

ELITE-AI

Integrating AI in ESP and Foreign Language Instruction

A Manual of Good Practices for Applied Linguistics Researchers

Partner institutions: UNITBV (Romania) · UM (Slovenia) · UNIUD (Italy) · UNIPU (Croatia)

Erasmus+ Grant No. 2024-1-RO01-KA220-HED-000248845

elite-ai.unitbv.ro

Table of Contents

About This Manual	3
Section 1 – Orienting the Research: AI as Object of Study, Not Solution	3
1.1 Framing the Research Question.....	4
1.2 Defining the Pedagogical Role of AI Before Selecting Tools	4
Section 2 – Designing Observation Instruments for AI-Enhanced Classrooms	5
2.1 What Observation Sheets Should Record.....	5
2.2 Triangulating Observation with Other Data Sources.....	6
Section 3 – Skills-Based Research Design in AI-Integrated Language Instruction	6
3.1 Reading Research with AI Tools	6
3.2 Writing Research with AI Tools	7
3.3 Speaking and Listening Research with AI Tools.....	7
Section 4 – Ethical Dimensions and Quality Assurance in AI-Integration Research	8
4.1 Equity and Access as Research Variables	8
4.2 Data Privacy and Tool Transparency	8
4.3 Academic Integrity as a Research Question, Not Only a Compliance Issue	9
4.4 Quality Assurance in Multi-Institutional Research	9
Section 5 – Lesson Plans as Research Outputs: Design, Documentation, and Dissemination	9
5.1 The Five-Stage Architecture as an Analytical Frame.....	10
5.2 Documenting Lesson Plans for Research Purposes	10
5.3 From Lesson Plan to Publication.....	11
Section 6 – From Classroom Data to Publication: Analytical and Writing Strategies	11
6.1 Selecting the Right Analytical Unit	11
6.2 Qualitative and Mixed-Method Approaches.....	12
6.3 Writing the Article: Positioning within the Field	12
Consolidated Researcher’s Checklist	13
Research Design.....	13
Data Collection.....	13
Analysis and Writing	13
Dissemination	14
A Note on Transferability	14

About This Manual

This manual is addressed to researchers working in applied linguistics, language pedagogy, and technology-enhanced language learning. It follows the theoretical apparatus and the tool-by-tool analyses performed by the members of the project's research team, distilling the collective experience into a set of concrete, transferable orientations – decisions and design choices that shaped the work from research question to manual releasing, and that are likely to make the subject of any study at the intersection of artificial intelligence and language instruction.

The manual is organised around six core areas of practice. Each section opens with a research rationale drawn from the project's findings, then moves to actionable guidance. A researcher's checklist closes each section. A consolidated checklist at the end of the document covers the full arc from project design to dissemination.

Who Should Use This Manual

Researchers designing empirical studies on AI tools in language classrooms; teams conducting Erasmus+ or similar collaborative educational research; educators transitioning from classroom practice into systematic inquiry; postgraduate students in applied linguistics or educational technology.

Section 1 – Orienting the Research: AI as Object of Study, Not Solution

One of the most consequential decisions any research team makes before the first data collection point is the decision about what, exactly, AI is in the study. The ELITE-AI project worked through this question explicitly, and the conclusions return to it repeatedly: AI is a tool, not a teacher, not an educational setting, and not a manual. This distinction is not semantic. It has direct methodological consequences.

When AI is treated as a solution – when the research question is, essentially, ‘does AI improve learning?’ – the study is almost always underpowered to answer it. The AI variable is too broad, the contexts too heterogeneous, the tools too different from one another to produce transferable conclusions. What the ELITE-AI team found, across four institutions and multiple ESP fields, is that the pedagogically meaningful unit of analysis is not the tool itself but the instructional decision behind its deployment: which

stage of the lesson plan it enters, what skill it targets, how the teacher edits or frames the output, and what the learner does with the feedback it generates.

1.1 Framing the Research Question

A productive research question in this field specifies the tool category or function (e.g., automated writing evaluation, adaptive vocabulary practice, conversational AI for speaking), the instructional stage (lead-in, teaching, practice, feedback, or independent work), the skill or competence targeted (lexical accuracy, genre awareness, oral fluency, listening comprehension), and the learner population, including proficiency level and ESP domain.

Questions that satisfy these four parameters are genuinely answerable within a single project cycle. Questions that do not – for instance, ‘What is the impact of AI on ESP learning?’ – require either a systematic review design or substantial narrowing before data collection begins.

Less productive framing	More productive framing
Does AI help ESP students?	Does AI-generated vocabulary feedback in a corpus tool improve collocational accuracy in legal writing at B2 level?
Is ChatGPT useful for language teaching?	How do teachers at C1-level ESP classes edit and frame ChatGPT output when teaching genre conventions in cover letter writing?
What do students think of AI tools?	Do students in mobility-themed ESP courses report higher autonomy when AI feedback is provided during, rather than after, practice tasks?

1.2 Defining the Pedagogical Role of AI Before Selecting Tools

Orientation Principle 1

The pedagogical function precedes the tool. Tools are chosen because they perform a specific instructional function at a specific lesson stage. They are not studied because they are available or popular.

The ELITE-AI research catalogues over one hundred AI tools across the four language skills. This richness is also a methodological risk: tool selection can drive research design rather than the other way around. The safer sequence runs in the opposite direction. Define what the tool must do pedagogically – generate authentic texts for reading analysis, provide pronunciation feedback without a human interlocutor, assist

in abbreviation recognition in specialised corpora – and then identify tools that perform that function.

Section 2 – Designing Observation Instruments for AI-Enhanced Classrooms

The primary data collection instrument in WP3 of the ELITE-AI project is the observation sheet, completed by instructors and students alike across all three stages of the instructional continuum: teaching, learning (student-side), and assessment. The decision to use observation sheets rather than questionnaires alone reflects a key methodological insight: self-reported satisfaction data and behavioural observation data diverge in ways that matter for applied linguistics research. Observation instruments capture what actually happens in the classroom; surveys capture what participants believe or feel happened.

2.1 What Observation Sheets Should Record

In the ELITE-AI design, each observation sheet is organised around lesson aim and lesson stage. Within that framework, instructors record positive outcomes, negative outcomes or friction points, cultural or domain-specific particularities, notes on personalised learning styles or individual learner behaviour, and any AI tool behaviour that departed from anticipated function. This structure – anchored to stage and aim, not to the tool alone – produces data that could be read across institutions and across ESP fields without losing local specificity.

Researchers designing similar instruments should resist the temptation to produce a single generic observation sheet. A sheet designed for the feedback stage, where the teacher and student interact around AI-generated corrections, looks quite different from one designed for the practice stage, where students engage with AI tools with less direct teacher mediation.

Design Principle for Observation Instruments

Each observation sheet should be anchored to a specific instructional stage and a specific skill focus. A sheet that covers the full lesson in a single instrument tends to produce vague, stage-indeterminate data that cannot support fine-grained analysis.

2.2 Triangulating Observation with Other Data Sources

Observation sheets alone are a first-person instructor account. They need triangulation. In the ELITE-AI project, this is achieved through student questionnaires after each activity cycle, and through the cross-institutional comparison that the standardised sheet format made possible. Researchers working in smaller-scale studies can triangulate through think-aloud protocols during AI-assisted practice tasks, through analysis of the AI tool's output itself (what the tool actually produced in response to student input), or through the learner's edited version of that output, which reveals the gap between AI generation and pedagogically valid text.

Orientation Principle 2

Observation data and self-report data should be collected in parallel and compared. Where they diverge – and in AI-integrated classrooms, they often will – the divergence is itself a research finding, not a methodological problem to be resolved.

Section 3 – Skills-Based Research Design in AI-Integrated Language Instruction

The ELITE-AI research perspective organises its tool analysis by language skill – reading, writing, listening, speaking – and its lesson plans follow the same architecture. This is not arbitrary. Skills-based organisation maps onto both the communicative approach that most contemporary ESP instruction follows and onto the types of AI tool that exist: reading comprehension tools function differently from pronunciation feedback tools, which function differently from conversational agents, which function differently from corpus and lexicography tools. A research design that conflates skills conflates tool categories and loses analytical purchase.

3.1 Reading Research with AI Tools

AI tools for reading in ESP contexts – such as Revita (University of Helsinki), Diffit, Text Inspector, and PrepAI – typically operate through adaptive text difficulty, automated cloze generation, vocabulary highlighting, and real-time feedback on comprehension tasks. Research questions in this area tend to concern the quality of the AI-generated adaptation (does automated text simplification preserve specialised register?), the learner's engagement with AI-generated vocabulary support (does

clicking for instant translation reduce or improve long-term retention?), and the teacher's role in selecting and contextualising source texts before AI processing.

A methodologically sound reading study will specify whether it targets literal comprehension, inferential comprehension, or vocabulary acquisition in context, since AI tools perform differently across these sub-skills and the evaluation criteria will differ accordingly.

3.2 Writing Research with AI Tools

Writing is the skill where the teacher's editorial role after AI use is most visible and most consequential. Tools such as Grammarly, Sketch Engine, GravityWrite, and ChatGPT operate at different levels of the writing process – surface-level error correction, corpus-based collocation analysis, text generation, and genre feedback respectively. A study that treats these as equivalent writing tools will produce uninterpretable results.

The ELITE-AI lesson plans for writing demonstrate the principle of layered tool use: Sketch Engine at the research and planning stage (corpus-based collocation extraction), ChatGPT at the drafting stage (genre model generation for student analysis), and Grammarly at the revision stage (surface accuracy). Research that maps tool use onto writing stage – rather than treating AI as a single intervention – produces data that is both more valid and more pedagogically useful.

3.3 Speaking and Listening Research with AI Tools

Speaking and listening studies with AI tools raise distinct challenges around ecological validity. Conversational AI platforms – Cambly, Talkpal AI, Heero.me, Eigo.AI – simulate interactional contexts, but the interlocutor is not human. Pronunciation feedback tools – SpeechAce, Podcastle – provide acoustic analysis, but the criteria for 'accuracy' are typically tied to native-speaker norms that may not be appropriate for English as a Lingua Franca contexts.

Researchers should therefore specify what variety of English the tool uses as its reference model and whether that model is appropriate for the learner population and professional context. An ESP course for maritime engineers whose working language is ELF should not be evaluated against a monolingual native-speaker standard, regardless of what the AI tool's default setting assumes.

Skill Area	Key research design considerations
Reading	Text difficulty calibration; vocabulary retention vs. immediate look-up; role of authentic vs. AI-simplified texts in ESP register acquisition
Writing	Stage-specific tool use (planning, drafting, revision); teacher's editorial role; genre awareness vs. surface accuracy
Speaking	ELF vs. native-speaker reference norms; conversational vs. monologic output; anxiety reduction claims require longitudinal data
Listening	Accent diversity in AI-generated audio; captioning as comprehension support vs. dependency; intercultural listening vs. linguistic accuracy

Section 4 – Ethical Dimensions and Quality Assurance in AI-Integration Research

The ELITE-AI project dedicates a full chapter section to ethical considerations in AI-supported language education. For researchers, these considerations are not merely a box to be ticked in an ethics review form. They affect research design, data interpretation, and the transferability of findings.

4.1 Equity and Access as Research Variables

AI tool access is not homogeneous across learner populations, institutions, or national contexts. A study conducted at a well-resourced European university with reliable broadband, institutional licences for corpus tools, and students with personal devices is not automatically transferable to contexts where any of these conditions are absent. Researchers should report access conditions explicitly and consider them when interpreting engagement or satisfaction data. Low engagement with a tool may reflect cost barriers, connectivity limitations, or device constraints rather than learner attitude or tool quality.

4.2 Data Privacy and Tool Transparency

Several AI tools used in language education collect learner data – writing samples, pronunciation recordings, reading behaviour logs – that may be stored on third-party servers outside the institution's data protection perimeter. Researchers must verify the data handling terms of each tool they introduce into a study and communicate these terms to participants in plain language. Informed consent forms for AI-integrated studies should specify which tools are used, what data each tool collects, how that data is stored, for how long, and whether it is used for model training.

In cross-institutional research involving multiple national legal frameworks – as in the ELITE-AI consortium – data handling agreements must be negotiated at the consortium level, not left to individual researchers.

4.3 Academic Integrity as a Research Question, Not Only a Compliance Issue

The ELITE-AI project addresses the risk of over-reliance and academic dishonesty in AI-integrated language instruction. For researchers, this creates an interesting inversion: the conditions that make AI use educationally effective – structured task design, explicit pedagogical framing, teacher mediation – are precisely the conditions that reduce the likelihood of dishonest use. A study that documents how these conditions function empirically contributes more to the field than a study that simply measures how often students use AI tools without permission.

Orientation Principle 3

The ethical dimensions of AI-integrated research are themselves research data. Documenting access conditions, tool transparency, and the teacher's role in framing AI use produces findings that are as theoretically significant as the learning outcomes data.

4.4 Quality Assurance in Multi-Institutional Research

The ELITE-AI project uses a standardised observation sheet format across four institutions to ensure data comparability. Without this standardisation, cross-institutional synthesis would not be impossible. Research consortia should invest time in instrument alignment before data collection begins. A pilot phase in which each partner completes the same instrument on the same task, then compares their records, reveals interpretive divergences that methodological training alone cannot anticipate.

Section 5 – Lesson Plans as Research Outputs: Design, Documentation, and Dissemination

One of the distinctive outputs of the ELITE-AI project is the design of AI-enhanced lesson plans structured across four language skills and five instructional stages: lead-in, teaching, practice, feedback, and homework. Twenty lesson plans were produced for B2–C1 level learners across four ESP domains (social media marketing and tourism, mobility and administrative language, legalese and criminal justice, education

and sport). For researchers, these lesson plans are not merely curriculum products – they are theoretical arguments made concrete.

A lesson plan that assigns AI tools to specific stages, specifies the skill focus and interaction type, and records the time allocation for each activity is, in effect, a documentation of a pedagogical decision-making process. When that process is recorded with enough precision that another teacher in a different institution could replicate it – which is the standard the ELITE-AI plans aimed at – it becomes a transferable research output, not merely local curriculum material.

5.1 The Five-Stage Architecture as an Analytical Frame

The lead-in / teaching / practice / feedback / homework structure serves a dual function: it organises instruction and it organises data collection. Researchers who adopt this frame can compare AI tool use across stages – and the ELITE-AI data suggest this comparison is informative. AI tools at the lead-in stage (Mentimeter for brainstorming, Educaplay for vocabulary activation) serve attention and activation functions. Tools at the teaching stage (Krajšavar for abbreviation analysis, ChatGPT for genre model generation) serve explanation and demonstration functions. Tools at the practice stage serve skill rehearsal under conditions of support. Tools at the feedback stage (Grammarly, SpeechAce, Classtime.com) provide formative evaluation. Tools at the homework stage support autonomous extension of classroom work.

The research implication is that the same tool deployed at different stages may produce different learning outcomes and different patterns of engagement. A study that controls for stage produces more interpretable data than one that does not.

5.2 Documenting Lesson Plans for Research Purposes

Documentation Standard

A lesson plan is research-grade when a researcher at a different institution, working in the same ESP domain and at the same CEFR level, can replicate the instructional sequence from the written record alone, without further clarification from the original designer.

A lesson plan designed for research purposes should include: the ESP domain and learner profile (year of study, level, group size, lesson duration), the specific skill focus and vocabulary or grammar target for each stage, the name and function of each AI tool introduced, the expected interaction type (individual, pair, group), the time

allocation for each stage, and a space for the instructor's post-hoc observation notes. The ELITE-AI template integrates all these elements and is available at elite-ai.unitbv.ro.

5.3 From Lesson Plan to Publication

The ELITE-AI project maps directly from observation sheets (primary data) to six peer-reviewed articles (two per activity in WP3: teaching, learning, assessment). The bridge between data collection and publication runs through the lesson plan: the plan defines what was done, the observation sheet records what happened, and the article interprets the gap between intention and outcome. Researchers who do not maintain this three-way alignment – plan, observation, interpretation – often find that their data do not have a clear analytical home when they sit down to write.

Section 6 – From Classroom Data to Publication: Analytical and Writing Strategies

The ELITE-AI consortium committed to publishing six specialised articles across the three WP3 activities, distributed across self-selected partner pairs. The decision to make publication a project deliverable – rather than an optional outcome – shapes the research culture in productive ways. It requires that data collection instruments be designed with analysis in mind from the outset, that the thematic focus of each article be agreed upon before data collection ends, and that writing begin while the implementation experience is still fresh.

6.1 Selecting the Right Analytical Unit

Applied linguistics research on AI integration often plunges at the analysis stage because the analytical unit was not specified in the research design. The question ‘What did students think of AI tools?’ is not answerable in a way that produces transferable knowledge, because thinking encompasses attitude, perceived utility, emotional response, and behavioural intention – constructs that require different instruments and different analytical approaches.

The ELITE-AI project's most analytically precise frames are those that focus on a specific tool function (adaptive text difficulty, collocational pattern generation, abbreviation recognition) within a specific instructional context. Researchers should identify their analytical unit at the design stage: is it the learner's linguistic output before and after AI tool use? The teacher's editorial decisions? The AI tool's output

quality relative to pedagogical criteria? The learner's self-assessment accuracy relative to teacher assessment?

6.2 Qualitative and Mixed-Method Approaches

Most existing research on AI in language education uses questionnaire-based survey designs. This creates a gap in the literature that mixed-method and qualitative studies can address productively. Instructor observation notes, think-aloud protocols, interview data, and discourse analysis of student-AI interaction logs are all underused in the field. The ELITE-AI project's observation sheets generate qualitative data – instructor reflections, domain-specific notes, individual learner observations – that support thematic analysis and can be combined with satisfaction scores in a mixed-method design.

Researchers should consider whether their research questions call for measurement (how much change? how often? how many?) or understanding (why? in what circumstances? with what meanings?). Many of the most important questions in this field – about teacher identity, about learner autonomy, about the conditions under which AI feedback is trusted or rejected – are understanding questions that survey instruments cannot answer.

6.3 Writing the Article: Positioning within the Field

An article on AI in language education needs to position itself within at least two bodies of literature: the applied linguistics literature on the skill or competence in question (reading comprehension, writing genre, oral fluency, listening strategy), and the AIED (Artificial Intelligence in Education) literature on the type of tool or function under investigation. Articles that position themselves only within one of these bodies tend to be published in journals where they speak only to half their potential audience.

Orientation Principle 4

Position the article at the intersection of language pedagogy and educational technology, not within one field alone. The most cited work in this area speaks to both audiences: it offers a finding about language learning that is also a finding about how AI tools function in educational contexts.

The ELITE-AI framework's reference base spans both traditions – from Krashen's input hypothesis and Vygotsky's Zone of Proximal Development on the language acquisition side, to Hwang et al. (2020) and Solorzano et al. (2024) on the AIED side.

Researchers should build their literature review across both bodies of work, ensuring that the article's contribution is visible to readers in both traditions.

Consolidated Researcher's Checklist

The following checklist covers the full arc of a research project at the intersection of AI and ESP/FL instruction. It is based on the design choices made across the ELITE-AI project and on the analytical lessons to be documented for the practical stage.

Research Design

1. Is the AI tool's pedagogical function specified independently of the tool's name or brand?
2. Is the research question anchored to a specific instructional stage, skill, and learner profile?
3. Are the claims the study will make – about learning outcomes, engagement, autonomy, or teacher practice – matched to instruments that can actually produce those claims?
4. In multi-institutional research: has a pilot phase been run to align interpretations of observation instruments across teams?

Data Collection

5. Are observation sheets anchored to lesson stage, not to the lesson as a whole?
6. Is self-report data triangulated with at least one form of behavioural or output data?
7. Are access conditions (device availability, connectivity, tool licensing) documented as potential confounds?
8. Have data privacy and consent procedures been verified for each AI tool used, including its data storage and third-party sharing practices?

Analysis and Writing

9. Is the analytical unit specified – learner output, teacher decision, tool function, or learner self-assessment?
10. Does the literature review span both the language pedagogy tradition and the AIED tradition?
11. Are divergences between observation data and self-report data treated as findings rather than methodological problems?

12. Does the article's contribution speak to both language educators and educational technology researchers?

Dissemination

13. Are lesson plans documented to replication standard – detailed enough that another researcher in the same ESP domain could reproduce the instructional sequence?
14. Have ethical considerations – access equity, tool transparency, academic integrity – been reported as part of the findings, not only as compliance statements?
15. Are recommendations written for practitioners who are not researchers – teachers, curriculum designers, institution-level policymakers – as well as for the research community?

A Note on Transferability

The ELITE-AI project is conducted across four European countries, four languages of instruction, and four distinct ESP fields, by nine academics with an average of over twenty years of professional experience in language acquisition research. The present manual reflects that breadth. At the same time, no manual based on a specific project can anticipate every context, every learner population, or every AI tool that will exist by the time a researcher reads it.

What transfers from this project is not a protocol but a set of analytical habits: specifying the pedagogical function before the tool, distinguishing between measurement and understanding questions, keeping the lesson plan, the observation instrument, and the publication aim in alignment, and treating the divergence between expectation and outcome as the most productive starting point for analysis. These habits apply regardless of which specific tools a future study uses or which ESP domain it works in.

The project's website at www.elite-ai.unitbv.ro contains all project's published outputs. Researchers who wish to build on, adapt, or critically extend the ELITE-AI framework are encouraged to engage with these resources directly.