause predicates);

- Encoding the referent as known rather than previously mentioned (first described by Ebert 1971 for Frisian, also present in Mina, Frajzyngier 2005).

8. A proposal for the structure of a database

The theoretical model of Systems Interactions (Frajzyngier & Shay 2003) could be translated into a configuration amenable to software implementation via a database allowing the investigation of the CorpAfroAs corpus.

Once the languages under consideration have been examined, the identified grammaticalized meanings would be grouped into subdomains within domains, and the encoding means (i.e. the forms) by which they are expressed would be listed. Constraints on the occurrence of those grammaticalized meanings would be given, as well as the appropriate labels reflecting grammaticalized meanings. If the same functions were found to be encoded in other languages, so much the better. If such parallels were not found, this would contribute to the creation of a non-aprioristic typology. Such an approach would be useful for linking the actual language-internal labeling of the discovered grammatical meaning to their equivalents in various linguistic or typological theories, or in different analytical traditions.

Each form belonging to a grammaticalized meaning would be retrievable in the corpus by means of a query pointing to annotations in the two glossing lines "*ge*", and "*rx*"[1]. Thus, each file in the database would correspond to a grammaticalized meaning in a particular language, and would contain instructions for retrieval of the associated structures or forms in the CorpAfroAs corpus. If the names of domains and subdomains were shared among languages, this would ensure the possibility of retrieving grammatical meanings (with their associated encoding means) among several languages of the corpus. If the names of domains and subdomains were different, that would demonstrate the extent of cross-linguistic differences.

For example, in Kabyle, 'identifiability of the head of a relative clause' is a grammaticalized meaning that is encoded by the use of relator *i* just after the head noun. This relator has been encoded as REL.REAL on the *ge* tier, and as PTCL on the *rx* tier. The following file in the database allows the identification of the grammaticalized meaning, of its insertion in a subdomain and a domain, and the retrieval of the relevant sequences in the Kabyle corpus.

[1] For more details about the contents of *ge* and *rx*, see the Introduction to this volume, as well as Mettouchi and Chanard (2010).

Domain	Reference
Subdomain	Reference in relative clause
Grammaticalized meaning	Identifiability of the head of a relative clause
Form	relator *i*
Constraints	Cannot appear with *irrealis* mood in the verb of the relative clause
Annotation in *ge*	REL.REAL
Annotation in *rx*	PTCL
Retrieval instruction	"look for REL.REAL in *ge* and PTCL in *rx*, both annotations vertically aligned on the same morpheme (*mb*) cell"

(11) *ufsinara jinjinənni* /

ur	*fsi-n*	*ara*	*jinjən-nni*	/
NEG	melt\NEGPFV-SBJ3PL.M	POSTNEG	hearth_stone\ANN.PL.M-CNS	/
PTCL	V24-AFFX	N.INDF	N.OV-DEM	/

iθəxðəm akkən fəlkanun //

i	*t-xdəm*	*akkən*	*af*	*lkanun*	//
REL.REAL	SBJ3SG.F-make\PFV	thus	on	hearth\ABSL.SG.M	//
DEMPRO	PRO-V23	ADV	PREP	N.COV	//

'the fireplace stones didn't melt, which she had thus put in the fireplace.'
(KAB_AM_NARR_02_420–422)

In Mina (see Section 9.1) relative clauses distinguish between two types of referentiality. If the head is not pronominal, the referentiality of the head of the relative clause is marked by the clause-final demonstrative *wàcín or wàhín*. If the head is pronominal, hence inherently referential, the clause-final demonstrative does not occur. The non-referentiality of the head of the relative clause is unmarked, and does not have a clause-final demonstrative. The grammaticalized meaning 'referentiality of the head of the relative clause' seems close to the 'identifiability of the head of the relative clause' of Kabyle. Mina is not part of CorpAfroAs, but had it been, the encoding means of the grammaticalized meaning, the relative clause marker *wàcín*, would have been annotated as DEM in *ge* and DEM.REL in *rx*. So it would have been systematically retrievable with its context, thus allowing a thorough investigation of possible similarities between Kabyle and Mina.

Domain	Reference
Subdomain	Reference in relative clause
Grammaticalized meaning	Referentiality of the nominal head of a relative clause

Form	clause-final demonstrative *wàcín or wàhín*
Constraints	If the head is pronominal (inherently referential), the clause-final demonstrative does not occur
Annotation in *ge*	DEM
Annotation in *rx*	DEM.REL
Retrieval instruction	"look for DEM in *ge* and DEM.REL in *rx*, both annotations vertically aligned on the same morpheme (*mb*) cell"

A representation of relative clause in the CorpAfroAs convention would look something like (12) a. for non-referential head:

(12) a.

hìdì	*mə̀*	*ɓám*	*mə̀kwádàk*	*gə̀r*	*kə́*	*nzə̀*	*kə́*
man	REL	eat	vulture	search	INF	be	like
N	PTCL	V	N	V	PTCL	V	PREP

ngámbə̀-n	*skù*
friend-1SG	NEG
N-PRO	PRT

'The man who ate/eats vulture cannot be a friend of mine.'

A representation of the relative clause with the referential head would have the form (12) b., where the clause–final demonstrative *wàcíŋ* (bolded) codes the existential status of the head, its referentiality:

(12) b.

séy	*hìdì*	*mə̀*	*ɓà*	*kə̀sáf*	***wàcíŋ***
SO	man	REL	cut	grass	DEM
PTCL	N	PTCL	V	N	DEM

à	*zá*	*wàcíŋ*	*tá*	*nàŋ*
3SG	COMP	DEM	GEN	1SG
PRO	PRT	DEM	PRED	PRO

'The man who cuts grass said, "This is for me."'

9. Application of the database to the reference system in Mina and Kabyle

9.1 The domain of reference in Mina

The data in this study comes from Frajzyngier et al. (2005), in some cases with an updated analysis. Mina does not have gender, one of the frequent formal means for encoding reference (Frajzyngier & Shay 2003). Neither is the plurality of nouns obligatorily marked. These two factors reduce the potential use of pronouns as reference markers. The third person pronoun, unmarked for gender, can refer to any noun in the previous discourse. Hence, one can reasonably expect the presence of some other means of encoding to enable reference across discourse. The encoding means relevant for the domain of reference in Mina include:

- Overt use of a noun;
- Absence of a noun;
- A noun followed by a determiner;
- Previous mention markers;
- Deictics;
- A deduced reference marker;
- Independent pronouns;
- Obligatory subject pronouns;
- Object pronouns.

Table 1 lists some grammaticalized meanings in the domain of reference and their associated encoding means in Mina.

Table 1. Subdomains in the Mina system of reference

Subdomain	Grammaticalized meaning	Encoding means
Deixis	Proximate	(N) *wà* (can occur alone)
Mention in discourse	First mention (non-topicalized)	Noun alone
	Remote previous mention	N (-POSS) *nákáhá*
Known	Instructs the listener to consider preceding noun as known	N wà (can follow remote previous mention)
	Locative	*mè-hín* (for inherently non-locative antecedents) *màcín* (only for inherently locative antecedents)
Deduced reference		(N) *ta*
Coreference/switch reference	Switches reference from immediately preceding antecedent to another, previously mentioned. The antecedent may be a noun phrase or an event	(PREP) *mbí*, or use of the third person singular subject pronoun
	Coreference	No subject pronoun
Indefinite	Always in the domain *de re*, 'one of a set'	N *dáhà*
Unspecified	Human	*ví* 'who', *í* '3PL'
	Non-human	*ú* (object)

Table 1. (*continued*)

Subdomain	Grammaticalized meaning	Encoding means
	Place	*váy* (directional) *tíkì* (stative), *ngíd* 'place other than the place of speech'
Head of relative clause	Non referential	No determiner after the relative clause
	Referential	Rel. Clause *wàcín or wàhín*

9.2 The domain of reference in Kabyle

The encoding means relevant for the domain of reference in Kabyle include:

- Gender marking on nouns and adjectives;
- Number marking on nouns and adjectives, both singular and plural;
- Overt use of a noun;
- Absence of a noun;
- Independent pronouns;
- Deictics;
- Obligatory subject pronouns;
- Optional object pronouns;
- Interrogative pronouns;
- Relators;
- Unspecified nouns.

Table 2 lists some grammaticalized meanings in the domain of reference and their associated encoding means in Kabyle.

Table 2. Subdomains in the Kabyle system of reference

Subdomains	Grammaticalized meanings	Encoding means
In all types of clauses		
Deixis	Proximal Deixis	Affixed demonstratives: *-a* ; *-agi* ; *-agini*
	Distal Deixis	Affixed demonstratives: *-in* ; *-inna* ; *-ihin* ; *-ihinna*
Known reference		Affixed demonstrative *-nni*
Coreference		bound pronouns
Switch reference		independent pronouns
Specification	Unspecified entity nouns	*jiwən* ('one', animate), *afəmma* ('thing'), *ara* ('thing' in negative contexts)

Table 2. (*continued*)

Subdomains	Grammaticalized meanings	Encoding means
In relative clauses		
Identifiability	Identifiability of the head in relative clause	relator *i* following the head.
Specification	Unspecified antecedent	animate pronoun (with gender and number distinctions: *wid, tid, win, tin* non-animate pronoun (no gender-number distinctions: *ajən*)
Identification	Unidentified referent	interrogative pronouns : animate (with gender and number distinctions: *anwa, anta, anwi, anti*), non-animate (no gender or number distinctions: *afu*), place (*anda*)

All of the grammaticalized meanings that constitute this domain are retrievable in the corpus, with the relevant search instructions. For instance, *-agi* can be found by looking for [PROXb in *ge* and AFFX in *rx*]; *win* by looking for [the_one\SG.M in *ge* and INDF.PRO in *rx*]; relator *i* by looking for [REL.REAL in *ge* and DEM.PRO in *rx*], etc. In this way, it is possible to examine in greater detail not only the general context of occurrence of the different encoding means and their frequency, but also their combination with other predications.

10. Comparison of the domain of reference in Mina and Kabyle

We can now compare the structure of the domain of reference, and the grammaticalized meanings in the domain of reference in two languages belonging to the same phylum. At this stage of analysis, we cannot be absolutely sure that the subdomains as listed above are indeed properly delimited nor can we be sure that all the grammaticalized meanings have been properly identified.

To give an example of the issues involved, Frajzyngier et al. (2005) analyzes the marker *ta* (phrase-final form *tàŋ* and *táŋ*) in Mina as a remote deictic marker for entities and a deduced reference marker. However, upon closer scrutiny, no natural language evidence emerges for the remote deixis function of *ta*. All examples that were taken to represent the remote deictics could equally well be analyzed as representing the deduced reference marker, which is why only the latter has been retained in Table 1. The table demonstrates an asymmetry in the subdomain of deixis: there is only the grammaticalized meaning of proximate deixis and there is no remote deixis.

The cross-linguistic comparison has at least two obvious benefits. The first is the possibility of discovering ways in which languages are similar and different, with respect to a given domain. The second benefit — much more important — is heuristic: it provides an indication of possible directions for future research. If there are differences across related languages, even remotely related ones, one would like to find the reasons for those differences, similarity among languages of the same family being the default assumption.

A comparison of the subdomains and the individual grammaticalized meanings within the domain of reference between Mina and Kabyle reveals the following similarities and differences:

Subdomain of deixis

The subdomain of deixis is present in both languages. While Mina has only a proximal deictic marker, Kabyle has two series, proximal and distal, each differentiated for gender and number. A possible avenue for future research would be to explore whether it is really the case that Mina does not have a remote deictic marker.

Mention in discourse

Kabyle does not appear to have the domain 'Mention in discourse'. Mina, on the other hand, has a specific form restricted to the first mention in discourse and remote previous mention. Possible research questions to be pursued are: is it indeed the case that (a) Mina encodes remote previous mention, and that (b) Kabyle has not grammaticalized any meanings related to previous mention?

Known reference

Both Kabyle and Mina have grammaticalized the category 'Known reference'. Although sharing of the category between two related languages is the default expectation, one should nevertheless be careful here given the fact that this category is otherwise quite rare across languages. One of the early descriptions of such a category was Ebert's (1971) work on Frisian. In Mina, the category 'known' is marked by the form *wa*, with its clause-final variants *wàcín* or *wàhín*. The marker indicates that the referent is to be treated as a known entity, regardless of whether the listener actually knows the referent. The source of the knowledge could be previous mention in discourse, regardless of the distance between the previous mention and the current mention, or it could be some other source of knowledge (see Frajzyngier et al. 2005). The source of knowledge could also be what is generally expected from any speaker of the language. The category 'known' is different from 'previous mention'. All previous mentions are known, but not every known element has a previous mention as its source, as the following Mina examples illustrate. In (13), baboon is mentioned in the immediately preceding sentence. In (14), the teacher was the topic of the previous paragraph, but last mention as *màllúm* was five sentences earlier. In between there were several other participants mentioned. In (15) the noun *mìšìl* was mentioned five clauses earlier, the act of stealing two clauses earlier.

(13) *kwáyàŋ à ndíŋ bə̀ làkáf wàcíŋ*
squirrel 3SG fear ASSC baboon DEM
'The squirrel was afraid of that baboon'

(14) *nd-á déw ká á bə́r màllúm wàcíŋ*
go-GO sit POS PRED side marabout DEM
'He came to sit next to this teacher.'

(15) *í kə̀́ mə̀l zá á n mìšíl wàhín*
3PL INF seize EE PRED PREP theft DEM
'They arrested him for stealing.'

Consider the following fragment, where in the first sentence the noun *fòrə́m* 'horn' is encoded by the remote previous mention marker *nákáhà*. In its mention in the next sentence, it is followed by the form *wà*:

(16) *í hók rə̀ wàcíŋ séy wàl wà*
3PL lift D.HAB DEM then wife DEM

ɓə̀t á ɓə̀t fòrə́m nákà bə́ vènjéh
take 3SG take horn REM ASSC pepper

díyà á dí kə́ nə̀ mà
put 3SG put in PREP mouth
'When they were lifting [the stones], the wife took the horn which contained pepper and put it in her mouth.'

íf á íf-é tə́ n fòrə́m wà dàp
blow 3SG blow-GO GEN PREP horn DEM just
'She just blew out what was in the horn.'

Mina, unlike Kabyle, has grammaticalized reference to known locative argument, with further differentiation between the reference to inherently locative and inherently non-locative noun phrases. The facts in Mina are consistent with the existence of locative predication in this language.

Deduced reference

Mina has the category of deduced reference and Kabyle does not. The obvious research question is to make sure that the analysis of the marker *ta* in Mina really holds. Another would be to explore whether natural discourse clauses containing this marker should not be analyzed as containing remote deixis, a category found in Kabyle but not in Mina.

Coreference and switch reference

Both languages have grammaticalized the distinction between coreference and switch reference. The interesting and not too difficult research question here is why both Kabyle and Mina use different means to encode the two grammaticalized meanings within this subdomain.

Unspecified reference

Both languages have grammaticalized the notion of unspecified reference. Kabyle makes a distinction between animate and inanimate, and Mina makes a distinction between human, non-human, place, and time. In addition, there is a further distinction in Mina between various types of unspecified place. Again this is consistent with the existence of locative predication in Mina.

Reference in relative clauses

Both languages encode the referential status of the head of the relative clause. Mina marks the non-referentiality of the head through the absence of a determiner after the relative clause, and referentiality through the presence of the determiner *wàcín* after the relative clause. Kabyle has a richer system, which encodes identifiability of the head: an unspecified head, and an unidentified head of the relative clause.

The research question that emerges here is why the two languages encode different categories, and why the categories are encoded by specific means. The answer to the first question appears to be related to the number of encoding means in relative clauses in each language. Kabyle has a rich system of relative and interrogative pronouns encoding the gender and the number of the head of the relative clause. Mina, on the other hand, does not have gender distinction, and the relative marker is the same for all heads of relative clauses.

This comparison of the various encoding means in Mina and Kabyle indicates a certain complementarity. Kabyle has gender and a robust number encoding on nouns; Mina does not have a gender system and productive encoding of number on nouns is limited to humans and larger domestic animals. Hence pronouns in Kabyle, which encode gender and number are a better means of encoding reference in discourse. Since Mina does not have gender and only a weak encoding of number, it has grammaticalized other means of encoding, namely the specific switch reference marker for the third person, and the deduced reference marker.

11. Conclusion

The study contributes to the theory and methodology for non-aprioristic typological research. By choosing a functional domain that is encoded in at least two of the languages studied, we can provide a systematic account of the typological differences between the languages under investigation. The differences pertain to grammaticalization of domains, grammaticalization of subdomains, and grammaticalization of specific meanings within each domain.

The proposed theory and methodology has considerable heuristic value as it indicates what should be investigated for each domain studied, in order to explain the differences among related languages. Moreover, the theory and methodology also indicates the potential cause and effect relationships for the differences among languages.

The proposed theory will be implemented in a follow-up project to CorpAfroAs, also funded by the Agence Nationale de la Recherche, entitled CorTypo.[2] The work done in CorpAfroAs will provide the basis for the retrieval of forms in context, and allows to conduct complex language-internal searches; the database will allow non-aprioristic cross-linguistic comparison among the languages of the corpus.

List of abbreviations

ASSC	associative	M	masculine
COMP	complementizer	NEG	negative
CONJ	conjunction	PAST	past tense
D	dependent	PL	plural
DEB	debitive	PRED	predicator
DED	deduced reference	PREP	preposition
DEM	demonstrative	PRF	perfective
EE	end-of-event	PRO	pronoun
F	feminine	PTCL	particle

[2] <http://cortypo.huma-num.fr/>

F.	Fula	PVG	point of view of goal
GEN	genitive	REF	referential
GO	goal	REFL	reflexive
H.	Hausa	REM	remote
HAB	habitual	SEQ	sequential
HYP	hypothetical	SG	singular

References

Bybee, Joan L., Perkins, Revere Dale & Pagliuca, William 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago IL: University of Chicago Press.

COCA - *Corpus of Contemporary American English*. <http://corpus.byu.edu/coca/>

Cohen, David. 1989. *L'aspect verbal*. Paris: Presses Universitaires de France.

Comrie, Bernard. 1976. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: CUP.

Dahl, Östen. 1985. *Tense and Aspect Systems*. Oxford: Blackwell.

Dixon, Robert M. W. 2010. *Basic linguistic theory, Vol 2: Grammatical topics*. Oxford: OUP.

Ebert, Karen. 1971. *Referenz, Sprechsituation und die bestimmten Artikel in einem nordfriesischen Dialekt* [Studien und Materialien 4]. Bredstedt: Nordfriisk Instituut.

Frajzyngier, Zygmunt. 1985. Logophoric systems in Chadic. *Journal of African Languages and Linguistics* 7: 23–37. DOI: 10.1515/jall.1985.7.1.23

Frajzyngier, Zygmunt. 1991. The de dicto domain in language. *Approaches to Grammaticalization*, Vol.1 [Typological Studies in Language 19], Elizabeth Closs Traugott & Bernd Heine (eds), 219–251. Amsterdam: John Benjamins. DOI: 10.1075/tsl.19.1.11fra

Frajzyngier, Zygmunt. 1993. *A Grammar of Mupun*. Berlin: Reimer.

Frajzyngier, Zygmunt. 1997. Pronouns and agreement: Systems interaction in the coding of reference. In *Atomism and Binding*, Hans Benis, Pierre Pica & Johan Rooryck (eds), 115–140. Dordrecht: Foris.

Frajzyngier, Zygmunt. 2008. *A Grammar of Gidar*. Bern: Peter Lang.

Frajzyngier, Zygmunt. 2010. Cross-linguistic comparison as a heuristic device: What are object pronouns good for? In *Essais de linguistique générale et de typologie linguistique*, Frank Floricic (ed.), 63–86. Lyon: ENS Éditions.

Frajzyngier, Zygmunt. 2011a. Les fonctions de l'ordre linéaire des constituants. *Bulletin de la Société de Linguistique de Paris* 107(1): 7–37.

Frajzyngier, Zygmunt. 2011b. Grammaticalization of the reference systems. In *Handbook of Grammaticalization*, Bernd Heine & Heiko Narog (eds), 625–635. Oxford: OUP.

Frajzyngier, Zygmunt. 2012. Theoretical bases for differential marking of grammatical and semantic relations of noun phrases: The proper domain for argument-adjunct distinction. Paper given at SLE Conference in Stockholm, September.

Frajzyngier, Zygmunt. 2013. Non-aprioristic typology as a discovery tool. In *Functional-Historical Approaches to Explanation: In Honor of Scott DeLancey* [Typological Studies in Language 103], Tim Thornes, Erik Andvik, Gwendolyn Hyslop & Joana Jansen (eds). Amsterdam: John Benjamins. DOI: 10.1075/tsl.103

Frajzyngier, Zygmunt. Submitted. Semantic prerequisites for typology of functional categories.

Frajzyngier, Zygmunt, Johnston, Eric with Edwards, Adrian. 2005. A Grammar of Mina. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110893908

Frajzyngier, Zygmunt & Mycielski, Jan. 1998. On some fundamental problems of mathematical linguistics. In *Mathematical and Computational Analysis of Natural Language* [Studies in Functional and Structural Linguistics 45], Carlos Martin-Vide (ed.), 295–310. Amsterdam: John Benjamins. DOI: 10.1075/sfsl.45.27fra

Frajzyngier, Zygmunt & Shay, Erin. 2003. *Explaining Language Structure through Systems Interaction* [Studies in Language Companion Series 55]. Amsterdam: John Benjamins. DOI: 10.1075/tsl.55

Hagège, Claude. 1974. Les pronoms logophoriques. *Bulletin de la Société de Linguistique de Paris* 69(1): 287–310.

Haspelmath, Martin. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology* 11: 119–132. DOI: 10.1515/LINGTY.2007.011

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3): 663–687. DOI: 10.1353/lan.2010.0021

Haspelmath, Martin, Dryer, Matthew S., Gil, David & Comrie, Bernard. 2005. *The World Atlas of Language Structures*. Oxford: OUP.

Izre'el, Shlomo. 2002. Preface. In *Semitic linguistics: The state of the art at the turn of the twenty-first century*. Shlomo Izre'el (ed.). Tel-Aviv: Eisenbrauns.

Jaggar, Philip J. 1994. The space and time adverbials NAN/CAN in Hausa: Cracking the deictic code. *Language Sciences* 16(3-4): 387–421. DOI: 10.1016/0388-0001(94)90010-8

Lazard, Gilbert. 2004. On the status of linguistics with particular regard to typology. *Linguistic Review* 21: 389–411. DOI: 10.1515/tlir.2004.21.3-4.389

Mettouchi, Amina & Chanard, Christian. 2010. From fieldwork to annotated corpora: The CorpAfroAs project. *Faits de Langues-Les Cahiers* 2: 255–265

Newmeyer, Frederick. 2007. Linguistic typology requires cross-linguistic formal categories. *Linguistic Typology* 11(1): 133–157. DOI: 10.1515/LINGTY.2007.012

NKJP -Narodowy Korpus Języka Polskiego. <http://nkjp.pl/poliqarp/nkjp-balanced/query/>

Seiler, Hansjakob. 1995. Cognitive-conceptual structure and linguistic encoding: Language universals and typology in the UNITYP framework. In *Approaches to Language Typology*, Masayoshi Shibatani & Theodora Bynon (ed.), 273–326. Oxford: Clarendon Press.