STRAND 3: Typology and dynamics of linguistic systems

Head: Katharina Haude, SeDyl; Dmitry Idiatov, LLACAN

Number of Work packages: 14

10 EFL teams involved: CRLAO, HTL, LACITO, LATTICE, LLACAN, LLF, LPP-P3, LPP-P5, MII, SEDYL

Number of participants: 91

Strand 3 investigates the diversity of the world’s languages. Most of its members are field linguists with a strong expertise in the description and documentation of lesser-studied languages. The research carried out here focuses on typology (i.e. the study of the universals and differences found in the world’s languages independently from their genealogical relation), as well as the variation in space and time found within and across areally or genealogically close languages. It involves, on the one hand, fine-grained cross-linguistic analyses of particular linguistic phenomena, which can only be carried out by experts on the individual languages concerned, and on the other hand, the investigation of the long-term and short-term effects of contact between related and unrelated languages.

Accordingly, Strand 3 is divided into two work packages. The first, “Grammar and Lexicon in Synchrony and Diachrony (GL)”, comprises nine projects (work packages) that deal with the analysis of selected grammatical and lexical phenomena, many of which have as yet been insufficiently studied from a typological perspective. GL1 investigates interjections and ideophones, categories that usually occupy a marginal place in grammatical descriptions, but that are interesting for their own sake as they are located between arbitrary and motivated language. Patterns in the semantic organization of the lexicon of diverse languages are the research topic of GL2, which investigates the cross-linguistic structure of several inadequately explored lexical fields and aims at establishing the historical trajectory of semantic change in these domains, including areal contact effects. GL3 focuses on the expression of associated motion and posture in languages from different linguistic areas and families. GL4 aims at a fine-grained typology of nonverbal predicates (“pseudo-verbs”), critically evaluating some of the proposals made in the literature to this respect. GL5 looks at noncanonical inflectional patterns underrepresented in morphological typology. GL6 takes a discourse perspective to investigate language-internal alternations between different grammatical constructions. GL7, a continuation of a project of Strand 6 (“Corpus”) of the previous LABEX-EFL, has as its aim to enlarge and refine the corpus of Mande languages. Similarly, GL8 is continuing work on the reconstruction of the Austronesian and Niger-Congo language phyla. Finally, the first work package also hosts a project dealing with the meta-level of grammatical description: GL9, which belonged to Strand 7 of the previous LABEX-EFL, looks at descriptive traditions beyond the Latin-based model. Crucially, all projects in this work package have as their data base a high number of lesser-described and partly endangered languages, and in this way will contribute greatly to a more detailed knowledge of these languages.

Of the five projects in the second work package, “Variation and Contact (VC)”, four are continuations of existing LABEX projects. VC1 continues its research on West-Asian languages, now focussing on word order and differential argument marking. VC2 takes a multidisciplinary view on the linguistic variation of the Central Gallo-Romance languages in
Central France, now also including a psycholinguistic dimension. VC3 investigates the striking speaker-dependent syntactic variation that is so frequent in creole languages. VC4, a new project, proposes a fine-grained analysis of contact between closely related languages, mostly Indo-European and Afroasiatic. VC5, finally, investigates code-switching through multifactorial analysis of plurilingual corpora.

Many of the projects in Strand 3 build on results of the first phase of the LABEX-EFL, e.g. by using databases that were created there. This empirical basis will be complemented by the creation of new databases designed for quantitative and qualitative analysis, benefiting from the expertise in more than 100 languages and language families.

The originality of Strand 3 lies in its combination of historical linguistics (macro-families reconstruction, long term typological change), contact linguistics, language documentation and typology, all grounded on both qualitative and quantitative approaches (intensive use of databases and electronic corpora to assess theoretical issues). The increasing interoperability between all these resources and the cross-checking of hypotheses in the different approaches will provide innovative scientific perspectives.
Work package 1:
Grammar and Lexicon in Synchrony
and Diachrony (GL)
1. Presentation

1.1. Abstract
We propose in this project to study interjections and ideophones: both are minor, invariable, and at least partly sound-symbolic word classes. Throughout their history, these word classes have been faced with definitional problems of a rather different nature than other word classes, in part because they are situated at the boundary of arbitrary vs 'motivated' language.

Starting with case studies, this project will take a typological, diachronic, areal and multimodal approach to the study of the semantics, morphology and syntax of interjections and ideophones.

1.2. Objectives
Interjections, invented by Latin scholars to fill a gap in the 8-part word classification after the dismissal of articles (which existed in Greek but not in Latin), are on the whole defined, throughout history, in negative terms on account of their phonological marginality, their non-participation in morphological processes and their syntactic autonomy.

Ideophones, the term for which was initially used interchangeably with 'interjection', are also difficult to define notionally, on account of the great variety of ideophonic lexemes in the world’s languages, and many definitions, as with interjections, highlight the non-standard phonology found with this word class.

We have decided to consider these two word classes within the same typological project with the hope that setting them up in opposition to each other will make us more efficient in searching for consistent (and contrasting) definitions.

Considering that in many languages, ideophones and interjections do not participate in morphological processes, it is reasonable to question whether they should, or even can be, topics in
grammatical descriptions. These words express emotions or reactions (interjections) and sensory imagery (ideophones), in other words subjective notions with great cultural variability. The difficulty in translating them and accurately capturing their meaning makes describing them difficult (for non-native linguists), marginalizing them even more compared to other word classes.

Another question of interest is that of the type of corpus which is needed to study these word classes: interjections and ideophones are more frequent in corpora of spontaneous speech, most often interactive, in other words types of linguistic production that are significantly different from those used for research into nominal or verbal morphology. Additionally, they will often only be interpretable to the fieldworker after a relatively long period in the field, adding to the complexity of their description.

Studies of interjections are certainly not a recent phenomenon, but more recent work has tended to be situated within the field of pragmatics (Ameka 1992, Cram 2008, Poggi 2009) rather than from the perspective of typology. Ideophones, on the other hand, have been given cross-linguistic consideration, through work on sound symbolism (Hinton et al. 1994), and in typology (Voeltz, F.K.E. & C. Kilian-Hatz 2001), and descriptive work (Dingemanse 2011, Reiter 2011). This work on ideophones has resulted in the following implicational hierarchy: sound < movement < visual patterns < other sensory perceptions < inner feelings and cognitive states (languages with ideophones covering a semantic field to the right will also have idophones in semantic domains to the left of that point).

Working from case studies on individual languages, this program will approach the analysis of ideophones and interjections from a typological, diachronic, areal and multimodal perspective in the consideration of the semantics, morphology, syntax and pragmatics of these word classes.

2. Bibliography

1 Presentation

1.1 Abstract
The project Lexical typology across time and space is at the crossroads between synchronic lexico-semantic typology and the diachronic study of semantic change. It aims at examining several semantic fields whose cross-linguistic structure has yet to be mapped in detail in the perspective of language change. This will be achieved through the study of cross-linguistic variations in the lexicon which show recurrent patterns and typological trends in a limited number of semantic domains (motion, impact and manipulation, cognition), using existing lexical databases and building a dedicated one. Understanding the dynamics of lexical typology is essential not only in order to explain the stability of statistical universals, but more broadly in order to shed light on one of the thorniest problems of historical linguistics: semantic reconstruction.

1.2 Objectives
Languages vary not just in their grammatical structure, not just in the form of lexical items, but also in the structure of their lexicon. Anyone speaking more than one language can easily come up with examples of cross-linguistic differences due to the internal organization of the lexicon – the way English lumps savoir and connaître together as know, the absurdity of translating idioms literally, the way a common word in one language has no lexicalized equivalent in another, the mistake in assuming that hello is used in the same contexts as bonjour... All these examples illustrate structural properties of the lexicon – i.e. aspects of a language’s semantic organization that are stored in its lexical component, and that can be compared across languages.

The lexicon is often defined in terms of unpredictability, by contrast with the regularities of grammar. However, unpredictability does not mean randomness. As suggested by the growing body of typological work summarised in Koptjevskaja-Tamm, Rakhilina and Vanhove (2015), cross-linguistic variation in the lexicon does show regularities and exceptions, recurrent patterns and typological trends, all of which are as much worthy of scientific scrutiny as the more often studied domains of grammar. Works such as Koptjevskaja-Tamm (2015), and even a special issue of Linguistics (50:3, 2012), illustrate a promising trend of mapping out the cross-linguistic typology of particular semantic domains, while works such as Urban (2011, 2013) show
that synchronic cross-linguistic data on colexification (François 2008) provides diachronic information on the nature and directionality of semantic change. Databases such as CLICS (http://clics.lingpy.org) and DatSemShifts (http://semshifts.iling-ran.ru/) illustrate the new possibilities opened up in this domain by digital methods. Meanwhile, semantic elicitation methods have advanced considerably, as illustrated by work on topics such as SMELL (Majid & Burenhult 2014).

Despite this significant progress, however, the literature available on the dynamics of lexical typology – a crucial aspect of semantic change – tends to be fragmentary, interested in one language area, or one semantic field, at a time. This scientific landscape contrasts sharply with the more fully developed theories available for phonological change or for grammaticalization. The time is ripe to take our understanding of the dynamics of lexical typology further.

1.3 Aims and methods

Our project will build on the results of the TUL project "Towards a typology of semantic associations" (2002-2009) directed by Martine Vanhove, as well as on recent advances in the domain. We also intend to make use of existing dictionaries – including the online dictionaries currently under development at LACITO – as well as online databases, such as the RefLex database for African languages at LLACAN. Based on these materials and more gathered in the field, we intend to examine several semantic fields whose cross-linguistic structure has yet to be mapped in detail in the perspective of language change. Understanding the dynamics of lexical typology is essential not only in order to explain the stability of statistical universals, but more broadly in order to shed light on one of the thorniest problems of historical linguistics: semantic reconstruction. This project will result in a better understanding of the forms taken and paths followed by semantic change within the lexicon.

Without wishing to restrict ourselves to a single semantic domain, we intend to explore in depth semantic domains of particular interest, including verbs of motion and associated motion (go, come, bring, carry, hold, send…), in association with Guillaume Jacques and Patrick Caudal’s working group on associated motion; verbs of impact and manipulation (put, take, shift, twist, raise, remove, throw…), extending the 19-language sample of Kopecka & Narasimhan (2012) to a more balanced range of languages and a wider range of notions; and verbs of cognition (know, feel, believe, understand, forget etc.).

1.4 Expected outcomes

This project will map out the cross-linguistic structure of several inadequately explored lexical fields, and establish as far as possible the historical trajectory of semantic change in these domains, while identifying areal contact effects. The results of the project will be published in the form of a peer-reviewed edited volume and a database.

1.5 Projects and contracts submitted in relation to one of the topics of the operation

Fedden, Sebastian. (demandé). History and Typology of Central New Guinea Languages. FNS-ANR.


GL3:
Associated Motion and Posture

<table>
<thead>
<tr>
<th>PIs</th>
<th>Patrick CAUDAL (LLF), Guillaume JACQUES (CRLAO)</th>
</tr>
</thead>
</table>
| Participants               | LACITO: Isabelle Bril (DR), Claire Moyse-Faurie (DR émérite), Paulette Roulon (DR), Guillaume Jacques (DR)  
LLF: Patrick Caudal (CR), Guillaume Fon Sing (MCF), Berthold Crystmann (LLF, CR),  
LLACAN: Marie-Claude SIMEONE-SENELLE (DR)  
CRLAO: Christine Lamarre (PRF Inalco), Tatiana Yakovleva (MCF U. Paris-Diderot)  
Lattice: Laure Sarda (CR), Benjamin Fagard (CR)  
HTL: Aimée Lahaussois (CR) |

1. Presentation

1.1 Abstract
This research project will be devoted to the typology of associated movement (Koch 1984; Wilkins 1991) and associated posture (Enfield 2002), with an emphasis on areal typological approaches (Austronesian, Australian, Sino-Tibetan languages, Niger-Congo, in particular). Associated posture and associated movement will be understood here in the broadest possible sense, i.e., ranging from dedicated lexical classes to specifically grammatical phenomena - whether related to particles, clitics, or affixes. Relations between associated posture/movement and description of events will be more specifically targeted. In many languages of the world, in fact, these categories are linked e.g. to serial verb systems, TAM auxiliaries/verbal periphrases, etc.

1.2 Objectives
It has been established that associated movement and posture markers frequently develop new functions, thereby forming complex grammatical (or lexico-grammatical) microsystems. Such systems sometimes combine markers (or former markers) of movement and associated posture, as e.g. in Arnhem Land languages (cf. e.g. Mawng (Singer 2006), Iwaidja (Caudal & Mailhammer 2016), Anindhilyakwa (Bednall to appear), Ndjébana (McKay 2000), or in some Amazonian languages, (Vuillermet 2012; Vuillermet 2015)), or add markers belonging to other notional micro-classes. Such grammatical systems can be endowed with discourse-structural functions (Bourdin 2008), or TAME functions (giving rise to tense / aspect meanings, or forming novel discursive connectors ...) – but also develop towards deixis, or voice categories e.g. the passive.... See (Newman 2002) or(Devos & van der Wal 2014) for some illustrations of such developments. The crosslinguistic consistency of these lexical and grammatical developments demonstrates the deep grammatization tendency (cf. Bourdin 2014: 119)) potential of such categories. This fact seems to take on a particular importance in the light of recent experimental work highlighting the intimate relationships linking postural perception, perception of movement and linguistic learning, cf.(Pfeifer 2006; Benitez & Smith 2012; Morse et al. 2015). In turn, this suggest that the categories of associated motion and associated posture offers vast...
and fecund research perspectives, involving fundamental connections between language, cognition/self-perception, and the world.

1.3 Projected activities
The new operation will first focus on collecting relevant data across as many relevant linguistic areas and phyla as possible; experimental and quantitative methods will be regarded as priority, whenever possible. It should be highlighted that the relevant data remains significantly underdescribed across most linguistic areas / phyla. Novel descriptive and (areal) typological results will be hopefully produced. Temporary contracts, as well as an engineer position to be shared with other operations will be considered.

This operation shall begin in 2020, and will capitalize on the work previously conducted within the GD4 operation (Strand 3), as well as through the ‘EvEs’ (Événements et Espace) Fédération TUL (Typology & Universals Linguistics Federation) project (2018; coordinators: P. Caudal (LLF) & M. Vuillermet (DDL)). The ultimate goal of this project is to help prepare an ANR project for 2021-2022.
1 Presentation

1.1 Abstract
The present project proposes to study for the first time the particular category of “pseudoverbs” from a cross-linguistic and typological perspective so as to deepen our understanding of the nature and functioning of the different part-of-speech categories in the world’s languages. In order to do so, the project puts forward to verify certain concepts from Role and Reference Grammar (RRG), which offer compelling alternatives to clarify the grammatical relations involving pseudo-verbs.

1.2 Objectives
Pseudo-verbs form a specific part-of-speech category of all sorts of non-verbal items dedicated to predication and displaying interesting and challenging morphosyntactic features relevant to theoretical and typological linguistics, which relate to the following primary issues:

First, even though none of the pseudo-verbs are “proper” verbs, all of them concern predication or co-predication. While sharing some features with verbs and co-verbs, like for instance having TAM or specialized pronominal paradigms or specific inflections, pseudo-verbs are in essence distinct grammatical forms with specific functions (see, for instance, in Berber: Aikhenvald 1995, Lafkioui 2001, 2011, Satzinger 2005, and in Semitic: Brustad 2000, Huehnergard 1987).

Second, the diachrony of pseudo-verbs shows interesting grammaticalisation phenomena relating to their development mostly out of non-verbal items which took on verbal forms and functions, such as the prepositional predicate in Berber and in Arabic which conveys ‘to have’ (e.g. with-3sg child ‘he has a child’). Other regular pseudo-verb developments concern demonstratives, locatives, adverbs and pronouns. Furthermore, certain languages show how some of these developments are cyclic in nature, since they are evolving in the reverse direction as they are losing some of the verbal features (de-grammaticalisation), causing a kind of cyclic movement (e.g. Berber, Lafkioui 2011).

Third, the pseudo-verbs cast doubt upon the use of traditional grammatical concepts like “subject” and “object” for their morphosyntactic relations, as the “logical subject” of this kind of configurations is marked in a different way in certain languages and contexts (e.g. Berber,
Arabic), often even matching with the “object position” in verb sentences, at the same time as having many of the morphosyntactic characteristics of subjects. The Unaccusative Hypothesis (Perlmutter 1978) might shed some light on certain aspects of this particular grammatical relationship.

The present project proposes to study for the first time these major issues relating to pseudo-verbs from a cross-linguistic and typological perspective so as to deepen our understanding of the nature and functioning of the different part-of-speech categories in the world’s languages. In order to do so, and hence to address the pseudo-verb category from a typological perspective, the project puts forward to verify certain concepts from Role and Reference Grammar (RRG), such as the concept of “macrorole” (Van Valin 2005, 2007), which offers compelling alternatives to clarify the grammatical relations involving pseudo-verbs.

1.3 References


1.4 Work plan

- A regular seminar (three sessions per year) to present and discuss the data and methodological and theoretical issues concerning pseudo-verbs.
- The study will be mainly based on the data and expertise of the participants. A small corpus sample of elicited data obtained from native informants, who will be paid for the task, will also be constructed for comparative purposes.
- 2 international workshops with invited speakers (2\textsuperscript{nd}, 3\textsuperscript{rd} year)
- 1 international conference which will be held in the 4\textsuperscript{th} year.

1.5 Output

- Paper presentations at seminars, workshops and conferences organized by the group as well as by other research groups and institutions.
- Publication of joint articles in highly classified journals
- 1 collective volume dedicated to the topic of the project with contributions of its member
1 Presentation

1.1 Abstract
The purpose of the project « Typology of inflectional systems with non-canonical inflection » is to create better typological knowledge of inflectional complexity in the world’s languages by studying non-canonical inflectional phenomena on new data coming from less-described languages and by adopting new approaches to better known morphological systems. The goal is to combine typological, formal and quantitative approaches for the system of inflectional systems. Empirical subjects of particular interest are non-concatenative flexion, the continuum between synthesis and periphrasis and the emergence of non-canonical flexion, as well as the emergence of uninflectedness.
1.2 Objectives

Theoretical: Create typological knowledge of inflectional complexity by the study of deviations from «canonical inflection» (Corbett, 2007), following approaches to morphology inspired by theories of morphology characterised as «word-and-paradigm» (Anderson, 1992; Aronoff, 1994; Baerman et al., 2015; Matthews, 1991; Hippisley and Stump, 2016; Stump, 1991, 2016; Stump and Finke, 2013; etc.). We will pay particular interest to the study of phenomena related to four topics (but there could be others), which can show a great deal of interconnection, namely: uninflectedness, non-concatenative inflection, the periphrasis-synthesis continuum and the emergence of non-canonical inflection.

Applied: Create a space to allow both inflectional morphologists in Paris and language experts with an interest in typological linguistics to exchange ideas about inflectional issues aimed at generating a better understanding of such phenomena from a typological point of view so that we can gain better descriptions.

1.3 Work plan

- Monthly seminar (open to MA and PhD students) to discuss the main theoretical issues about non-canonical inflection, discuss new data from different languages or revisit known data from new perspectives.
- Monthly project meeting next to the monthly seminar to discuss management and progress of the project.
- Organization and participation in workshops and publications
- Field trips to collect relevant data and creation of databases of inflectional morphology on poorly described languages.

1.4 Outcomes and dissemination

- Creation of an on-line database with new inflectional data on lesser-described languages.
- Individual and collective publications on specific inflectional systems in different languages with a variety of approaches
- Organisation and participation in workshops

1.5 References cited

Corbett, Greville G. 2007, Canonical typology, suppletion and possible words. Language 83, 8-42.
1 Presentation

1.1 Abstract
Based on discourse corpora of genealogically and typologically diverse languages, we will investigate the grammatical constructions that speakers actually use when describing events – in particular, events involving two or more participants – in a larger context. This includes the analysis of verbal voice alternations, the ways in which event participants are referred to (or not), and the linear order of clausal constituents. Particular attention will be paid to factors such as animacy, definiteness, and different types of topicality of discourse referents, as well as to the employment of verbal categories (voice, aspect) used to direct the hearer’s attention.

1.2 Objectives
The morphological, syntactic, and semantic properties of the different constructions of an individual language are generally well understood by the participating language experts. This programme’s goal is to get a better idea of the conditions under which a particular construction or argument expression is used. Our research basis will be data from naturally occurring speech, also because structured data like those resulting from psycholinguistic experiments are usually not available for the languages under study here. We will use annotated corpora of spontaneous oral discourse which the project members already dispose of. They stem from typologically and genealogically diverse languages, which have different grammatical possibilities to allow a variable presentation of an event and its participants. The project is a partial continuation of the Axe3/GD1 project of the previous Labex EFL, “The Typology and Corpus Annotation of Information Structure and Grammatical Relations”, at which most of its members participated, and relies to a large degree on the corpora annotated there.

As a starting point, we will analyze how previously identified event representation patterns in a particular individual language are distributed in actual discourse. This will necessarily involve the identification of discourse-structuring strategies in the respective
languages. Once the criteria that may influence the choice of a construction are identified, the functions of the morphosyntactic strategies used by one language can be compared with those of others (e.g. voice alternations; overt vs. covert vs. zero argument expressions; elaborate event vs. elaborate participant descriptions), while bearing in mind the overall morphosyntactic system and the grammatical possibilities of each language.

International collaborations are envisaged with the Collaborative Research Center (SFB) 1252 “Prominence in Language” at the University of Cologne (dir. Klaus von Heusinger), with Stefan Schnell at Melbourne University and his colleagues who developed the GRAID corpus annotation strategy, and with Eva Schultze-Berndt of Manchester University who works on the corpus-based analysis of reference-tracking strategies in lesser described languages.

The planned outcome of the project consists in single-authored articles, at least one co-authored article by project members, and a collective volume with contributions by project members and collaborators.
1. Summary of the first phase

1.1. Activities since 2015

The main efforts were directed to the building of new corpora for Mande languages and to the increasing the Bambara Reference Corpus (which exists on line in open access since 2011), cf. http://cormand.huma-num.fr/mandeica/

Bambara Reference Corpus

In January 2015, the Bambara Reference Corpus included 2 165 534 words, of these 339 699 in the manually disambiguated subcorpus. In April 2018, it reached 5 220 107 words, of these 1 075 790 words in the disambiguated subcorpus. A program of partial automatic disambiguation collocation-based program has been developed, it allowed to decrease the ambiguity rate in the main (non-disambiguated manually) subcorpus from about 70 % to about 30%. Some studies have been carried out on automatic statistical POS and tone-disambiguation; the results of these studies are being integrated into the program of morphological analysis of Bambara texts.

In April 2018, a Bambara-French parallel corpus has been published on line. At the moment of the first publication, it counted some 100,000 words in each part (French and Bambara); further texts for more than 100,000 words are ready for integration.
Also in April 2018, a syntactically annotated subcorpus (according to the Universal Dependencies model) has been published. Its starting volume is about 1000 phrases; further texts are being treated.

**Maninka Reference Corpus**

Maninka Reference Corpus was published in April 2016. It is composed of two subcorpora: in N’ko (an original writing system Manding languages created in 1949), some 3,100,000 words, and in Latin orthography, about 400,000 words. Since that time, the main effort was directed to the improvement of the automatic annotation, in particular to the cleaning and completion of the lexical database (an electronic Maninka-French dictionary, Malidababa).

**Eastern Dan corpus**

Published in April 2018, this corpus includes texts in a language spoken by some 600,000 people in Côte d’Ivoire. At the beginning, it counted some 20,000 words; it is planned to increase it soon to some 300,000 words.

Before the end of 2018, it is planned to publish on line corpora for 3 or 4 other Mande languages (among the nearest candidates are Mwan, Mano, Kakabe, and Kpelle languages).

On the base of the Bambara and Maninka corpora, orthography spellcheckers for these two languages have been developed (for Open Office).

Electronic libraries for Bambara and Maninka have been created and put online, in cooperation with Malian Academy of Languages (Bambara) and Koffi Annan University & N’ko Academy (Maninka).

**1.2. Publications**


2. Presentation of the follow-up project

2.1. Abstract

The operation "Text corpora for Mande languages" is reaching a stage of maturity. Main electronic tools necessary for the functioning of the parallel corpora, syntactically annotated corpora and audiocorpora developed in the first phase will facilitate, on the one hand, the growth of the existing
text corpora and improvement of their quality and, on the other hand, to create quickly text corpora of additional Mande languages and, eventually, of languages from other families. In the new phase, the NLP aspects of the project will be emphasized, in particular: automatic (statistically-based) disambiguation; automatic clusterization of polysemy (on the basis of vectorial analysis); development of OCR tools for the concerned languages; furthermore, first attempts of speech recognition and automatic translation problematics will be made.

2.2. Objectives
During the second phase of the project, three types of works are planned:
1) Increase of the existing corpora. In particular, it is intended to bring the Bambara Reference Corpus to 8-10 million words. This corpus will include most of the existing literature and press available in this language. Automatic disambiguation of the entire Corpus will be carried out. The statistically (manually) annotated subcorpus will reach 10,000 phrases; based on this, automatic syntactic annotations with an acceptable error rate will be carried out for the entire Bambara Reference Corpus. The parallel Bambara-French/French-Bambara corpus will reach 1,000,000 words (in each part). An annotated and synchronized audio corpus of Bambara will be built. The Maninka Reference Corpus will reach at least 5 million words. A disambiguated Maninka subcorpus will be launched; by 2024, it will have reached 400,000 words. A parallel Maninka-French/French-Maninka corpus will be built and reach 200,000 words. An audiocorpus of Maninka will be created. The Eastern Dan corpus will reach 500,000 words, of which some 15% will constitute the disambiguated subcorpus. A parallel Dan-French/French-Dan corpus will be built (target size of 100,000 words).

2) Compiling new corpora. Among the big Mande languages, Jula of Côte d'Ivoire (spoken by more than 10 million people) is the first candidate. A Ivorian Jula corpus will be built according to the same model as the Bambara Reference Corpus; it will be based on a 12,000-word Jula-French dictionary produced by Ivoirian colleagues, external collaborates of the project. A series of small corpora for Mande languages will be built: apart from Mwan, Guinean Kpelle, Kakabe and Mano, there are good chances to build corpora for Guro, Gban, Jogo, Kla-Dan. For these languages, the size of corpora will vary between 50,000 and 300,000 words, depending on the amount of the texts available. For each language, furthermore, a parallel corpus and an audio corpus will be built. A Soninke corpus may also be built in collaboration with colleagues from Hamburg and Lyon. Our operation is open for cooperation with other colleagues willing to build corpora for their respective languages with our informatics tools.

3) Application of NLP methods to the corpora will be enhanced: automatic disambiguation of texts (parts of speech and glosses); application of distributive analysis to the sense discrimination; advanced studies in syntax will be carried out; tools for OCR and automatic spellchecking for the concerned languages will be developed; first approaches to automatic translation and speech recognition will be undertaken.

2.3. Evolution with respect to the first phase
The corpora of the big languages (Bambara, Maninka) will be brought to maturity, so that they can be used both in linguistic studies (fine-grained grammars; lexicological studies) and in practical applications. It is intended to simplify the process of building annotated corpora for Mande languages (and, eventually, for languages of other families), which will serve to advance corpus linguistics for African languages.

2.4. Projects and contracts submitted in relation to the project
It is planned to submit a project directly connected with our operation for an ERG grant in 2019.
1. Summary of the first phase

1.1. Activities since 2015

At the first stage of Labex, a great effort was made to create an international network of researchers for the reconstruction of Niger-Congo. With the support of Labex, members of the group organized two international congresses "Towards Proto-Niger-Congo: Comparison and Reconstruction" in 2012 and in 2016, and launched a series "Niger-Congo Comparative Series" in Language Science Press. During this period, the work of the Niger-Congo team turned around the lexical database Reflex, http://reflex.cnrs.fr/database/index.php, which exceeded the limit of 1,100,000 words in 2017. Apart from the quantitative growth, Reflex has evolved qualitatively too. It represents today a powerful and comfortable tool for comparative studies, in open access.

An important result for the Niger-Congo reconstruction is the publication of a book on reconstruction of Niger-Congo numerals (Pozdniakov 2018, Language Science Press). A Niger-Congo etymological base (K. Pozdniakov, G. Segerer) has been built in Reflex. Considerable results have been reached in the reconstruction of the proto-languages of the Atlantic and Mande families. For the Mande family, the publication of the volume Mande Languages (1150 pages, edited by V. Vydrin) was an important milestone.

In the period 2009-2018, the Austronesianists in our team have worked both on the synchronic description of the languages of our expertise, and on what can be reconstructed of their history (see the bibliography, under Bril, François, Sagart).

Isabelle Bril has been conducting fieldwork on one northern New-Caledonian language, Zuanga-Yuanga, and working on a dictionary and grammar of those two divergent dialects of the same
language. I. Bril is also documenting and analysing northern Amis, a Formosan language of the east coast of Taiwan (Bril 2016). Alexandre FRANÇOIS is compiling a dictionary of Mwotlap (Banks, Vanuatu), of which a current draft is already available online; he expects to complete it by 2020. Another well advanced descriptive project is a dictionary of Teanu (Vanikoro, Solomon); he plans to have it ready for publication in 2019.

Major achievements in our group include the following:
- participation in a special issue of 2017, Studies in Language, on Lexical flexibility in Oceanic languages (ed. by E. van Lier)
- publication of the Linguistic Atlas of French Polynesia (Charpentier & François 2015)
- a collective volume The Languages of Vanuatu: Unity and Diversity (François et al. 2015)
- the creation of two open-access academic series in the Austronesian domain: Studies in the Languages of Island Melanesia (2012–) and Studies in the Languages of Wallacea (2017–), both within Asia-Pacific Linguistics.
- a contract has been signed with OUP for editing the Guide to the Malayo-Polynesian languages of Southeast Asia by Antoinette Schapper and Sander Adelaar.
- a volume Contact and substrate in the languages of Wallacea, ed. by A. Schapper.

1.2. Publications


Valentin Vydrin, Mamadi Diané. L’interrogation en maninka de Guinée. Mandenkan 56, 2016, pp. 95-117. halshs-01481035

Valentin Vydrin, Mamadi Diané. L’interrogation en maninka de Guinée. Mandenkan 56, 2016, pp. 95-117. halshs-01481035


2. Presentation of the follow-up project

2.1. Abstract
The rising prominence of typological approaches to the study of language universals and of evolutionary approaches to explanation in linguistics entails a renewed interest in the classification of the languages of the world as well as in the reconstruction of their proto-languages. This in turn requires careful lexical and syntactic analysis of a great sample of languages based, as much as possible, on first-hand data collected by specialists.

This project aims at a better understanding of the history and internal genealogy of the two biggest language phyla of the world, viz. Niger-Congo and Austronesian. Bringing together some of the world’s leading experts, we will strengthen the empirical basis of comparative work in these phyla by describing and documenting individual languages, and we will propose better classifications and genealogies. In every domain of our research project (description, classification and reconstruction), specialists of Austronesian and Niger-Congo will exchange methods and insight from their own field.

2.2. Objectives
In the next two years, I. BRIL is planning a collaboration with Stavros Skopeteas on the prosody of Amis, in particular in relation to informational hierarchy. Collaboration with Elizabeth Zeitoun (Academia Sinica, Taipei) is expected to start in 2020 on a joint new project on Sakizaya, a neighbouring, nearly extinct Formosan language, with some comparative aims. Various articles, a grammar and a dictionary of northern Amis are planned for publication from 2019 onward. Some additional fieldworks are planned for Amis and Sakizaya.

Alexandre FRANÇOIS plans to finish writing a reference grammar of Teanu (Vanikoro, Solomons) in 2022. The reference grammar of Hiw (Torres, Vanuatu), is planned for publication in 2024.

Antoinette SCHAPPER, who just joined the LACITO team, will be joining the RT1 group in 2018. Her work aims to describe and explain the typology of the languages of Wallacea, at the crossroads of the Southeast Asian and Melanesian worlds. She has worked on documenting Papuan and Austronesian languages in the Timor region and on the Austronesian languages of the Aru Islands, south of New Guinea.
Valentin Vydrin is planning to finish writing reference grammars of Guinean Maninka and Eastern Dan languages, and also to work on dictionaries and annotated text corpora for these languages and Bambara. He is also planning to coordinate work on completion and publication of a set of dictionaries for Southern Mande languages (Mwan, Guro, Mano, Kla-Dan) and Southwestern Mande languages (Guinean Kpelle, Looma).

Guillaume Segerer will continue collecting data for underdescribed Atlantic (and also ex-Atlantic) languages, in particular, Jola varieties and Sua.

Pascal Boyeldieu, Dmitry Idiatov, Mark van de Velde and their PhD students will continue collecting data on Adamawa languages, and Nicolas Quint, on Kordofanian languages (Heibanian and Koalib).

2.3. Evolution with respect to the first phase

The new Labex project will put to work the data accumulated in the RefLex database and other electronic sources developed at the first stage. It is planned to carry out reconstructions of phonology and morphology of protolanguages at various levels (proto-languages of groups, of mid-range families, and, finally, of Proto-Niger-Congo). Electronic annotated corpora of texts will serve a major support to this reconstruction.

Among the tasks of that stage is a verification of the hypothesis on belonging of the Mande family to Niger-Congo (recently put in doubt by some scholars). A very prospective direction is a study of the regular homophony by K. Pozdniakov and G. Segerer on the base of RefLex that opens a path to a reliable semantic reconstruction of proto-languages and to a search of regular phonetic correspondencies. Another major direction will be the verification of the validity of the Adamawa family (whose splitting into two or more mid-range families cannot be excluded) and its link to the Gur family.

Compared to other language families, the Austronesian phylum – including the Oceanic subfamily within it – has been relatively well studied by historical linguists, since the end of the 18th century. While the comparative method has proven its efficiency, a question that remains is the degree to which the internal genealogy of the Austronesian family really does fit the structure of a tree. Most historical scholars assume that the Tree model provides the best approach to reconstructing genealogy; this is especially the case among Austronesianists. Building on a new model called Historical glottometry (François 2014, 2017; Kalyan & François 2018), the new RT1 project (2020-2024) should help us question further which model is most suitable to account for the history of Austronesian family.

The new project will also investigate interferences between Papuan and Austronesian languages. Antoinette Schapper will use the significant documentary materials that exist on the West Papuan languages to develop a comparative database of West Papuan languages. This will permit a preliminary reconstruction of the phonology, the morphosyntax and the lexicon of Proto-West Papuan, giving us a picture of a Papuan language that was probably one of the first encountered by speakers of pre-Proto-Oceanic and earlier. In turn, this will enable the first principled, systematic comparison between a reconstructed Papuan language before Austronesian influence and reconstructed members of the Austronesian language family.
GL9: Extended Grammars

<table>
<thead>
<tr>
<th></th>
<th>2015 – 2019</th>
<th>2020 - 2024</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strand</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Work package</td>
<td>EM1c</td>
<td>GL9</td>
</tr>
<tr>
<td>Project title</td>
<td>Extended grammars</td>
<td>Extended grammars</td>
</tr>
<tr>
<td>PI</td>
<td>Émilie AUSSANT</td>
<td>Émilie AUSSANT</td>
</tr>
</tbody>
</table>
LabEx members: SeDyL: A. Donabedian (PU)
HTL: M. Farina (CR), S. Auroux (émérite, HTL), É. Aussant (CR), A. Lahaussois (CR), C. Puech (PU émérite)
LaCITO: F. Jacquesson (DR émérite)
LLACAN: T. Nikitina (CR), N. Quint (DR), M. Vanhove (DR)
MII: P. Samvelian (PU, Paris 3),
LLF: C. Donati (PU), B. Colombat (PU émérite)

1. Summary of the first phase

1.1. Activities since 2015

Workshops:

a) In collaboration with Aimée Lahaussois, organisation of two workshops « Grammaires étendues » on November 3-4, 2016 (November 3rd: « Extended Sanskrit Grammar », speakers: É. Aussant, N. Balbir, J.-L. Chevillard, G. Ciotti, V. D’Avella, A. Griffiths, D. Patel (the proceedings of this workshop were published as n° 39.2 of HEL); November 4th: « Descriptive Linguistics », speakers: A. Lahaussois, F. Jacquesson, M. Vanhove, T. Nikitina, R. Lambert-Bretière, N. Quint (the proceedings will be published as a special issue of Faits de langues).


c) In collaboration with Jean-Luc Chevillard organisation of a workshop « Extended Grammars » at the conference ICHoLS XIV in Paris in 2017 (22 proposals received, 17 presentations, 7 articles as a result of these presentations will be selected and published in a special issue of Beiträge zur Geschichte der Sprachwissenschaft in 2019).


e) organisation of a planned workshop "Grammaire grecque étendue", to take place in November 2018. Provisional liste of participants: S. Archaimbault, G. Bohas, M. Farina, C. Conduché (to be completed).

Submitted projects:

a) In line with the call for projects for the ERC Consolidator Grant 2017, E. Aussant (PI) submitted the project TransTools (« How human languages have been described - A global history of transfers »). The project was not selected.

b) Aussant submitted a project "Typologie et terminologie" for the period 2019-23 of the Fédération de recherche TUL, which was accepted. As the Fédération will not continue, the project will form part of the project "Grammaires étendues" of the new Axe 3 of LABEX EFL.

c) In line with the call for proposals for the ERC Synergy Grant 2017, Aussant has submitted, as on of three PIs1, the project MultiLingICA ("Translation, Adaptation, Amalgamation: Multilingual Interactions in the Indic Cultural Area"), which is still in course of evaluation (in Phase 1 the project has been accepted).

d) Aussant has submitted a doctoral research project within the campaign « contrat doctoral avec mobilité internationale 2018 de l’INSHS », entitled « Extended Sanskrit Grammar and the description of X [Indian language] ». The project was not selected.

Participations at various conferences

1.2. Publications


---

1 PI 1 coordinateur: Eva Wilden (Univ. d’Hambourg), PI 2: Harunaga Isaacson (Univ. d’Hambourg), PI 4: Hugo David (EFEO Pondichéry).
2. Presentation of the new project

2.1. Abstract
The history of language description can be seen as a major, more or less successful, enterprise of adapting linguistic tools that were initially shaped to describe or represent a (very) small amount of languages: of the many thousands of languages known in the world, only a few have produced « reflexive » linguistic tools (i.e. tools which describe the language in which they are composed); this is the case for Arabic, Chinese, Greek and Sanskrit. Spearheaded by S. Auroux, a number of works have been carried out on the transfer of the Latin grammatical model onto European vernaculars and « exotic » languages – a phenomenon called « Extended Latin Grammar » by Auroux. It is now appropriate to 1) document widely the other cases of « extended grammars »; 2) test the commensurability of these different transfer phenomena; 3) investigate and reconstruct the long and complex history of linguistic terminology.

2.2. Planned activities
1) Documentation of the various « extended grammars »: This will take the form of workshops with proceedings. Directly linked to the Extended Sanskrit Grammar is the project MultiLinglICA, which has been submitted to the ERC (Synergy Grant 2017) and is still under evaluation, as well as a PhD project (proposed for an INSHS PhD grant with international mobility, but not selected).
2) For a better understanding of the history of language description all around the world, it is necessary to test the commensurability of the different transfer phenomena, which has never been done. A re-submission to the ERC of the project TransTools, which precisely aimed at comparing the different transfer phenomena, is planned.
3) The reflection on the transfer of descriptive models and terminologies meets some contemporaneous descriptivists’ and typologists’ concerns: if one wants to grasp all the issues addressed by a number of questions today in descriptive linguistics and in linguistic typology, and if one wants to contribute meaningfully to these debates, one has to investigate carefully the linguistic terminology in order to reconstruct its long and complex history; this inevitably involves a history of terminological transfers. Activities carried out in this direction are those which had been planned within the TUL Federation.

2.3. Changes with respect to the previous period
Two changes: 1) the comparison of the different transfer phenomena, 2) the articulation with some issues currently addressed in descriptive linguistics and linguistic typology. The goals set at the beginning of the program (in 2015) were primarily linked to the documentation of the various extended grammars.

2.4. Submitted and/or obtained projects/contracts linked to the program
Documentation of the various « extended grammars »:
a) Project MultiLinglICA (submitted to the ERC, Synergy Grant 2017, still under evaluation).
b) PhD research project "Extended Sanskrit Grammar and the description of X [Indian language]" (submitted to the INSHS in 2018, not selected).

Testing the commensurability of the different transfer phenomena: project TransTools (submitted for an ERC Consolidator Grant 2017, not selected).

Investigation into linguistic terminology: project « Typology and terminology » (submitted–and selected– for the 2019-23 mandate of the TUL Federation).
Work package 2:  
Variation and Contact (VC)
## VC1: Languages, Dialects and Isoglosses of West Asia

<table>
<thead>
<tr>
<th></th>
<th>2015 – 2019</th>
<th>2020 - 2024</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strand</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Work package</td>
<td>LC3</td>
<td>VC1</td>
</tr>
<tr>
<td>Project title</td>
<td>Languages, Dialects and Isoglosses of the Iran-Caucasus-Anatolia-Near-East Area</td>
<td>Languages, Dialects and Isoglosses of the West Asia</td>
</tr>
<tr>
<td>PIs</td>
<td>Anaïd Donabédian, Pollet Samvelian</td>
<td>Anaïd Donabédian, Pollet Samvelian</td>
</tr>
</tbody>
</table>
| Participants | LABEX members
Victoria Khurshudyan (SeDyL)
Jonny Cheung (MII)
Agnès Korn (MII)
Anton Antonov (CRLAO)
Fida Bizri (SeDyL)
Emmanuel Giraudet (MII)
Matteo De Chiara (MII)
Pollet Samvelian (MII)
Anaïd Donabédian (SeDyL) | LABEX members
Victoria Khurshudyan (SeDyL)
Agnès Korn (MII)
Anton Antonov (CRLAO)
Fida Bizri (SeDyL)
Emmanuel Giraudet (MII)
Matteo De Chiara (MII)
Pollet Samvelian (MII)
Anaïd Donabédian (SeDyL) |
| Post-doc with Labex contract: | Thomas Jügel | End of Phd thesis, Post-docs
Pinar Karakilcik (SeDyL)
Hélène Gerardin (membre individuel)
Katherine Hodgson (SeDyL)
Masud Mohammadirad (MII) | |
| Doctorants: | | Expected PhD candidates
Arayik Hairiyan (SeDyL) |
| PhD defended during the project: | | |
| Hélène Gerardin (membre individuel) | | |
| Ophélie Gandon (MII) | | |
| Pegah Faghiri (MII) | | |
| PhD defense expected: | | |
| Katherine Hodgson (SeDyL) | | |
| Pinar Karakilcik (SeDyL) | | |
| Masud Mohammadirad | | |
| Collaborations extérieures: | | |
| Hosep Dolatian | | |
| External participants: | | |
| Nina Sumbatova (ex- chaire EFL) | | |
| Yuri Koriakov (Académie des Sciences de Russie) | | |
| Georgi Moroz (HSE, Moscou) | | |
| Ioanna Sitaridou (Cambridge) | | |
| Christiane Bulut (Nicosie) | | |
| Pegah Faghiri (U. de Cologne) | | |
1. Summary of the first phase

1.1. Activities since 2015

Three topics have been investigated by the participants of the operation since 2015:

1) The diversity of relative constructions in the languages of the area has been one of our main topics of investigation since 2015:

- We followed up on the series of international workshops initiated in 2013 by organizing 4 new workshops on July 3, 2015, December, 10-11, 2015, December 12, 2016, October 12, 2016.

- One of the major outcomes of the operation was the PhD thesis of Ophélie Gandon (Doctoral funding, Université Sorbonne Nouvelle), entitled *La relativisation dans une perspective aréale: l’aire Caucase – Anatoe de l’Est – Iran de l’Ouest*, defended on December 12, 2016. The thesis was a thorough overview of relativization strategies in the languages of the area and also proposed a questionnaire including more than 90 parameters and features in order to describe various aspects of relative constructions.

- Katherine Hodgson (International doctoral funding Inspire – Marie Curie) has almost completed her PhD on Subordination in Armenian dialects: an experimental study from an areal-typological perspective. This is an extensive investigation of relative constructions in spoken and colloquial Armenian, combining field work and experimental methods. The defense is planned for February 2019.

- A volume on relative constructions in the languages of West Asia is in preparation. It will include a selection of the talks given during our international workshops and also a few chapters by renowned specialists of the languages of the area.

- About 19 presentations at workshops and conferences.

2) Several participants of the operation investigated clitic pronouns:

- Thomas Jügel was hired as a postdoctoral researcher. He investigated the evolution of clitics from pronouns into agreement markers in West Iranian languages. He co-authored 2 papers with Pollet Samvelian: an article in Bulletin de la Société de Linguistique de Paris in 2016 and a book chapter in a volume edited by Richard Larson. There were also several presentations at conferences.

- Masud Mohammadirad obtained an international doctoral funding, Cofund - Marie Curie, in 2016 to work on clitics in Western Iranian languages.

3) Cartographic modeling of language contact. A training session was organized by the operation on January 19, 2018, in order to initiate our participants to the cartographic package of R. We also organized a workshop on January 22, 2018. We are currently collaborating with researchers from Russia and Armenia for the realization of an atlas of Armenian dialects, with new data never investigated in previous studies. The ongoing collaboration with Ioanna Sitaridou (University of Cambridge) aims at the elaboration of georeferenced map of some other languages of the area (e.g. Greek, Aramaic). Data for these languages are contributed by our partners. The purpose is to achieve interoperability of various maps in order to investigate language contact in the area in a more systematic way. The chapter on language contact in Anatolia will be co-authored by A. Donabedian and I. Sitaridou.
1.2. Publications

Journal papers

Book chapters and proceedings
Gandon, O. [in press] “Linguistic divergence with respect to relativization strategies in East Anatolia”. Turkish, Turkic and the languages of Turkey, UMass University, Amherst, November 21st-22nd.
2. Presentation of the new project

2.1. Abstract
As in the first phase, the main purpose of the operation is the modeling of the notion “linguistic area” and the interaction between dialectology, contact linguistics and linguistic typology. The linguistic configuration in West Asia is extremely promising for investigating these topics. The operation LC3 brings together several specialists of the languages of the area (such as Armenian, Iranian languages, Caucasian languages) and aims to actively contribute to a better understanding of contact phenomena.

2.2. Objectives
In the new phase of the Labex EFL (2020-2024), we will investigate 2 new aspects of the languages of the West-Asian area:

1) OV word order/variation/convergence: A set of typological properties have been shown to converge with SOV basic order (e.g., agglutination, preverbal wh- interrogatives, preverbal focus, pro-drop, lack of morphological agreement in NPs, non-finite relativization, and, maybe less reported, bare objects). In the Anatolia-Caucasus-Iran continuum, several languages share these features, and in addition show a similar pattern of information packaging, characterized by a default focus preverbal position bearing sentence nuclear stress, and a focus marked by prominent stress on any constituent of the clause regardless of its linear position. Furthermore, the linear order in those languages is flexible since it depends on discourse configuration and not on syntactic constraints, as accounted for as discourse configurationality.

   Our research will focus on word order variations in the OV languages of the era in order to identify the parameters intervening in word order preferences, in particular information structure, and also the influence of linguistic contact in these variations.

   Although many languages of the area are described as SOV (Armenian, Persian, Kurdish, Turkish, Georgian), they display significant differences with respect to the degree of flexibility in word order. In Eastern Armenian, for instance, VO order seems to be quite frequent, without “backgrounding” postverbal constituents, while in Persian and in Turkish, postposed constituents are generally claimed to be backgrounded. Kurdish, in the other hand, seems to display strict OV order, and no postposing is possible.

2) Differential object marking (DOM) is another feature shared by many languages of the area. While it has been extensively investigated in Persian, in other languages, such as Armenian, it has received less attention. Apart from studying the parameters intervening in DOM, we will especially focus on the relation between DOM and word order.

2.3. Outcome
- An interoperable map of the languages and dialects of Anatolia (follow up of the work initiated in the first phase of the project)
- Theoretical modeling of word order variations, information structure and differential object marking

2.4. Evolution with respect to the first phase
The two main topics of investigation are new and have not been investigated during the previous phase.
### VC2: Central Gallo-Romance: linguistics and ecology of a transitional zone

<table>
<thead>
<tr>
<th></th>
<th>2015 – 2019</th>
<th>2020 - 2024</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strand</strong></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>Work package</strong></td>
<td>LC4</td>
<td>VC2</td>
</tr>
<tr>
<td><strong>Project title</strong></td>
<td>The linguistic Crescent: a contact area between Occitan and Oilic varieties</td>
<td>Central Gallo-Romance: linguistics and ecology of a transitional zone</td>
</tr>
<tr>
<td><strong>PI</strong></td>
<td>Nicolas Quint (LLACAN)</td>
<td>Nicolas Quint (LLACAN)</td>
</tr>
<tr>
<td><strong>Participants</strong></td>
<td>LABEX members:</td>
<td>LABEX members:</td>
</tr>
<tr>
<td></td>
<td>Nicolas Quint (LLACAN), Heather Burnett (LLF), Judit Gervain (LPP), Maximilien Guérin (LLACAN), Barbara Hemforth (LLF) Jean-Léo Léonard (Université Paris 4), Sylvain Loiseau (LACITO, UMR7107), Stefano Manfredi (SeDyL), Marijana Petrovic (LACITO)</td>
<td>LLACAN: Nicolas Quint (DR), Maximilien Guérin (postdoc) LLF: Heather Burnett (CR), Barbara Hemforth (DR) LPP: Judit Gervain (DR) Jean-Léo Léonard (PU Paris 4) LACITO: Sylvain Loiseau (MdC, Paris 13) SeDyL: Stefano Manfredi (CR) Marijana Petrovic (LACITO)</td>
</tr>
<tr>
<td></td>
<td>External participants:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gilles Adda (LIMSI), Philippe Boula de Mareuil (LIMSI), Guylaine Brun-Trigaud (BCL), Jean-Christophe Dourdet (FORELL), Marc-Olivier Hinzelin (Université d’Hambourg), Philippe Maurer (Université de Zurich), Stephan Mietzke (Université de Rennes 1), Jérôme Picard (UPS2259) Rafèu Sichel Bazin – postdoc (CLLE, UMR5263)</td>
<td>External participants: Gilles Adda (IR, LIMSI), Philippe Boula de Mareuil (DR, LIMSI), Guylaine Brun-Trigaud (IA, BCL), Jean-Christophe Dourdet (MdC, FORELL), Louise Esher (CR, CLLE), Marc-Olivier Hinzelin (Université d’Hambourg), Philippe Maurer (Université de Zurich), Stephan Mietzke (MdC, Université de Rennes 1), Jérôme Picard (IR, CLT), Rafèu Sichel Bazin (postdoc, CLLE)</td>
</tr>
</tbody>
</table>
1. Summary of the first phase

The first funding the LC4 project received from EFL came in 2016, but our research group managed to secure an external funding from DGLFLF (see below), which allowed us to begin fieldwork in 2015.

Since the beginning of the project, LC4 members have been active on the field:

- The descriptivists of the team have been collecting data from the Crescent varieties spoken in the following communities: Anzême (23), Azérables (23), Bonnat (23), Bussières-Poitevine (87), La Celle-Dunoise (23), Dompierre-les-Églises (87), Fresselines (23), Fursac (23), Lourdoueix-Saint-Michel (36), Saint-Agnant Versillat (La Prugne, 23). For most of these locations, monographs (grammatical analysis + lexicons) are now in progress.

- At least two typological surveys have been initiated: (i) one on phonological data, which were used for specific publications and contributed also to enrich the online data base on regional languages (https://atlas.limsi.fr/; the Crescent area appears in grey on the map); the phonological data were collected in places not previously contemplated by descriptivists (eg. Eguzon-Chantôme (36), Naves (63), Pleuville (86)); (ii) a few preliminary psycholinguistic tests, which will be elaborated upon in the proposed new phase of the project.

Each fieldwork resulted in the recording and archiving of several hours of audio/video material which are systematically stored on a hard disk and will be made available on a specific section of HumaNum. This linguistic material will be used as a common resource both for researchers of the project (in particular for comparative analyses) and for the general public.

Regular meetings were held both in Paris and in the Crescent region. The first international scientific conference dealing exclusively with Crescent varieties took place on March 25-27 2017 in Le Dorat.

The high commitment and scientific excellence of the team members allowed us to secure two new fundings, from DGLFLF in 2016 and ANR in 2017 (see details below).

At present, various monographs are being prepared and at least two should be released before the end of 2019. Several papers will also be published in 2018 and in the following years.

Overall, the asset of the team resides in its multidisciplinarity, as it regroups both typologists/descriptivists and specialists of various branches of Language Sciences (notably Morphology, Natural Language processing, Phonetics and Phonology, Psycholinguistics, Semantics).

1.1. Publications

Chapters in edited volumes


Conferences and seminars


Organisation of conferences and summer schools

2. Presentation of the follow-up project

2.1. Abstract
Our research group intends to develop a multidisciplinary approach to the study of the Gallo-Romance varieties of the linguistic Crescent. These varieties represent fascinating case studies for contact linguistics, as they display simultaneously Oïlic and Occitan features. Special emphasis will be put on the following three topics: (i) the linguistic Crescent and its environment (study of the Oïlic or Occitan varieties spoken around the Crescent area and of the regional French varieties spoken alongside the Crescent varieties), (ii) checking the non-linguistic (anthropological, historical etc.) correlates of the isoglosses observed in the field, (iii) in cooperation with members of Strand 4, development of psycholinguistic studies devoted to the Crescent. All in all, this new project aims at proposing a global, multidisciplinary and ecological approach to the linguistic Crescent.

2.2. Planned activities and outcomes
Central Gallo-Romance: linguistics and ecology of a transitional zone will continue and enrich the work initiated within the scope of LC4. The following deliverables are planned:
- 10 monographs (grammatical analyses + basic lexicon) devoted to diverse Crescent varieties.
- 20 papers dealing with various features displayed by these varieties in several domains (semantics, morphology, phonetics/phonology, natural language processing, psycholinguistics etc.).
- Organisation of several international conferences within and outside the Crescent area (the next one is scheduled for February 2019 in Montluçon).
- Participation at several international conferences (general and Romance linguistics).
- Development and on-line publication of questionnaires specifically devoted to the Crescent varieties.
- Archiving on Huma-Num of over 50h of audio and video recordings of data from least 30 different Crescent varieties.
- Development of a searchable online data base combining linguistic data and cartographic functions.
- Collective volumes dealing with the Crescent varieties.

2.3. Evolution with respect to the first phase
LC4 ‘Les parlers du Croissant’ was begun and developed during the second half of the first phase of Labex EFL (see above), and the project proposed here will be a direct continuation of this enterprise. The varieties of the linguistic Crescent will be the main focus of our multidisciplinary team. At the same time, we intend to enlarge our perspective by taking into account new, more interdisciplinary approaches regarding the Crescent.

The linguistic Crescent and its environment
The Gallo-Romance linguistic varieties of the Crescent, which display typical Oïlic and Occitan features simultaneously and are traditionally used in the central part of Gallo-Romania, represent essentially a transitional area between two diasystems. Today, only a minority of the population of the linguistic Crescent still speak a variety of the Crescent as their mother tongue.
In order to understand the actual practices of these speakers, we must take into account the overall ecology of the transitional zone into which they are inserted. This means that:
- Occitan (Auvergnat, Limousin) and Oïlic (Poitevin-Saintongeais, Berrichon) varieties spoken in the vicinity of the Crescent area will be taken systematically into consideration in order to understand what exactly happens in this traditional zone.
- Local varieties of standard French are spoken on a daily basis by all speakers of Crescent varieties as well as by most of the population who do not speak any Crescent variety. It is therefore crucial to study the main characteristics of these local varieties of standard French in more detail in order to ascertain if any of their features can be traced back to the Crescent varieties; for instance, compare local Crescent standard French ‘ça pleut’ (ça = neutral expletive pronoun distinct from 3rd masculine singular) and Saint-Priest-la-Feuille Crescent variety ‘qua pleu’ (qua = neutral expletive pronoun distinct from 3rd masculine singular) vs. Paris standard French ‘il pleut’ (expletive pronoun identical to 3rd masculine singular).

The linguistic Crescent: linguistic and non-linguistic aspects of the transitional zone
Apart from linguistic features, other traits seem to characterize the Crescent area as a transitional zone. For instance, the limit between Occitan and Oïlic varieties closely follows the limit between the places where people used to plough with oxen (Southwards) and the places where people used to plough with horses (Northwards). Within the framescope of an ecological approach of the linguistic Crescent, the non-linguistic correlates of the Crescent will be systematically identified and analysed (possibly in cooperation with non-linguists, e.g. historians or anthropologists, thereby enhancing the interdisciplinarity of our project).

Psycholinguistic characterization of bilingualism in the Crescent
Today, psycholinguistics clearly represents a new frontier for language sciences and linguistic research. The Crescent area is of particular interest for psycholinguistic research. As a matter of fact, virtually no study was ever produced about the effects that bilingualism has on the brain of people practicing simultaneously French and a regional language (in particular a Romance variety). Good experimental data about regional languages are still missing in the domain of psycholinguistics (most cases of bilingualism that are studied within this framescope regard people who practice simultaneously two official languages, e.g. Spanish and French, English and German etc.). All of today’s speakers of a Crescent variety also speak French on a daily basis and generally with (near)native proficiency. The psycholinguistic study of their productions and abilities is liable to provide us with a theoretical model accounting for the bilingual profile national language + regional language. This is the reason why typologists and psycholinguists involved in LC4 have agreed to further develop their synergy and propose a PhD project ‘bidialectalism’ (supervisors: Judit Gervain & Nicolas Quint) about the psycholinguistic aspects of bilingualism between a local language and French in the Crescent area.
1. Summary of the first phase
The research project GD2 “The syntax of complex sentences in creole languages” only began officially in 2016 (first year when funding were made available), although its members already started to collaborate informally during the second half of 2015. The main aim of the project is to develop a typology of subordinate/coordinate clauses in creole languages on the basis of a searchable database.
Since the beginning of the project,
- we have organised 2 internal meetings per year that allowed us to discuss the main problems of the analysis and comparison of complex sentences in creole languages. During the meetings we had the possibility to collaborate with invited experts in the field (Sonia Cristofaro, Susanne Michaelis – LABEX international chair 2017). Until now,
we have focussed our attention on subordination; we plan to deal with coordination and serial verbs during the last two years (2018 and 2019) of our programme.

- Following the discussions held during the meetings, we have developed a typological database devoted to the syntax of complex sentences in Creole languages and related languages (lexifiers and substrates).
- In 2018, we set up a workshop on Creole syntax for the Syntax of the World’s Languages conference held in Paris, September, 3rd–5th (https://swl8.sciencesconf.org/)
- Furthermore, we (i) presented our topic of studies to many scientific conferences, (ii) did linguistic fieldwork in order to collect data for our database and check our hypotheses and (iii) published a considerable bulk of material dealing with Creole languages and in particular their syntactic characteristics; (iv) note also that several PhD students are integrated to GD2, three of which at least have syntax as the main or one of the dominant focus of their research.

1.1. Publications

International journals
Fon Sing, Guillaume, forth. Creoles are not typologically distinct from non-Creoles. In Language Ecology. John Benjamins.


Quint, Nicolas, 2015, Capeverdean words derived from Portuguese non-infinitive verbal forms: a descriptive and comparative study, Papia (Revista Brasileira de Estudos Crioulos e Similares), N°25/2, 189-215.


Books


Edited volumes

Chapters in edited volumes


Quint, Nicolas, 2015, Let's Speak Capeverdean. Londres/Colombo: Battlebridge


Databases


Conferences and seminars

Fon Sing, Guillaume. 2015. About some innovative features in Mauritian Creole. Workshop GDRI SEEPiCLa. CLUL/FLUL, Universidade de Lisboa, Portugal.

Fon Sing, Guillaume. 2015. Morphological and syntactical innovations in Mauritian Creole: extraction and analysis from regional French data. Tenth Creolistics Workshop: “Innovations” - with special attention to parallels between creole and sign language creation. Aarhus University, Denmark.


Manfredi, Stefano 2015. The history of kedé in Sudanese Arabic and Arabic-based pidgins and creoles. From speaker oriented modal marker to general subordinator. 11ème conférence
Migge, Bettina. 2015. The role of multilingual practices in negotiating social change. Minor Translingualism. London: Institute of Modern Languages Research, School of Advanced Study, University of London. 28 October.
Nunez, Joseph Jean François, 2016, Casamancese creole in contact, Invited lecture 25 April, Bremen University.
Nunez, Joseph Jean François, 2016, The sociolinguistic situation in Senegal, Invited lecture 27 April, Bremen University.

Quint, Nicolas & Eliane Araújo, Frase complexa em cabo-verdiano: estudo baseado em corpus. ACBLPE annual meeting, Universidade de Cabo Verde, Praia, 2016/06/25.


Quint, Nicolas & Joseph Jean-François Nunez, 2015, Contact de langue chez les créolophones casamançais: l’exemple des déterminants. Society for Pidgin and Creole Linguistics (SPCL) and Associação de Crioulos de Base Lexical Portuguesa e Espanhola (ACBLPE) joint conference, organized by Institut für Romanistik de la Karl-Franzens-Universität Graz (Graz University), Graz (Austria).


Quint, Nicolas, 2015 (keynote speaker), L'atlas linguistique du capverdien. Séminaire de l'équipe "systèmes linguistiques et dialectologie" du Gipsa-Lab (Grenoble University /CNRS), Grenoble.


1.2. Organisation of conferences and summer schools


One of the main problems in typologizing syntactic structures on the basis of (semi-)spontaneous data is represented by the high degree of individual variation affecting these structures. This is especially true for creole languages, whose syntactic structures are affected by different linguistic processes such as substrate/adstrate interference, creole-lexifier attrition, as well as innovative ongoing changes. This new project aims at integrating the analysis of syntactic variation with the typologisation of different kinds of syntactic structures (subordinated and coordinated clauses) in creole languages as well as improving the categorization of the semantic values and origins of creole coordinating and subordinating conjunctions.

The project “Syntactic variation in creole languages” is a prolongation of the project GD2 “The syntax of complex sentences in creole languages”, which aimed at producing a database of complex sentences in a number of creole languages with different lexifiers and substrate languages. While compiling our database, it immediately became clear that a standard synchronic comparison of subordinate clauses was particularly difficult because of the concurrence of many constructions for encoding the same meaning. The main aim of this operation is thus to analyze different factors intervening in the syntactic variation of creole languages and to integrate them within a typological approach to syntactic structures. Firstly, we will focus our attention on ongoing internal changes linked to different degrees of grammaticalization and reanalysis. When necessary, we will also take into account the outputs of different kinds of creole-lexifier bilingualism. Lastly, we will analyze instances of matter and pattern transfer from substrate/adstrate languages into creole languages. The analysis of the aforementioned factors of variation will be implemented while expanding our database on the syntax of complex sentences in creole languages (http://llacan.vjf.cnrs.fr/labexGD2/form/index.php)

**2.3. Evolution with respect to the first phase**
- Development and integration of a multifactorial grid for the analysis of syntactic variation within the GD2 database.
- Development of semantic maps for visualizing the grammaticalization of subordination and coordination markers.
- Development of a concordancer for doing complex queries in the database.
1. Presentation

1.1. Abstract
This project aims at developing a fine-grained analysis of the contact dynamics involving closely related languages. Based on heterogeneous (i.e. multistylistic and multidialectal) corpora, its ambition is to typologize different kinds of variation at play in contact settings so as to revise the notion of “language boundaries” and to refine the analysis of the social implications of language contact.

1.2. Objectives
In the attempt to embrace different synchronic aspects of bi/multilingualism, we will work on diverse sociolinguistic settings involving contact between closely related languages. On the one hand, we will investigate contact between creole languages and their lexifiers while adopting a dynamic perspective of analysis which goes beyond the idea of “decreolization”. The theoretical reflection about this particular contact setting will be founded on the annotation and study of heterogeneous corpora including, among others, French, English, Portuguese, Arabic-based creoles, and their respective lexifiers. On the other hand, we intend to analyze the practices characterizing contact between varieties of the same language. For this part of the project, we will work on multidialectal corpora of Spanish, English, Arabic, French, Occitan, and Scottish Gaelic. This will eventually allow us to identify a number of indexical differences between dialects in contact and their role in the selection of linguistic features. By typologizing the previous contact settings, we will also contribute to the debate about the social structuring of linguistic diversity, which goes under the name of “sociolinguistic typology”, as well as to our understanding of the synchronic emergence of new linguistic varieties. On a methodological ground, we will adopt both quantitative and qualitative methods of investigation for working on our corpora of annotated first-hand transcriptions.
1. Summary of the first phase

1.1. Activities since 2015
Language contact phenomena have received growing attention in the last 30 years. Different brands of Contact Linguistics (Thomason, 2001, Winford, 2003) proposed typologies of language contact settings based on parameters such as the type of setting (language maintenance or
loss), social factors (length and intensity of contact, degree of bilingualism), language contact mechanisms and results (borrowing, shift, calque, grammaticalization, reorganization etc.; see Croft (2000), Field (2002), Heine and Kuteva (2003 & 2005), and Matras and Sakel (2007)). Most of the studies with a typological perspective focus on the description of the outcomes of contact and its impact on the languages involved – they were realized with a diachronic perspective, and to date little attention has been paid to synchronic phenomena. Whereas most of the studies with a synchronic perspective focus on code-switching and code-alternation from a structural and cognitive perspective in that they try to determine the linguistic patterns and constraints that govern code-switching (cf. Poplack, 1980, Myers-Scotton, 1993, Auer, 1998, Muysken, 2000), research on multilingualism and synchronic contact settings is more heavily concerned with the social constraints and the meaning-making potential of code alternation; this research is mostly conducted by researchers interested in the sociology of language, interactional sociolinguistics or second language acquisition. A few projects proposed in the last years the interaction and cross-fertilisation between these two lines of research. The research project LC1 within the first phase of LABEX EFL is one of them. It proposed to identify multiple causation for both internally motivated changes and contact-induced processes, through a fine-grained approach integrating both social and linguistic factors (see e.g. Chamoreau & Léglise 2012, Migge & Léglise 2013). This involved devising a multifaceted methodology and a multi-model approach to explain contact-induced language changes. It involved furthermore the development of common computer-based tools as an XML editor and a multilingual concordancer that were tested by the various participants in the LC1 research program (see Vaillant & Léglise 2014 and publication lists) available on a platform: http://lc1-labexefl.vjf.cnrs.fr.

1.2. Publications

International journals


2. Presentation of the follow-up project

2.1. Abstract
This research program investigates codeswitching and contact phenomena through multifactorial analysis of plurilingual corpora. It aims at 1) critically examining the typology of terms proposed in the literature and 2) devising a typology of codeswitching phenomena from corpora from a great range of situations and social contexts.

2.2. Objectives
Typologies of code-switching or code mixing were proposed 20 years ago (see e.g. Auer 1998, Muysken 2000) and are still being discussed. Most studies in contact linguistics and code-switching assumed languages to be discrete and bounded systems in contact; terms like matrix language or code-alternation are good examples of this hypothesis (Léglise 2017). Within the last decade, there has been a shift in focus from the investigation of how relatively bounded communities come into contact and the kind of linguistic outcomes that gives rise to, to that of how social actors produce and negotiate social meanings, social positions and relations through their language practices. Hall & Nilep (2015), for instance, shows very clearly how contemporary
research, much centered on linguistic repertoires, wonderfully illustrates disruption of presumed connections between language, community, identity and spaces. Within the scope of multilingualism, lots of terms were coined in the last 10 years to give voice to the variety of practices that are observable such as crossing (Rampton 2005), trans-idiomatic practices (Jacquemet 2005), translanguaging (Garcia 2009, Garcia & Li 2013, Creese & Blackledge 2010) or (poly)languaging (Jørgensen et al 2011), heterogeneous language practices (Léglise & Alby 2013). One of the aims of this research program is to critically examine the typology of terms proposed in the literature in the light of the codeswitching phenomena observable in our plurilingual corpora that include data from a range of contexts, such as formal, educational and informal settings and a diverse range of language pairs. A second aim is to devise a typology of codeswitching phenomena from corpora from a great range of situations and social contexts. This typology will crucially aim to also integrate structural-cognitive and socially oriented approaches to code alternation phenomena. Finally, it involves exploring and developing computer-based tools for analysing plurilingual and linguistically heterogeneous language corpora. It will facilitate identifying social and structural factors that condition the appearance of zones of bivalency between languages, types of switches, word order phenomena or linguistic forms belonging to neither of the languages in contact.

2.3. Evolution with respect to the first phase
During the last years, LC1 proposed to identify multiple causation for both internally motivated changes and contact-induced processes, through a fine-grained approach integrating both social factors and linguistic factors. The computer-based tools developed in that project as well as the multifaceted methodology and multi-model approach that have been set up will be further applied to synchronic issues in order to better handle, describe and explain codeswitching phenomena and ongoing contact-induced language changes.