Expert Evaluation Protocol (Heuristic Evaluation + Human–AI Checklist + Light Persona-Based Walkthrough)

Goal

Evaluate a prototype MT interface augmented with LLM-generated support, focusing on:

1. Usability and interaction quality using Nielsen's usability heuristics
2. Human–AI interaction risks using a mini-checklist derived from Amershi et al.'s "Guidelines for Human–AI Interaction" (CHI 2019)
3. Persona-based task evaluation through light walkthrough

Out of scope here (already assessed elsewhere):
– pedagogical appropriateness of the instructional content
– translation accuracy/quality as such

1) Method and materials

Method

- Heuristic Evaluation (Nielsen) + Task-based walkthrough
- Targeted Human–AI checklist (Amershi et al., CHI 2019)
- Persona impact tagging + light persona-based walkthrough

Materials provided to each expert

- Prototype access link (the prototype version works without login, assuming a default user profile). In this evaluation: Mother tongue: Italian, English proficiency: B1, first use of the ILA Interface
- Short task script
- Persona cards (each expert receives the same set of cards)

2) Severity rating (0–4)

- **0** No usability problem
- **1** Minor issues
- **2** Moderate (causes friction or confusion)
- **3** Major (causes errors or significantly blocks progress)
- **4** Critical (prevents task completion and/or strongly undermines trust)

3) Tasks (what experts do)

Perform tasks in order. Use **think-aloud**: state what you expect and what you find.

T1 — Onboarding & settings

1. Open the user interface using the provided link.
2. Set Language source (ITA) and target (ENG) among those available.
3. Check whether settings are clear, editable, and persistent as expected.

T2 — Translation + "noticing" (highlighting)

4. Paste the provided Italian text and run translation from Italian to English (IT → EN).
5. Identify highlighted segments and understand why they are highlighted and the color coding.
6. Click a highlighted segment to open the associated panel.

T3 — Interaction with AI-generated content

7. Read the AI-generated explanation, including the image representing the meaning (Language Insight panel)
8. Check whether content is clear, understandable, informative, and coherent

T4 — Interaction with expansions

9. Open/close any expansions
10. Check whether the affordances to open the expansions are consistent
11. Check whether the expansion content is clear, understandable, informative, and coherent
12. Evaluate anchoring to text spans, control, and error/ambiguity handling.

T5 — Closing the loop: Useful / Not useful

13. Use **Useful** and **Not useful** for at least one cue.
14. Verify the system state change is **visible and understandable** (e.g., history / save-for-later).
15. Retrieve the saved item (not yet applicable) and verify what was recorded.

4) What to check (Nielsen's 10 usability heuristics — correct terms)

During tasks, identify violations of Nielsen's heuristics and log issues:

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

5) Human–AI mini-checklist (derived from Amershi et al., CHI 2019)

A. Set appropriate expectations (G1–G3)

- G1 Make clear what the system can do (capability framing: cosa fa il sistema oltre alla traduzione)
- G2 Make clear how well the system can do what it can do (trust calibration: output AI fallibile, casi limite)
- G3 Time services based on context (cute/suggerimenti mostrati senza interrompere)

B. Support efficient, controllable interaction (G4, G7–G12)

- G4 Show contextually relevant information (ancoraggio allo span source/target; collegamento evidenziazione↔cue)
- G5 – "Match relevant social norms." (tono appropriato a contesto lavorativo/utente; evitare output "out of place")
- G6 – "Mitigate social biases." (evitare stereotipi/assunzioni indesiderate negli esempi/testo AI)

- G7 Support efficient invocation (aprire cue/espansioni in modo rapido e prevedibile)
- G8 Support efficient dismissal (ignorare/chiudere in modo efficiente: fondamentale per persona la velocità, essendo un'attività condotta durante il task di primario di traduzione)
- G9 Support efficient correction (se l'AI non è utile: modi chiari per riorientare o correggere)
- G10 Scope services to match users' needs (gestione ambiguità: disambiguazione o graceful degradation)
- G11 Make actions and outcomes understandable (perché un segmento è evidenziato / perché una cue è proposta)
- G12 Maintain short-term memory (history e riferimenti efficienti a ciò che è stato accettato, ancora non implementato)

C. Personalization without surprise (G13–G14)

- G13 Learn from user behavior (personalizzazione basata su Accept/Discard e profile, ancora non implementato)
- G14 Update and adapt cautiously (avoid abrupt, unexplained changes, ancora non implementato)

D. Handle uncertainty and failure safely (G15–G18)

- G15 Encourage granular feedback (Useful/not useful come feedback locale; eventualmente motivazione/flag)
- G16 Convey the consequences of user actions (e.g., what cosa comporta Useful vs Not useful sul futuro del Sistema changes)
- G17 Provide global controls (e.g., controlli globali: disable AI cues, adjust verbosity)
- G18 Notify users about changes (when behavior/capabilities change)

6) Persona impact tagging + light walkthrough

Persona cards

Each expert receives persona cards **P1/P2/P3** (derived from the survey). Use them as lenses for identifying breakdowns.

Persona impact tagging (required per issue)

For **every issue**, add:

- **Persona impacted:** Persona **P1 / P2 / P3** (multiple allowed) or **All**
- **(Optional) Rationale (one line):** e.g., “low proficiency → confusion interpreting highlights”, “goal: speed → friction in opening/closing cues”, “time pressure → ignores cue but can’t dismiss fast”

Light persona-based walkthrough (recommended)

After your first pass, do a quick second pass:

- For each persona, re-check **one core flow** (T2–T4) and note any **persona-specific breakdowns** (time pressure, low confidence, high efficiency needs, etc.).
  Keep it light: aim for **2–3 minutes per persona**.

EVALUATION FORM

A) Session info

- Expert: E1
- Expertise (HCI / UX / other): HCI, UX

B) Task outcomes (quick status)

Mark **OK / Issues / Blocking** and add short notes per any specific issue. Consider in particular possible issues with each Persona.

Severity score (0–4) (overall task-level severity)

0 = smooth, no issues

1 = minor friction, task completed easily

2 = noticeable friction/confusion, but task completed

3 = major problems, task completion difficult / error-prone

4 = task could not be completed / breakdown of trust or contro

- **T1 (Settings):** ☐ OK × Issues ☐ Blocking
  Severity (0–4): 1

  Notes (incl. Persona impact if relevant):

  Language settings are not fully clear. The interface suggests the presence of language direction controls through a visual indicator in the top-right area, implying that language direction may be changeable in a future version. However, in the current prototype the interaction appears to be intentionally scoped to writing in Italian and receiving a translation in English, with two predefined and non-editable text boxes.

  This design choice is not explicitly communicated to the user. As a result, users must infer that only the Italian text box is intended for input and that language direction cannot be changed at this stage. This creates ambiguity between perceived affordances and actual

system capabilities (violation of H1 – Visibility of system status and H3 – User control and freedom), and requires users to rely on inference rather than recognition (violation of H6 – Recognition rather than recall).

Persona impact:

P1: May experience uncertainty about where interaction is allowed, increasing hesitation during first use.

P2: Reduced clarity and transparency may negatively affect trust in the system's intentional scope and reliability.

- **T2 (Highlighting / noticing):** ☐ OK × Issues ☐ Blocking
  Severity (0–4): 2

  Notes (incl. Persona impact if relevant):

  Phrasal verbs in the English translation are visually highlighted using colored underlining and are clearly clickable, supporting initial noticing and interaction. The highlighting mechanism effectively draws attention to relevant linguistic elements and allows users to recognize interactive items without relying on prior knowledge (H1 – Visibility of system status, H6 – Recognition rather than recall).

  However, the criterion used to assign different colors is not made explicit. It is unclear why each phrasal verb is displayed with a different color and whether color variation conveys semantic, functional, or pedagogical meaning. As a result, users must infer that colors are used only to visually distinguish multiple phrasal verbs within the same text, introducing ambiguity and inconsistency (H2 – Match between system and the real world, H4 – Consistency and standards).

  From an accessibility perspective, the use of multiple colors without semantic explanation may negatively affect readability and increase cognitive load. A single high-contrast highlight color would better support legibility and accessibility, especially for users with visual impairments or low tolerance for visual complexity (H8 – Aesthetic and minimalist design). While color differentiation can be useful to identify the presence of multiple phrasal verbs, the approach may scale poorly: in cases where four or more phrasal verbs are present, the accumulation of different colors could lead to visual clutter and reduced clarity.

  Persona impact:

  P1: Higher risk of confusion and visual overload due to unexplained color variation.

P3: Likely to engage with highlighted items, but may initially misinterpret color differences as indicators of difficulty or importance.

- **T3 (AI-generated cues + expansions):** ☐ OK × Issues ☐ Blocking
  Severity (0–4): 1

  Notes (incl. Persona impact if relevant):

  The content is generally clear, relevant, and well aligned with the user's immediate task, supporting comprehension without requiring users to leave the translation workflow (H2 – Match between system and the real world, H6 – Recognition rather than recall).

  A particularly significant strength of the interaction is the possibility to access multiple alternative contents and to expand the available explanatory resources. Allowing users to explore additional examples, related phrasal verbs, and alternative formulations supports deeper understanding and accommodates different learning and usage needs (H7 – Flexibility and efficiency of use, G10 – Scope services to match users' needs).

  However, the system does not explicitly communicate the nature or reliability of the AI-generated content. There is no indication that the explanations are AI-produced, nor any signal of uncertainty or potential ambiguity in interpretation. This limits transparency and may lead users to over-trust the provided explanations, especially in professional or institutional contexts (G2 – Make clear how well the system can do what it can do, G11 – Make actions and outcomes understandable).

  In addition, the fixed order and cumulative presentation of information may increase cognitive load, particularly for first-time users or users under time pressure. While the richness of content is a strength, the lack of prioritization or progressive disclosure can reduce efficiency (H8 – Aesthetic and minimalist design).

  Persona impact:

  P2: Benefits from access to alternative formulations and contextual examples, but may still be concerned about the absence of reliability cues.

  P1: May find the amount of information overwhelming without clearer guidance on which content is essential versus optional.

- **T4 (Expansion + history/trace):** ☐ OK X Issues ☐ Blocking
  Severity (0–4): 1

  Notes (incl. Persona impact if relevant):

Expandable sections in the side panel allow users to access additional examples and related phrasal verbs. The interaction is coherent and predictable: expansions are clearly associated with the selected phrasal verb and can be opened and closed without losing control of the main task (H3 – User control and freedom, H4 – Consistency and standards, G7 – Support efficient invocation).

There is not enough information to assess how history or trace of previously explored items is managed; therefore, no conclusive evaluation can be made regarding short-term memory support at this stage.

A relevant design opportunity concerns personalization: allowing users to customize the order or prominence of information within the side panel could improve information retrieval and efficiency, especially for expert or high-demand users (H7 – Flexibility and efficiency of use). This is not a current usability issue but a potential enhancement.

Persona impact:

P2: Would particularly benefit from customizable information order to prioritize accuracy- and terminology-related content.

P1: Less affected, but could benefit from simplified or reordered views.

- **T5 (Useful/Not useful + history/trace):** ☐ OK X Issues ☐ Blocking
  Severity (0–4): 1

  Notes (incl. Persona impact if relevant):

  The interface provides Useful and Not useful controls at the bottom of the side panel, allowing users to give quick binary feedback on the AI-generated content. The interaction is simple and does not interrupt the primary task, supporting lightweight feedback during use (H3 – User control and freedom, G15 – Encourage granular feedback).

  However, the system does not communicate the consequences of providing feedback. After selecting either option, the side panel simply closes, with no visible confirmation, explanation, or trace of what has been recorded. This limits transparency and makes it unclear how feedback will influence future system behavior (H1 – Visibility of system status, G16 – Convey the consequences of user actions).

  Persona impact:

  P1: May be unsure whether feedback has been successfully registered.

  P2: Reduced trust due to lack of traceability and accountability of feedback.

P3: Missed opportunity for reuse and consistency in institutional contexts where traceability matters.

C) Nielsen heuristics checklist (tick if violated; add location)

X H1 Visibility of system status
X H2 Match between system and the real world
X H3 User control and freedom
X H4 Consistency and standards
☐ H5 Error prevention
X H6 Recognition rather than recall
X H7 Flexibility and efficiency of use
X H8 Aesthetic and minimalist design
X H9 Help users recognize, diagnose, and recover from errors
X H10 Help and documentation

Where / evidence :

H1: The interface suggests that language direction and user profiling features are available (e.g., top-right indicator, login button), but these functions are not active and not explicitly labeled as unavailable. In addition, after selecting Useful or Not useful, no visible confirmation or system state update is provided, making outcomes unclear.

H2: Highlight colors applied to phrasal verbs do not correspond to meaningful linguistic or functional distinctions. Users must infer that color variation has no semantic value, reducing alignment with real-world expectations.

H3: Visible but inactive controls (language direction, login/profile) create an expectation of user control that cannot be fulfilled. Feedback actions cannot be reviewed, undone, or traced, limiting users' sense of control over their interactions.

H4: Multiple highlight colors are used without consistent or explained semantics, potentially leading users to expect different behaviors or meanings associated with different colors.

H6: Users must rely on inference to understand which text area is editable, that the language pair is fixed, and why phrasal verbs are highlighted, due to the absence of explicit labels or guidance.

H7: AI-generated content and expandable sections follow a fixed structure and order, limiting efficiency for experienced users who may want quicker access to specific information.

H8: The combination of multiple highlight colors and dense side-panel content may cause visual clutter, particularly as the number of phrasal verbs increases. The side panel also reduces space for the main texts, with potential impact on smaller screens.

H9: After user actions such as submitting Useful or Not useful feedback, no feedback or explanatory message is shown, leaving users uncertain about whether the action was successfully completed.

H10: The interface provides no onboarding, legend, tooltip, or inline documentation explaining system scope, highlighting logic, or the role of AI-generated content.

D) Human–AI checklist (Amershi et al., CHI 2019) — for AI touchpoints only

For each item: **Yes / Partial / No / N.A.** + evidence.

Expectations & timing

- G1 Make clear what the system can do: ☐ Yes X Partial ☐ No ☐ N.A. Evidence: The system's core capability (highlighting and explaining phrasal verbs) is understandable through interaction, but it is not explicitly framed or introduced.
- G2 Make clear how well it can do it: ☐ Yes ☐ Partial X No ☐ N.A. Evidence: There is no indication of AI involvement, reliability, or potential limitations of the explanations provided.
- G3 Time services based on context: X Yes ☐ Partial ☐ No ☐ N.A. Evidence: AI-generated cues appear only upon user interaction and do not interrupt the primary translation task.

Control, transparency, efficiency

- G4 Show contextually relevant info: X Yes ☐ Partial ☐ No ☐ N.A. Evidence: Explanatory content is clearly anchored to the selected phrasal verb and relevant to the current text span.
- G7 Support efficient invocation: X Yes ☐ Partial ☐ No ☐ N.A. Evidence: Cues and expansions can be accessed through direct interaction with highlighted elements.
- G8 Support efficient dismissal: X Yes ☐ Partial ☐ No ☐ N.A. Evidence: The side panel can be easily closed, allowing users to ignore AI suggestions without friction.
- G9 Support efficient correction: ☐ Yes X Partial ☐ No ☐ N.A. Evidence: Users can dismiss content via Not useful, but cannot provide details.

- G10 Scope services to users' needs: ☐ Yes X Partial ☐ No ☐ N.A. Evidence:The system offers rich explanatory content, but does not adapt or prioritize information based on user goals or context
- G11 Make actions/outcomes understandable: ☐ Yes X Partial ☐ No ☐ N.A. Evidence: While the link between highlighted text and explanations is clear, the rationale behind feedback effects is not explained.
- G12 Maintain short-term memory: ☐ Yes ☐ Partial ☐ No X N.A. Evidence: _____

Adaptation

- G13 Learn from user behavior: ☐ Yes ☐ Partial ☐ No X N.A. Evidence: _____
- G14 Update and adapt cautiously: ☐ Yes ☐ Partial ☐ No X N.A. Evidence: _____

Failure & recourse

- G15 Encourage granular feedback: ☐ Yes X Partial ☐ No ☐ N.A. Evidence: Binary feedback (Useful / Not useful) is supported, but without qualitative input.
- G16 Convey consequences of user actions: ☐ Yes X Partial ☐ No ☐ N.A. Evidence: The system does not explain how user feedback influences future behavior. In other cases there is a certain grade of predictability.
- G17 Provide global controls: ☐ Yes X Partial ☐ No ☐ N.A. Evidence: Users can control the depth of interaction by choosing whether to expand additional explanatory content, but no global controls are available to adjust AI behavior (e.g., disable cues, reduce verbosity, or manage preferences).
- G18 Notify users about changes: ☐ Yes ☐ Partial ☐ No X N.A. Evidence: No system changes or updates are communicated in the prototype.

E) Persona impact tagging (required) + light walkthrough notes

Persona impact tagging rule

For every issue logged in section F, fill:

- Persona impacted: **A / B / C / D / All**
- Optional rationale: 1 line

Light persona-based walkthrough summary (2–3 lines per persona)

- P1: Most impacted by unclear system status and unexplained visual cues (language direction, highlight colors, feedback effects). Benefits from contextual explanations and examples, but may experience uncertainty and cognitive load when system behavior is not explicitly stated. Also, absence of history, feedback traceability, and explicit system framing limit usefulness in institutional contexts where consistency and accountability matter.
- P2: Least impacted by usability issues. Impacted by lack of transparency regarding AI-generated content, absence of reliability cues, and limited control over content prioritization.
- P3: Appreciates rich explanations and alternative examples: highlighting, expansions, and multiple examples align well with multiple options goals. However, trust and efficiency are reduced by missing traceability and personalization. Minor frictions in usability (fixed structure, lack of customization) do not significantly hinder task flow.