OAR Journal

Firenze University Press – USiena PRESS
https://riviste.fupress.net/index.php/oarj

Oral Data Production

L2FOC: a crosslinguistic speech corpus for French and Italian Second Language Acquisition

Bianca Maria De Paolis*, Simone Stroppiana

Università di Torino, Italy
biancamaria.depaolis@unito.it; simone.stroppiana@edu.unito.it
*Corresponding author

OPEN ACCESS

Citation: De Paolis, B. M., & Stroppiana, S. (2026). L2FOC: a crosslinguistic speech corpus for French and Italian Second Language Acquisition, *Oral Archives Journal*, 2: 5-25. doi: 10.36253/oar-3497

Received: May 13, 2025
Accepted: January 26, 2026
Published: May 12, 2026

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

ORCiD
BMDP: 0000-0001-7725-9617

Ai soli fini concorsuali, l'attribuzione dei paragrafi è la seguente: Bianca Maria de Paolis per i §§ 1, 3, 4 e 5; Bianca Maria De Paolis e Simone Stroppiana per il § 2.

Abstract. This paper presents L2FOC, a crosslinguistic corpus of spoken Italian and French developed to investigate the syntactic and prosodic realisation of focus in both native and non-native speech. The corpus includes approximately 10 hours of recordings from 65 speakers across four groups (L1/L2 Italian and French), collected through three tasks designed to elicit speech with a ranging degree of spontaneity. Recordings were made with laboratory equipment and are accompanied by orthographic transcriptions and multi-tier phonetic and phonological alignment. The corpus enables fine-grained analysis of how information structure is encoded across languages and proficiency levels, and supports applications in both theoretical linguistics and speech technology. To illustrate its analytical potential, two studies are briefly discussed: one examining the syntax–prosody interface in L2 focus strategies, and one assessing automatic speech recognition performance on learner speech.

Keywords: second language acquisition, information structure, romance languages, prosody-syntax interface, task-elicited speech.

1. INTRODUCTION

1.1. Corpora and linguistic resources in Second Language Acquisition

In the study of second language acquisition (SLA), corpora are indispensable tools for observing how linguistic structures emerge, stabilize, and vary across contexts. Corpora allow researchers to access naturalistic language data at scale, enabling analyses that are empirically grounded and reproducible. However, not all corpora are created for the same purpose: linguistic resources vary in design and annotation depth, ranging from large-scale, semi-spon-

taneous speech corpora intended to approximate authentic usage, to highly controlled experimental datasets aimed at isolating specific variables such as syntactic complexity or prosodic contour. This distinction introduces a longstanding methodological tension between the ecological validity of naturalistic corpora and the experimental control of elicited datasets (for an overview, see Gilquin, De Cock and Granger 2010 or Mackey and Gass 2016). While natural corpora offer invaluable insight into naturalistic usage, they rarely provide the structural balance or annotation granularity required to analyze tightly constrained phenomena – particularly when investigating subtle interfaces such as prosody and syntax in focus realisation (on the notion of focus see Lambrecht 1994; Krifka 2008, among others). Furthermore, while it is true that for a truly spontaneous communication everything, including the setting, should be naturalistic, it is also true that without a microphone and a minimally soundproof place, it is very hard to obtain the audio quality for a reliable phonetic analysis. Conversely, experimental data often lacks the spontaneity and diversity of authentic interaction, potentially limiting generalizability.

In this landscape, the first, major problem is that the SLA corpus research remains dominated by works on L2 English, with large-scale resources such as the Cambridge Learner Corpus (OpenCLC, 2017), EFCAMDAT (Geertzen, Alexopoulou and Korhonen 2014), and ICLE (Granger et al. 2020) providing extensive data across multiple L1 backgrounds. Efforts to expand the SLA research corpora to languages different from English are being made, e.g., for Chinese, the International Corpus of Learner Chinese, which aims to balance and widen the range of learners of the existing corpora (Zhang and Tao 2018). Equivalent corpora for Romance languages, though, remain rare (particularly for Italian), often lacking prosodic annotation, controlled elicitation protocols, or public accessibility. The lack of suitable resources becomes particularly problematic when investigating complex information-structural categories such as focus in L2 speech, where syntax, prosody, and crosslinguistic influence all come into play. While general-purpose learner corpora like FLLOC (Myles 2006), the Newcastle Corpus (Allen et al. 2007), or PAROLE (Hilton 2009) offer valuable data, they were not designed to elicit specific information-structural configurations, and do not provide a balanced representation of native and non-native speech across both Italian and French. Many recordings included in such corpora also fall short in terms of audio quality, making them unsuitable for prosodic analysis–an essential requirement not only for tracking strategies involved in focus marking, but also for investigating broader domains, such as pragmatic competence and/or discourse organization in second language learning.

On the other end of the spectrum, some corpora specifically designed to examine discourse-related phenomena in L1/L2 Italian and French (e.g., Turco, Dimroth and Braun 2013; Anastasio 2021) offer a suitable combination of speaker groups and high audio quality. However, they consist of very tightly controlled datasets, which limits their applicability to other research domains. As a result, they lack a sufficient number of relevant occurrences to support robust, generalizable conclusions for research aims beyond their original scope.

All these imbalances pose a major obstacle to the specific field of comparative SLA research: in the absence of corpora that combine cross-linguistic coverage, phonetic detail, and task-based control, it is difficult to test hypotheses about transfer effects and discourse in L2 acquisition. The L2FOC corpus was developed precisely to fill this gap, as it offers a carefully constructed, richly annotated dataset designed to support empirical investigation at the

syntax-prosody interface in both native and learner speech, and to move beyond the Anglocentric focus of most current SLA resources.

1.2. Aim and purpose of L2FOC corpus

In response to the limitations outlined above, the L2FOC corpus was conceived as a resource specifically designed to support contrastive analyses of focus realisation in second language speech within Romance languages, combining high-quality recordings that enable detailed acoustic investigation of prosody with a theoretically grounded, nuanced treatment of information structure. It was developed in the context of the PhD project of the first author, jointly supervised by the University of Turin and the University of Paris 8 (De Paolis 2024). The research aimed to explore how focus is encoded both syntactically and prosodically in Italian and French, comparing native speakers with L2 learners of each language. Particular attention was given to cross-linguistic differences in focus strategies, the relative difficulty and acquisition trajectory of different structures, and the non-linear patterns often observed in the development of syntactic competence in L2 production. To operationalize these principles, the L2FOC corpus was structured to systematically represent different speaker profiles and elicitation conditions.

From the outset, the corpus was conceived with the intention of being made publicly available and open access, in line with current best practices promoting data re-use, as well as the reproducibility and comparability of results. Another key idea that guided its development was the potential for expandability – both in terms of speaker populations and elicitation conditions – so that the resource could be enriched and adapted for future research purposes.

In the next section (2), we describe in detail how the corpus was built and how it is structured. In the final section (3), we outline two studies that have been conducted using the L2FOC corpus, illustrating its versatility and potential for addressing a range of research questions.

2. THE CORPUS

The corpus is organized into four sub-sections, each corresponding to a specific speaker group: ITL1, which includes recordings of native Italian speakers; FRL1, containing recordings of native French speakers; ITL2, composed of native French speakers with Italian as a second language; and FRL2, which includes native Italian speakers with French as a second language[1].

Each subsection is further divided into three parts, reflecting the three elicitation tasks completed by every participant. This task-based organization is aligned with the overall design of the corpus, which was conceived to capture a broad spectrum of speech types – from semi-spontaneous utterances to highly controlled productions – enabling the study of how focus is realised under different cognitive and interactional conditions.

[1] The specific areas of origin of the Italian and French participants are detailed in the following section on participant characteristics.

Beyond the speech data itself, the corpus was developed to meet several key criteria for linguistic analysis, including balance across languages, participant types, and age groups, as well as the collection of detailed metadata concerning learners' linguistic background, acquisition context, and proficiency levels. The following sections outline these aspects in greater detail.

In addition to the speech data, the corpus is enriched with orthographic transcriptions and comprehensive metadata, both of which play a crucial role in facilitating linguistic analyses. The transcriptions provide a reliable basis for syntactic and prosodic annotation, while the metadata offers essential contextual information about each speaker. These include details such as age, gender, language background, length and context of L2 exposure, and self-assessed as well as independently evaluated proficiency levels. The following paragraphs provide a detailed account of the transcription conventions and metadata structure, outlining how they contribute to the usability and extensibility of the L2FOC corpus for a range of research purposes.

2.1. Participants

The selection and organization of participants in the L2FOC corpus were informed by established methodological principles in second language acquisition (SLA) research. Within this field, it is widely acknowledged that cross-sectional designs – comparing multiple learner groups at a single point in time – can yield insights comparable to those obtained through longitudinal studies, especially when investigating interlanguage development (see, in particular, Jarvis and Pavlenko 2010; Gass and Selinker 2001).

Participants were recruited between summer 2021 and autumn 2022 in two locations, Turin (Italy) and Paris (France). Italian native (ITL1) and L2 speakers (ITL2) were recorded in Turin, while French native (FRL1) and L2 speakers (FRL2) were recorded in Paris. In order to minimize the impact of diatopic variation[2], only speakers from the Turin area have been selected for the Italian groups; for the French groups, native and non-native participants were drawn from the Paris region. All participants included are adults (aged 18 or older), to avoid including learners in the so-called critical period of language acquisition (Lenneberg 1967). In total, 65 participants were recorded, distributed across the four subgroups: FRL1 (n = 18), ITL1 (n = 14), ITL2 (n = 17), and FRL2 (n = 15).

All participants were volunteers who received no financial compensation for their involvement, and they provided informed consent prior to participation. To ensure anonymity and confidentiality, each participant was assigned an anonymous identification code, which was used in place of names in all data files. Personal information was stored separately from research data, and documents related to language assessment were kept distinct from consent forms to avoid any possibility of identification. Any occurrences of personal identifiers within the recordings or transcripts were removed during data processing, and only the

[2] We are aware that referring simply to 'Italian' and 'French' is insufficient, given the strong diatopic variation in both languages – though for different reasons. In Italian, variation is primarily local and regional, whereas in French it also reflects pluricentricity and the historical spread of the language beyond the metropolitan area. For this reason, we do not claim our sample to be representative of all possible varieties of Italian and French. Instead, we deliberately restricted the points of inquiry to minimize within-sample variability, at the cost of not capturing the full breadth of these highly diverse languages.

research team had access to the anonymized dataset, in order to guarantee that participants' privacy was rigorously protected throughout the study, in accordance with Regulation (EU) 2016/679[3].

Corpus metadata includes socio-biographical information about each participant: age, gender, education, language background. All this information was used to evaluate inter-group comparability and ensure sufficient diversity in learning trajectories, and can also be exploited to conduct further investigation on the materials. The full version of the linguistic questionnaire is available in Appendix.

Non-native speakers also completed a multi-component proficiency assessment, combining three measures: (i) a self-assessment questionnaire; (ii) a written cloze test (Vedder 2008 for Italian; Tremblay and Garrison 2010 for French); and (iii) an oral evaluation of task-based speech samples, conducted by experienced L2 instructors. The oral evaluations focused on semi-spontaneous responses during the picture comparison and picture story tasks, which were particularly informative for assessing real-time language use. Evaluation was performed considering both CEFR descriptors (CEFR 2020) and the CAF framework (Complexity, Accuracy, Fluency, see Pallotti 2009; Norris and Ortega 2009). Table 1 summarizes the demographic characteristics and CEFR proficiency levels of the participants[4].

Table 1. Summary of the final sample characteristics.

Group	Sex (M/F)	Age Range	Mean Age	SD Age	CEFR Levels	N. Speakers
FRL1	4 / 14	31	27.5	9.5	-	18
ITL1	3 / 11	10	25.6	10	-	14
FRL2	7 / 8	25	32.5	7.4	A1–2: 1 B1–2: 7 C1–2: 7	15
ITL2	7 / 10	35	27.4	8.6	A1–2: 4 B1–2: 7 C1–2: 6	17

Recordings for the ITL1 and ITL2 sub-sections were conducted in a soundproof booth at the Laboratorio di Fonetica Sperimentale "Arturo Genre" (University of Turin), while recordings for the FRL1 and FRL2 sub-sections took place in the soundproof booth of the SFL – Structures Formelles du Langage laboratory (Université Paris 8). In cases where access to university facilities was restricted due to COVID-19 health measures, participants were recorded in alternative quiet locations using portable equipment, while adhering to safety protocols.

[3] This study did not require ethical approval from an institutional ethics committee, since at the time (academic year 2020–2021), both Université Paris 8 and Università di Torino did not require formal ethical review for the type of non-invasive, low-risk data collection employed in this study.

[4] The corpus is not fully balanced for gender, particularly in the FRL1 and ITL1 groups. Gender was not controlled as a primary factor because, to our knowledge, studies of focus marking in Italian and French have not reported consistent gender effects; priority was therefore given to controlling other factors such as literacy, age, and region of origin.

The total corpus comprises approximately 10 hours of speech, corresponding to around 10 minutes per speaker across 65 participants. All audio files were recorded in mono .wav format at 44.1 kHz sampling rate.

2.2. Tasks

The corpus includes three elicitation tasks, each designed to test different aspects of focus realisation under controlled or more naturalistic conditions.

2.2.1. Read-Aloud (RA)

The first task consists of three short scripted dialogues between the interviewer and the participant, centered on everyday situations. An example is given in (1); the full script is available in the Appendix section.

(1)
A. Ciao Nina! Benvenuta.
ciao Nina ben-ven-uta
hello Nina well-come-PTCP.F.SG
'Hi Nina! Welcome.'

B. Ciao Giulio! Che profumino! Hai cucinato le lasagne?
ciao Giulio che profum-ino
hello Giulio what smell-DIM
'Hi Giulio! What a nice smell!'
hai cucinato le lasagne
AUX.2SG cook.PTCP DEF.PL lasagna.PL
'Have you been cooking lasagna?'

A. No, ho cucinato la parmigiana.
no ho cucinato la parmigiana
no AUX.1SG cook.PTCP DEF.SG parmigiana
'No, I cooked the parmigiana.'

B. Davvero? L'hai fatta tu?
davvero l' hai fatta tu
really it= AUX.2SG do.PTCP you
'Really? Did you make it yourself?'

A. Sì, la parmigiana l'ho fatta io. È Giovanna che ha fatto il dolce, invece.
sì la parmigiana l' ho fatta io
yes DEF.SG parmigiana it= AUX.1SG do.PTCP I
'Yes, I made the parmigiana myself.'
è Giovanna che ha fatto il dolce invece
be.3SG Giovanna REL AUX.3SG do.PTCP DEF.SG dessert though
'It is Giovanna who made the dessert, though.'

Each participant produces 18 utterances, corresponding to 18 original turns by the interviewer[5]. The task was designed to elicit focus constituents in both marked and unmarked syntactic positions, including subjects, verbs, and direct or indirect objects. For example, in the read-aloud task participants produced corrective focus on objects in sentences such as *A: Hai cucinato le lasagne? – B: No, ho cucinato la parmigiana* ('Did you cook the lasagna? – No, I cooked the parmigiana'). Subject focus was elicited through clefts, as in: *È Giovanna che ha fatto il dolce* ('It is Giovanna who made the dessert'). Verb focus was elicited through corrective exchanges contrasting different actions, as in: *A: Hai già mandato il messaggio a Elena? – B: No, ho telefonato a Elena* ('Have you already sent the message to Elena? – No, I called Elena'). Similarly, indirect object focus was elicited in: *È a Giulia che ho telefonato* ('It was Giulia that I called'), while determiner phrases were tested in exchanges like: *Non il cedro, grazie. Sono i limoni che mi servono oggi* ('Not the citron, thanks. It's the lemons that I need today'). Whenever possible, target phrases were composed of voiced segments to facilitate reliable f0 tracking in prosodic analysis. Lexical choices were likewise constrained to maximize crosslinguistic comparability, with items matched across French and Italian for semantic and denotative meaning, lexical frequency, phonetic composition, metrical weight, etc.

A summary table (Table 2) provides the distribution of focus types elicited in this task.

Table 2. Occurrences of target constituents in the read-aloud task.

Target Const.	Focus type	Word order	N. occurrences
Subject	Correction	Marked	1
		Unmarked	1
	Identification	Unmarked	1
Direct object	Correction	Unmarked	5
	Correction	Marked	2
Indirect object	Correction	Unmarked	1
Verb	Correction	Unmarked	1

2.2.2. Question-Answer (QA)

The second task presents participants with a short illustrated story followed by a series of scripted questions designed to elicit three focus types: broad focus, narrow identificational focus, and narrow corrective focus. Stimuli and procedures are adapted from the methodology developed by Gabriel (2010), and subsequently employed in comparable crosslinguistic studies (e.g., Feldhausen and Vanrell 2014). An example excerpt is shown in Figure 1. This task was chosen for its adaptability across languages and populations, allowing consistent elicitation of comparable structures in typologically distinct L1/L2 combinations. The QA task targets focus placement on verbs, arguments, subjects, and adverbials. Participants were

[5] Only participants were equipped with a microphone during the recording sessions; the interviewer's prompts were therefore not captured in the audio files. However, all interviewer turns are preserved in the transcripts and metadata, so the dialogical structure can be fully reconstructed.

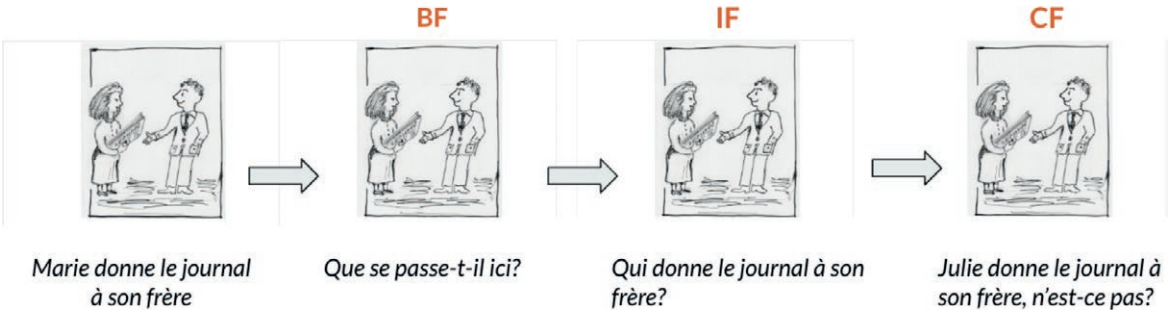

Figure 1. Some slides from the "picture story" task (French version).

simply instructed to avoid elliptical responses; no metalinguistic explanation or training was provided, to preserve some degree of spontaneity.

Each participant completed two illustrated scenes, responding to 26 target questions and three filler items. In contrast to RA, syllabic matching across languages was not feasible due to the fixed visual narrative. Nevertheless, etymologically related lexical items (e.g., 'giornale' and 'journal') were selected wherever possible to ensure semantic equivalence and partial phonetic comparability between French and Italian.

A summary table (Table 3) below details the distribution of focus subtypes and target constituents for this task. The full script is available in the Appendix section.

Table 3. Questions in the picture-story task.

Target const.	Focus type	N. occurrences
-	Broad	4
Subject	Identification	3
	Correction	3
Direct object	Identification	2
	Correction	2
Indirect object	Identification	2
	Correction	2
Adverbial	Identification	2
	Correction	2
Verb	Identification	2
	Correction	2

2.2.3. Picture Comparison (PC)

The third task was designed to elicit spontaneous[6], interaction-driven speech. The participant and the experimenter each view an image – identical in structure but differing in key

[6] We describe the speech in this section of the corpus as 'spontaneous' to distinguish it from the non-scripted, semi-spontaneous speech elicited in the second task (QA). We are, however, aware that laboratory speech cannot be regard-

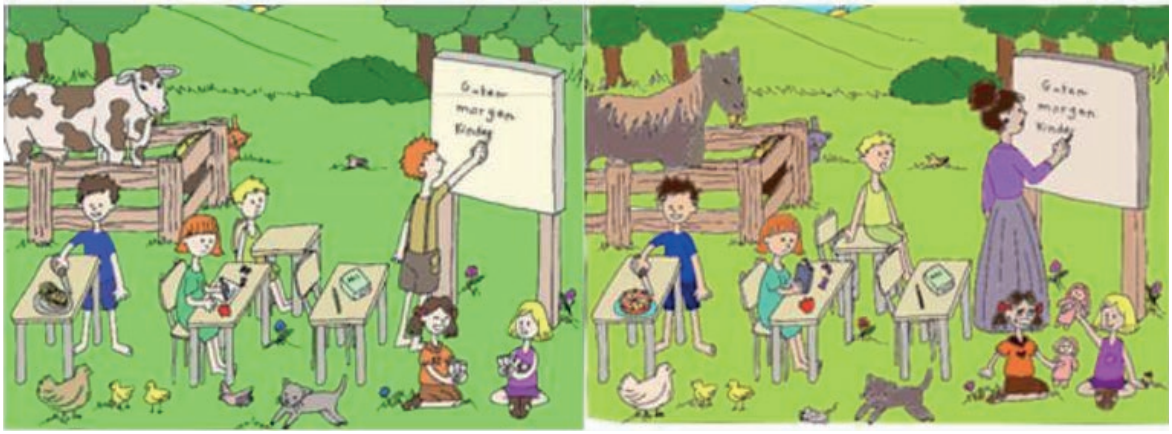

Figure 2. Picture of the experimenter (left) and of the participant (right).

visual details (see Figure 2) – without seeing each other's version. The experimenter follows a semi-controlled script[7], avoiding syntactically marked constructions to prevent priming effects. The participant's role is to identify the differences between the two scenes, thus naturally producing contrastive focus in real-time interaction. This setup allows for prosodic and syntactic observations under cognitively active, socially embedded conditions.

2.3. Annotation

Each recording is accompanied by a manual orthographic transcription. In addition to the transcription, extensive automatic segmentation was performed and subsequently manually corrected, allowing for direct use in phonetic and prosodic analysis. For the QA tasks, segmentation was carried out using EasyAlign (Goldman 2011) for FRL1_QA and FRL2_QA, and WebMAUS (Kisler, Reichel and Schiel 2017) for ITL1_QA and ITL2_QA.

The same segmentation process was applied to the RA task, using WebMAUS across all four sub-sections. Each resulting Praat TextGrid file includes four annotation tiers (see Figure 3), offering progressively broader levels of linguistic structure:

- Phoneme segmentation (with X-SAMPA transcription)
- Syllable segmentation (with X-SAMPA)
- Word-level segmentation (with X-SAMPA)
- Orthographic transcription of the utterance

---

ed as fully spontaneous: true spontaneous speech is typically obtained through non-participant observation, a method which, although fully naturalistic, raises serious privacy concerns and is therefore not feasible for corpora intended for public access and use.

[7] As in the read-aloud task, only participants were recorded with a microphone; the experimenter's speech is therefore not audible in the corpus files. However, all interviewer turns are preserved in the transcripts and metadata, ensuring that the interactional structure can be reconstructed.

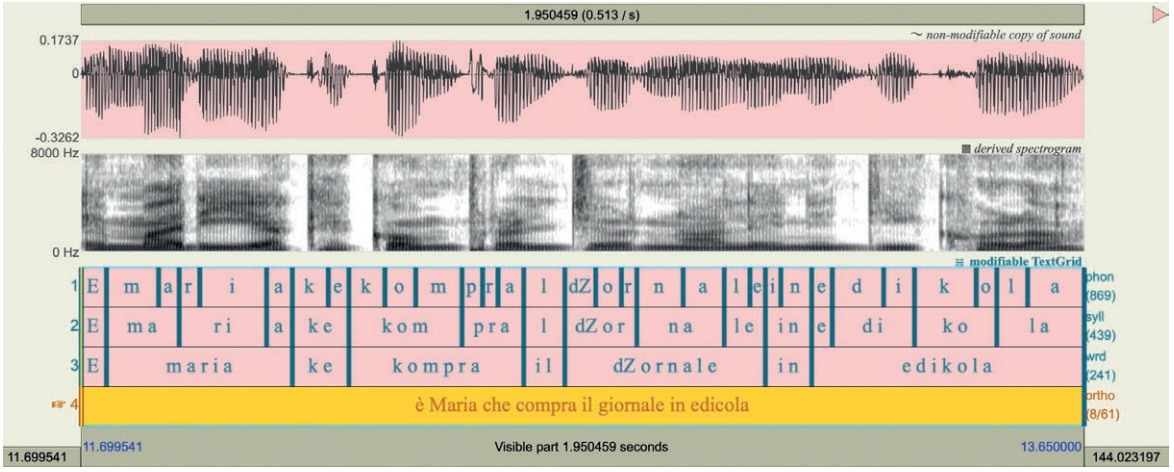

Figure 3. Example of TextGrid with 4 tiers of annotation on Praat (Boersma and Weenink 2021).

This rich multi-tier annotation greatly facilitates further research, allowing users to access well-prepared data that is ready for phonetic, phonological, and prosodic analysis without the need for extensive preprocessing. Alignments and transcriptions were manually verified and adjusted, especially since mainstream forced-alignment tools are trained on target-like native speech and often fail with non-target pronunciations typical of L2 productions[8].

3. RESEARCH ENABLED BY THE CORPUS

3.1. Focus realisation in L1 and L2

The L2FOC corpus has already supported research in both linguistic theory and applied domains. It was first used in the doctoral dissertation for which it was designed (De Paolis 2024), where it served to examine focus marking, crosslinguistic influence, and the syntax-prosody interface in L1 and L2 speech. Thanks to its controlled elicitation format and parallel cross-linguistic design, L2FOC enables direct, quantitatively grounded comparisons between native and learner productions across languages, focus types, and target constituents.

One of the most robust findings of De Paolis (2024) concerns the cross-linguistic organisation of syntactic strategies for focus marking. Across a dataset of over 1,000 annotated sentences, French and Italian display sharply different profiles. French relies on clefts almost cat-

[8] Most of the resulting errors involved complete misalignment of entire syllables or larger segments, which were systematically corrected during manual verification. During the manual revision phase, segment boundaries were refined through visual inspection of the waveform and spectrogram, combined with auditory-perceptual feedback. Temporal boundaries were adjusted according to characteristic acoustic transitions – for instance, boundaries were placed at the point of maximum formant variation in sonorous portions such as diphthongs, vowel sequences, laterals, nasals, and rhotics. For rhotic segments exhibiting clear periodic cycles, boundaries were set at the onset or release of stable periodicity preceding the following segment. For initial plosives, adjustments also considered speaker- and utterance-specific factors such as speech rate, the duration of adjacent segments, and the realization of closure phases in comparable contexts. All these adjustment criteria were applied by the annotators on the basis of established phonetic practice and were consistent with guidelines proposed in works such as Savy (2006), which were used as a reference during alignment verification.

egorically for subject focus – 94% of subject targets – and shows no functional distinction between identification and correction: clefts are used systematically in both contexts. Italian, in contrast, uses clefts much more selectively – 17% overall – and crucially specialises them for corrective focus: subject clefts rise to 55% in correction but drop sharply in identification, and virtually disappear for non-subjects (0.5%).

Learners reflect these tendencies in nuanced ways. Overall cleft frequency in both L2 groups converges on an intermediate rate, 25% in L2 Italian and 23% in L2 French, showing that formal similarity between Italian and French clefts facilitates their acquisition. As in the L1 data, clefts are used predominantly for subject focus (81% in L2 Italian; 75.3% in L2 French), whereas they remain rare for other roles. At the same time, both L2 groups show a weaker differentiation between identification and correction than Italian natives: although clefts tend to increase in corrective contexts, the contrast remains attenuated. Italian learners of French approximate the French target most closely, with systematic subject clefting regardless of focus type, whereas French learners of Italian face the additional task of acquiring a restriction on cleft use, since Italian deploys clefts primarily in correction. These results, grounded in a large and controlled dataset, illustrate how cross-linguistic similarity, structural complexity, and the nature of the information-structural contrast jointly shape learners' syntactic preferences. The results reported here are summarized graphically in Figure 4.

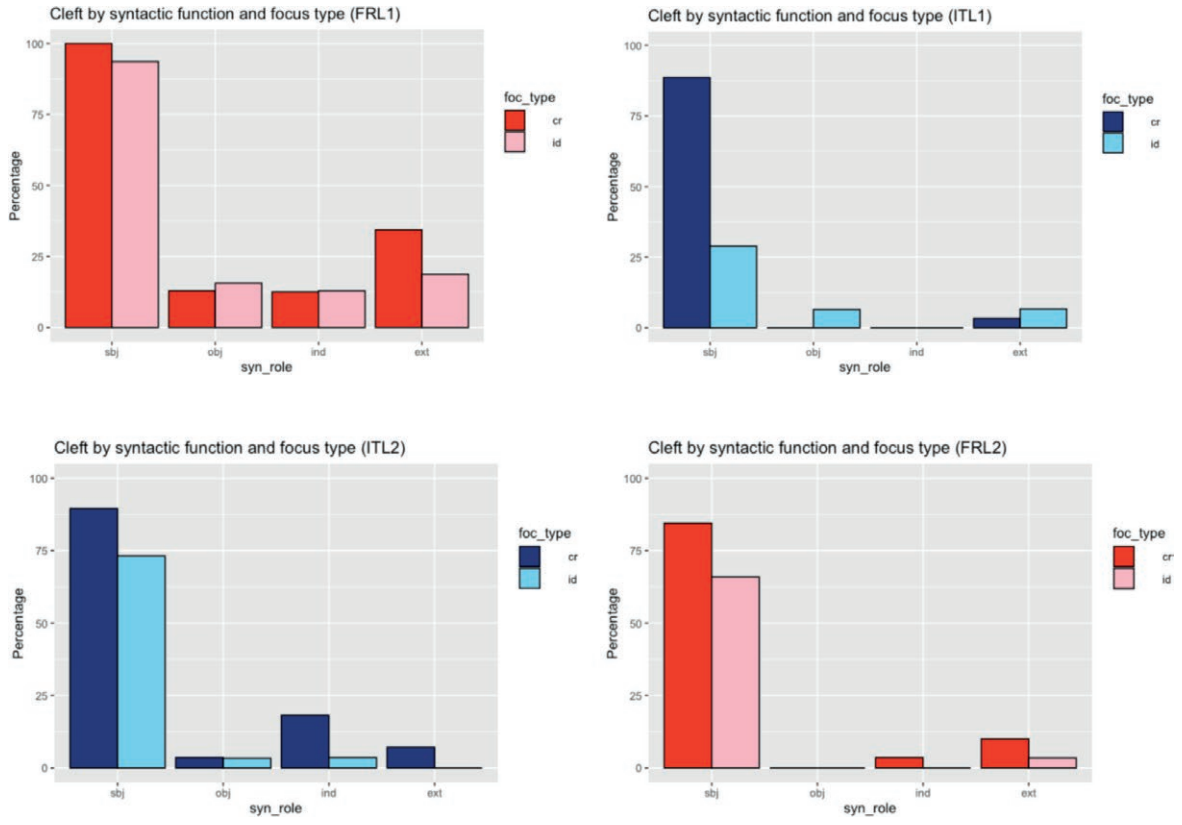

Figure 4. Proportion of cleft constructions across syntactic roles (subject, object, indirect object, external argument) and focus types (identification vs. correction) in the four groups: native French (FRL1), native Italian (ITL1), French learners of Italian (ITL2), and Italian learners of French (FRL2).

Beyond patterns of use of specific syntactic devices, L2FOC has also made it possible to track how prosodic and syntactic marking interact and vary across broad, identificational, and corrective focus in both native and non-native speech. This second analysis included only subject-focus items; each participant produced 10 elicited subject targets, so the analysis draws on around 600 tokens in total (15 speakers × 4 groups × 10 items). Results consistently reveal a gradient pattern: broad focus is predominantly unmarked (over 40–50% unmarked tokens in both L1 groups), identification introduces moderate (non-systematic) marking, and corrective focus elicits the most frequent and most complex combinations of syntactic and prosodic cues. In L1 French, narrow-focus subjects are almost systematically marked, with 80% of corrective-focus tokens showing joint marking through syntax and intonation, and the remaining 20% marked syntactically. L1 Italian shows the same gradient, but distributes marking differently: overall narrow-focus data contain 26% intonational marking, 10% syntactic marking, and 12% combined marking, but corrective subject focus triggers 89% syntactic marking and 64% additive marking, with no unmarked tokens.

Crucially, these results demonstrate that syntax and prosody are not used as alternatives, but as additive strategies, creating a distinction between the two narrow-focus subtypes, i.e. identification and correction. Increased contrastiveness leads speakers to accumulate cues rather than to replace one with the other: French relies on syntax as the baseline and adds prosodic prominence only in highly contrastive contexts, whereas Italian uses prosody as the primary resource and adds syntactic restructuring mainly under correction. No marking device is tied to a single focus subtype; what differentiates identification from correction is the degree of cumulative recruitment of syntax and prosody.

Learners follow a similar logic, though less robustly. Italian learners of French (FRL2) already adopt syntax as the default for subject focus in identification but fail to increase prosodic marking in correction, unlike L1 French. French learners of Italian (ITL2) approximate the Italian target more closely: in corrective subject focus, almost all tokens are marked, and combined syntax+intonation reaches nearly two thirds of cases. However, both learner groups underuse marking on non-subjects, leaving 50–59% of such items unmarked, a pattern reflecting both L1 influence and structural complexity. A graphical overview of these results is given in Figure 5.

Crucially, all these findings are enabled by the corpus' multi-level annotation (syntax, prosody, information structure) and its matched cross-linguistic design, which ensures comparability across groups. The robustness of the results is further supported by the large number of tokens, which provides sufficient statistical power across different analytical scopes.

3.2. Speech technology and L2

Beyond linguistic analysis, L2FOC has also proven valuable for speech technology. In a recent ASR evaluation (Aiello et al., 2025), the corpus was used to assess WhisperX performance on L2 Italian, alongside other publicly available L2 datasets. The study showed that Word Error Rates (WERs) are consistently higher for non-native speech than for native benchmarks: WhisperX large-v2 reports 18.3% WER on VoxPopuli and 6.0% on mTEDx (both L1 Italian corpora), whereas aggregated L2 Italian corpora reach 22.4% (v2) and 22.1% (v3), i.e. clearly above native performance and approximately 3.7–3.9× the mTEDx baseline.

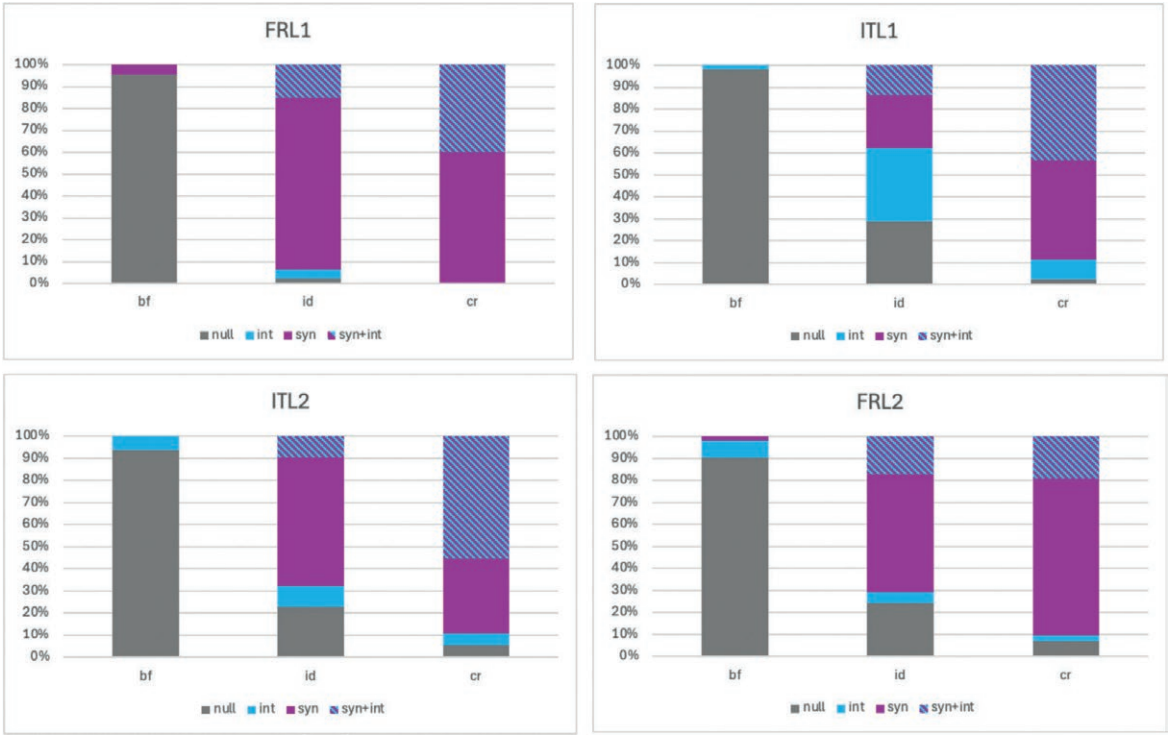

Figure 5. Distribution of marking strategies across focus types (broad = bf, identification = id, correction = cr) in the four groups: native French (FRL1), native Italian (ITL1), Italian learners of French (FRL2), and French learners of Italian (ITL2). Bars represent the proportion of tokens realised with no marking (null, grey), intonation only (int, blue), syntax only (syn, magenta), or combined syntax and intonation (syn+int, striped).

The evaluation further revealed a strong effect of learners' L1 background. Speakers with Romance L1s (including those in L2FOC) showed the lowest WERs (around 13.5%), whereas users with Germanic L1s reached much higher error rates (around 31.3% with WhisperX v2). Proficiency interacted with L1 in a non-linear way: among Romance speakers, advanced users did not consistently outperform intermediate ones (e.g. 17.2% vs. 18.4% in v2), showing that phonetic and prosodic transfer can outweigh proficiency alone.

L2FOC could contribute to this work in a valuable way thanks to its multi-layered content and structure: corpora including Romance languages as both L1 and L2 remain rare in ASR evaluation, and the corpus' structure (featuring controlled elicitation, parallel cross-linguistic design, and multiple proficiency levels) allowed effects of L1 similarity and proficiency to be examined in effective and varied ways.

4. CONCLUSIONS

The L2FOC corpus offers a robust and versatile resource for investigating a wide spectrum of research questions across syntax, prosody, phonetics, and second language acquisition. Its architecture enables fine-grained, crosslinguistic comparisons between native and

non-native speakers of Italian and French, with a design that systematically controls for linguistic variables while preserving naturalistic speech conditions.

One of the key strengths of the corpus lies in its high-quality audio recordings, which allow for detailed acoustic and prosodic analysis, including f0 tracking, phrasing, and prominence. The data are accompanied by manual orthographic transcriptions and multi-tier annotations, including phoneme- and syllable-level segmentations (with X-SAMPA), as well as word-level alignments. These annotations were generated using state-of-the-art tools (EasyAlign and WebMAUS), and were manually verified to ensure reliability and precision.

Crucially, the corpus includes rich metadata for each participant, covering age, gender, language background, proficiency level (via CEFR and CAF frameworks), education, and acquisition context. For non-native speakers, proficiency was assessed through a triangulated method combining cloze tests, self-assessments, and teacher-evaluated oral performance, providing researchers with a nuanced profile of learner variation.

In addition, the elicitation tasks were carefully designed to target specific focus types under varying levels of control and spontaneity. This makes the corpus especially valuable for exploring the syntax-prosody interface, crosslinguistic transfer, and the development of discourse-related structures in L2 speech. The inclusion of parallel tasks across two Romance languages further enables contrastive research on typological patterns and language-specific focus strategies.

At the same time, we are aware of certain limitations. As with most learner corpora, L2FOC was conceived with a specific research goal in mind-namely, the investigation of focus realisation in native and non-native Romance speech. This objective inevitably shaped both the structure of the dataset and the elicitation methods employed. While L2FOC can be fruitfully applied to other domains (e.g. pragmatics, discourse organization, phonetics), it was not designed, for example, to be perfectly phonetically or prosodically balanced across languages, nor to capture all possible dimensions of learner production. These constraints should be borne in mind when interpreting the data, yet they also point to possible directions for future extensions aimed at broadening its scope beyond its original aim.

Overall, the features of L2FOC make it a methodologically rigorous and richly annotated dataset that serves both empirical and applied purposes, supporting research in syntax and phonetics, second language acquisition, prosody, corpus linguistics, and speech technology development for under-resourced languages and varieties.

5. DATA AVAILABILITY

The L2FOC corpus is openly accessible at the repository *Ortolang* (officially recognized as a CLARIN B-centre) at the following link:

https://www.ortolang.fr/market/corpora/bianca-maria-de-paolis

It includes all audio files, aligned TextGrids, and accompanying metadata. Researchers are welcome to reuse the data under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

ACKNOWLEDGEMENTS

We wish to thank Ilinca Francisca Cojan for realising some of the drawings for the Picture Story task and the two illustrations used for the Picture Comparison task. We also wish to thank all participants who contributed to the recordings.

FUNDING

The data collection benefited from the support and infrastructures of the Laboratorio di Fonetica Sperimentale "Arturo Genre" (Università di Torino) and the Structures Formelles du Langage laboratory (Université Paris 8), where the totality of recordings were carried out. This research was funded by a doctoral fellowship from the Università di Torino and a Vinci Project grant (Université Franco-Italienne / Università Italo-Francese), both awarded to the first author.

REFERENCES

Aiello, Annachiara, Wenwei Dong, Catia Cucchiarini, and Helmer Strik. 2025. "Evaluating Automatic Speech Recognition on Non Native Italian". In *XXI convegno annuale AISV*, Urbino (Italy), 6-8 February 2025.

Allen, Will, Joan C. Beal, Karen P. Corrigan, Warren Maguire, and Hermann L. Moisl. 2007. "A Linguistic 'Time Capsule': The Newcastle Electronic Corpus of Tyneside English". In *Creating and Digitizing Language Corpora*, edited by Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, 16-48. London: Palgrave Macmillan.

Anastasio, Simona. 2021. *Parler de déplacement en L2: perspectives acquisitionnelles dans une approche translinguistique.* Roma: Aracne Editore.

Paul Boersma and David Weenink. 2021. *Praat: Doing Phonetics by Computer,* version 6.1.56, University of Amsterdam.

Büring, Daniel. 2010. "Towards a typology of focus realization". In *Information Structure*, edited by Malte Zimmerman and Caroline Féry, 177–205. Oxford: Oxford University Press.

CEFR. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment–Companion Volume.* Strasbourg: Council of Europe Publishing.

De Paolis, Bianca Maria. 2024. *Focus-induced variations in prosody and word order in native and non-native Italian and French*. PhD Thesis, Università di Torino / Université Paris 8.

Feldhausen, Ingo and Maria Del Mar Vanrell. 2014. "Prosody, Focus and Word Order in Catalan and Spanish: An Optimality Theoretic Approach". In *10th International Seminar on Speech Production.*

Gabriel, Christoph. 2010. "On Focus, Prosody, and Word Order in Argentinian Spanish. A Minimalist OT Account". *Revista Virtual de Estudos da Linguagem*, Special issue (4): 183–222. https://doi.org/10.5565/rev/isogloss.404

Gabriel, Christoph and Jonas Grünke. 2018. "Focus, prosody, and subject positions in L3 Spanish: analyzing data from German learners with Italian and Portuguese as heritage languages". In *Focus realization in Romance and beyond*, edited by Marco García García and Melanie Uth, 358–86. Amsterdam: John Benjamins.

Gass, Susan and Larry Selinker. 2001. *Second Language Acquisition: An Introductory Course.* Mahwah, NJ: Erlbaum.

Geertzen, Jeroen, Theodora Alexopoulou, and Anna Korhonen. 2014. "Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat)". In *Selected Proceedings of the 2012 Second Language Research Forum*, edited by Ryan T. Miller, 240–54. Somerville, MA: Cascadilla Proceedings Project.

Gilquin, Gaetanelle, Sylvie De Cock, and Sylviane Granger. 2010. *The Louvain International Database of Spoken English Interlanguage (LINDSEI)*. Louvain-La-Neuve: Presses universitaires de Louvain.

Goldman, Jean-Philippe. 2011. "Easyalign: An Automatic Phonetic Alignment Tool under Praat". In *Proceedings of InterSpeech,* 3233-6. https://doi.org/10.21437/Interspeech.2011-815

Jarvis, Scott and Aneta Pavlenko. 2010. *Crosslinguistic Influence in Language and Cognition*. London: Routledge.

Granger, Sylviane, Maïté Dupont, Fanny Meunier, Hubert Naets, and Magali Paquot. 2020. *The International Corpus of Learner English*. Version 3. Louvain-la-Neuve: Presses universitaires de Louvain. Université catholique de Louvain.

Hilton, Heather. 2009. "Annotation and analyses of temporal aspects of spoken fluency". *CALICO Journal*, 26: 644–61.

Kisler, Thomas, Uwe Reichel, and Florian Schiel. 2017. "Multilingual Processing of Speech via Web Services". *Computer Speech and Language*, 45: 326–47. https://doi.org/10.1016/j.csl.2017.01.005

Krifka, Manfred. 2008. "Basic notions of information structure". *Acta Linguistica Hungarica,* 55(3–4): 243–76.

Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topics, Focus, and the Mental Representations of Discourse Referents.* Cambridge: Cambridge University Press.

Lenneberg, Eric. H. 1967. *Biological Foundations of Language*. New York: Wiley.

Mackey, Alison, and Susan M. Gass. 2016. *Second Language Research: Methodology and Design*. New York: Routledge.

Myles, Florence. 2006, *French Learner Language Oral Corpora (FLLOC)*, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2495

Norris, John and Lourdes Ortega. 2009. "Towards an organic approach to investigating CAF in instructed SLA: The case of complexity". *Applied Linguistics*, 30(4): 555–78. https://doi.org/10.1093/applin/amp044

OpenCLC. 2017. Distributed by Lexical Computing Limited on behalf of Cambridge University Press and Cambridge English Language Assessment.

Pallotti, Gabriele. 2009. "CAF: defining, refining and differentiating constructs". *Applied Linguistics*, 30(4): 590–601. https://doi.org/10.1093/applin/amp045

Tremblay, Annie, and Meryl D. Garrison. 2010. "Cloze Tests: A Tool for Proficiency Assessment in Research on L2 French". In *Selected Proceedings of the 2008 Second Language Research Forum*, edited by Matthew T. Prior, 73–88. Somerville, MA: Cascadilla Proceedings.

Savy, Renata. 2006. *Specifiche per la trascrizione ortografica annotata dei testi raccolti. Progetto CLIPS-W1-a4*. Retrieved at: https://it.scribd.com/document/314785610/11-Specifiche-Trascrizione-Ortografica.

Turco, Giuseppina, Christhine Dimroth, and Bettina Braun. 2013. "Intonational means to mark verum focus in German and French. Language and Speech", 56(4): 460–90. https://doi.org/10.1177/0023830912460506

Vedder, Ineke. 2008. "Competenza pragmatica e complessità sintattica in italiano l2: l'uso dei modificatori nelle richieste". *Linguistica e Filologia*, 25(1): 99–123.

Zhang, Jie and Hongyin Tao. 2018. "Corpus-based research in Chinese as a second language". In *The Routledge Handbook of Chinese Second Language Acquisition*, edited by Chuanren Ken, 48–62. London; New York: Routledge.

APPENDIX

I. Read aloud
I.1 Italian

Dialogo 1. Leggi la parte di Giulio!
Giulio è a casa sua e sta cucinando il pranzo per lui e la sua amica Nina. Nina arriva e Giulio le apre la porta.

Giulio: Ciao Nina! Benvenuta.
Nina: Ciao Giulio! Che profumino! Hai cucinato le lasagne?
Giulio: No, ho cucinato la parmigiana.
Nina: Davvero? L'hai fatta tu?
Giulio: Sì, la parmigiana l'ho fatta io. È Giovanna che ha fatto il dolce, invece.
Nina: Chi ha fatto il dolce, scusa?
Giulio: Giovanna ha fatto il dolce.
Nina: Mi ricordo di lei. La settimana scorsa ha preparato la torta al limone. E oggi che cosa ha fatto?
Giulio: Oggi ha fatto la torta caprese.
Nina: Peccato! La fa benissimo la torta al limone.
Giulio: Ma no, è la torta caprese che fa benissimo. Vedrai che non mi sbaglio.
Nina: Allora non vedo l'ora di mangiarla!

Dialogo 2. Leggi la parte di B!
A e B sono in macchina, A sta guidando.

A: Dobbiamo avvertire del nostro ritardo. Preferisci telefonare a Giulia o a Elena?
B: Preferisco telefonare a Giulia.
A: D'accordo. Fallo subito.
B: L'ultima volta, però, è a Giulia che ho telefonato. Magari è meglio cambiare.
A: Fai come vuoi, basta che avverti.
B: Ecco fatto.
A: Ma come? Hai già mandato il messaggio a Elena?
B: No, ho telefonato a Elena.
A: Sei stato rapido. Non è che mi dici una bugia?
B: Che ho telefonato è vero. Ma alla fine ho richiamato Giulia!

Dialogo 3. Leggi la parte di B!
Dal fruttivendolo.

A: Buongiorno! Le è piaciuta la frutta che ha comprato ieri?
B: Sì, mi è piaciuta molto. I mandarini li ho mangiati già tutti.
A: Benissimo. Ne vuole ancora?
B: No, oggi vorrei prendere dei limoni.
A: Preferisce il cedro o proprio il limone?
B: Non il cedro, grazie. Sono i limoni che mi servono oggi.
A: Eccoli qui. Vuole altro? Delle arance, visto che le ha mangiate tutte?
B: Ho ancora molte arance: i mandarini sono finiti.
A: È vero, che sciocca, me lo ha appena detto. A lei capita che la memoria non funzioni?
B: La memoria, quella funziona sempre.
A: Forse dovrei mangiare più spinaci. Gli spinaci aiutano la vista o la memoria?
B: Gli spinaci aiutano la memoria.
A: Ha ragione! Ecco qui la sua frutta. Arrivederci!
B: Arrivederci!

I.2 French

Dialogue 1. Lisez la partie de Jules !
Jules est chez lui et prépare le déjeuner pour lui et son amie Nina. Nina arrive et Jules lui ouvre la porte.

Jules : Salut Nina ! Bienvenue.
Nina : Salut Jules! Ça sent bon ! Tu as préparé un gâteau ?
Jules : Non, j'ai préparé de la marmelade.
Nina : Vraiment ? C'est toi qui l'as faite ?
Jules : Oui, la marmelade, je l'ai faite moi-même. C'est Jean-Marie qui a fait le dessert, par contre.
Nina : Pardon, qui a fait le dessert ?
Jules : Jean-Marie a fait le dessert.
Nina : Je me souviens de lui. La semaine dernière, il a fait une tarte au citron. Et aujourd'hui, qu'est-ce qu'il a fait ?
Jules : Aujourd'hui, il a fait un gâteau meringué.
Nina : Dommage ! Il la fait très bien, la tarte au citron.
Jules : Mais non, c'est le gâteau meringué qu'il fait très bien. Tu vas voir, je ne me trompe pas.
Nina : Alors j'ai hâte de le manger !

Dialogue 2. Interprétez B !
A et B sont dans la voiture, A conduit.

A : Nous devons prévenir de notre retard. Tu préfères appeler Julie ou Hélène ?
B : Je préfère appeler Julie.
A : D'accord. Fais-le de suite.
B : La dernière fois, par contre, c'est à Julie que j'ai téléphoné. Peut-être qu'il serait mieux de changer.
A : Tu peux faire ce que tu veux, pourvu que tu les préviennes.
B : C'est fait.
A : Comment ça ? Tu as déjà envoyé le message à Hélène ?
B : Non, j'ai téléphoné à Hélène.
A : Tu as été rapide. Tu ne me dis pas un mensonge ?
B : Que j'ai téléphoné, c'est vrai. Mais au final j'ai rappelé Julie !

Dialogue 3. Interprétez B !
A l'épicerie.

A : Bonjour ! Avez-vous aimé les fruits que vous avez achetés hier ?
B : Oui, j'ai beaucoup aimé. Les mandarines, je les ai déjà toutes mangées.
A : Très bien. Vous en voulez encore ?
B : Non, aujourd'hui je voudrais acheter des melons.
A : Préférez-vous la pastèque ou les melons ?
B : Je ne veux pas de pastèque, merci. Ce sont les melons que je veux aujourd'hui.
A : Les voici. Ce sera tout ? Voulez-vous des oranges, puisque vous les avez toutes mangées ?
B : J'ai encore beaucoup d'oranges : les mandarines sont terminées.
A : C'est vrai, je suis bête, vous venez de le dire. Cela vous arrive-t-il que votre mémoire ne fonctionne pas ?
B : Ma mémoire, elle marche toujours bien.
A : Peut-être que je devrais manger plus d'épinards. Les épinards aident-ils à la vue ou à la mémoire ?
B : Les épinards aident à la mémoire.
A : Vous avez raison ! Voici vos fruits. Au revoir !
B : Au revoir !

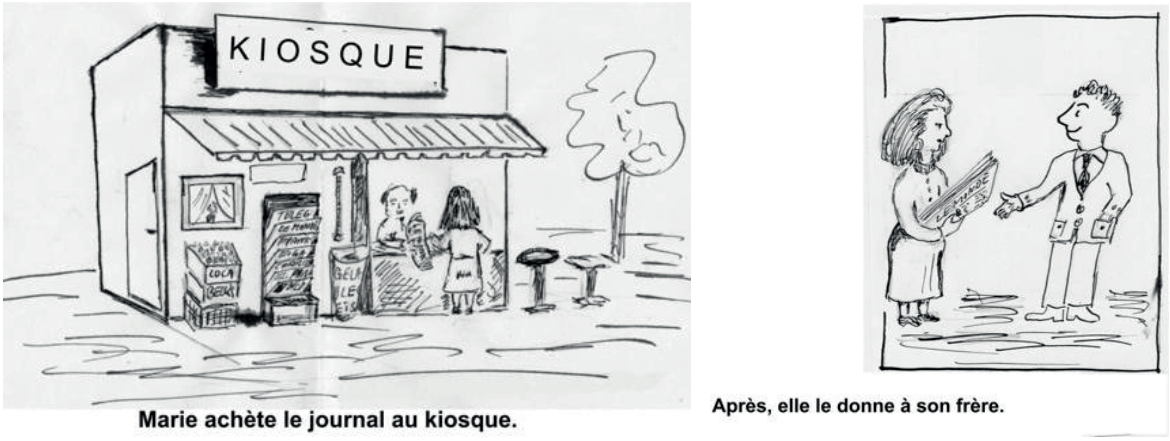

Figure I. Baseline-picture slides for story n. 1 (French version): picture 1a (left) and picture 1b (right).

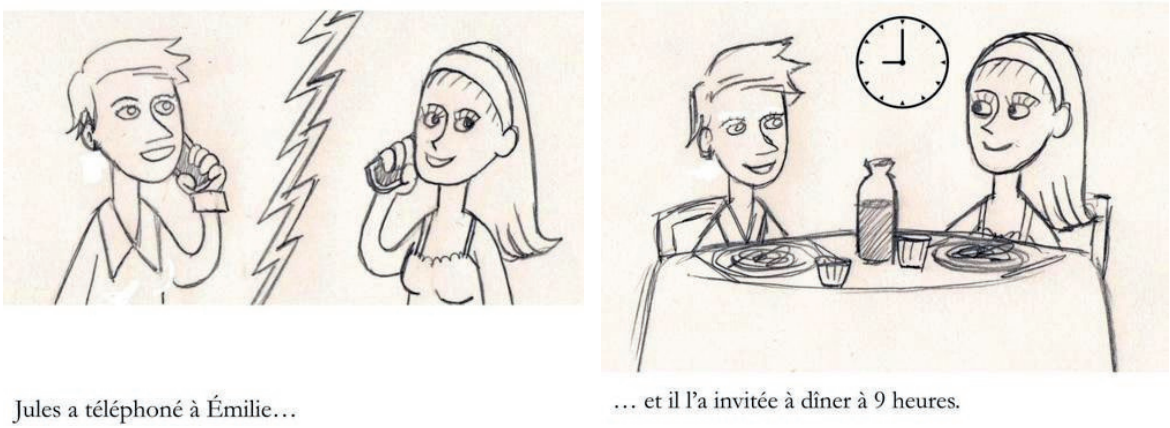

Figure II. Baseline-picture slides for story n. 2 (French version): picture 2a (left) and picture 2b (right).

II. Picture story
II.1 French

Stimulus 1a.
Qu'est-ce qu'il se passe ici ? *Broad focus*
Qu'est-ce que Marie achète au kiosque ? *Identification focus; object*
Qui achète le journal au kiosque ? *Identification focus; subject*
Marie achète des mots croisés au kiosque, non ? *Correction focus; object*
Où est-ce que Marie achète le journal ? *Identification focus; adverbial*
C'est Marie qui achète le journal, n'est-ce pas ? *Confirmation*
Que fait Marie avec le journal ?
Marie achète le journal au supermarché, non ?
Que fait Marie au kiosque ?
Marie vole le journal au kiosque, non ?

Stimulus 1b.
Qu'est-ce qu'il se passe ici ?
Que fait Marie ?
Qui est en train de donner le journal à son frère ?

À qui Marie donne-t-elle le journal ?
Julie donne le journal à son frère, n'est-ce pas ?
Qu'est-ce que Marie donne à son frère ?
Marie donne à son frère des mots croisés, non ?
Marie donne le journal au frère de Julie, n'est-ce pas ?

Stimulus 2a.
Qu'est-ce qu'il s'est passé ici ?
À qui a téléphoné Jules?
Jules à téléphoné à Christine, non?
Qui a téléphoné à Émilie?
Qu'est-ce que Jules a fait avec Émilie?
Marc a téléphoné à Émilie, n'est-ce pas?
Jules a envoyé un message à Émilie, non?

Stimulus 2b.
Qu'est-ce qu'il s'est passé ici?
À quelle heure Jules a invité à dîner Émilie ?
Jules a invité à dîner Émilie à 7 heures, n'est-ce pas ?
Stéphane a invité à dîner Émilie à 9 heures, non?

II.2 Italian

Stimulus 1a.
Che cosa succede qui?
Che cosa compra Maria in edicola?
Chi compra il giornale in edicola?
Maria compra una rivista di cruciverba in edicola, giusto?
Dove compra il giornale Maria?
È Maria che compra il giornale, giusto?
Che cosa fa Maria con il giornale?
Maria compra il giornale al supermercato, no?
Che cosa fa Maria in edicola?
Maria ruba il giornale in edicola, giusto?

Figure III. Baseline-picture slides for story n. 1 (Italian version): picture 1a (left) and picture 1b (right).

Figure IV. Baseline-picture slides for story n. 2 (Italian version): picture 2a (left) and picture 2b (right).

Stimulus 1b.
Che cosa succede qui?
Che cosa fa Maria?
Chi sta dando il giornale a suo fratello?
A chi dà il giornale Maria?
Giulia dà il giornale a suo fratello, vero?
Che cosa dà Maria a suo fratello?
Maria dà a suo fratello una rivista di cruciverba, giusto?
Dà il giornale al fratello di Giulia, no?

Stimulus 2a.
Che cosa è successo qui?
A chi ha telefonato Giulio?
Giulio ha telefonato a Cristina, giusto?
Chi ha telefonato ad Emilia?
Che cosa ha fatto Giulio con Emilia?
È Marco che ha telefonato ad Emilia, vero?
Giulio ha mandato un messaggio a Emilia, no?

Stimulus 2b.
Che cosa è successo qui?
A che ora Giulio ha invitato a cena Emilia?
Giulio ha invitato Emilia a cena alle sette, giusto?
Stefano ha invitato Emilia a cena al ristorante, no?