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. Check that Language source is set to ITA and target is ENG.
3. Check whether settings are clear and persistent as expected.

T2 — Translation + “noticing” (highlighting)

4. An Italian text and its translation are provided, simulating a translation activity.
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 and control.

T5 — Closing the loop: Useful / Not useful

13. At the moment **Useful** and **Not useful** in this version is not active yet, but imagine using it for at least one cue.

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: E2
- 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): X** OK □ Issues □ Blocking
  Severity (0–4): 0

  Notes (incl. Persona impact if relevant):

  The first use of the interface seems clear and user-friendly.

- **T2 (Highlighting / noticing):** □ OK **X** Issues □ Blocking
  Severity (0–4): 0

  Notes (incl. Persona impact if relevant):

  The highlighted sections of the text catch the eye

- **T3 (AI-generated cues + expansions): X** OK □ Issues □ Blocking
  Severity (0–4): 0

Notes (incl. Persona impact if relevant):

The content is clear and well explained

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

  Notes (incl. Persona impact if relevant):

  The content of expansions is consistent and clear. I find very useful the examples showed by clicking the images, but I would add an instruction like “would you like to have some more examples about the subject, click the images below/above”

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

  Notes (incl. Persona impact if relevant):

  Personally, I find the system and the Interface very useful, it seems clear, user friendly, and complete.

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

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

Where / evidence :

The unticked checkboxes refer to elements of the list that I couldn’t check during the evaluation. The ticked ones are referred to the details given in the right window. There are the images that help the person memorize the meanings of up used in phrasal verbs and the language used is simple, and comprehensible. Finally, the appearance of the interface is simple and very easy to understand and to use.

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: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: the overall view of the interface
- G2 Make clear how well it can do it: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: the translation and the details on the right window
- G3 Time services based on context: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: ________

Control, transparency, efficiency

- G4 Show contextually relevant info: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: the highlighted verbs in the text
- G7 Support efficient invocation: ☐ Yes ☐ Partial ☐ No **X** N.A. Evidence: ________
- G8 Support efficient dismissal: ☐ Yes ☐ Partial ☐ No **X** N.A. Evidence: ________
- G9 Support efficient correction: ☐ Yes ☐ Partial ☐ No **X** N.A. Evidence: ________
- G10 Scope services to users' needs: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: ________
- G11 Make actions/outcomes understandable: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: ____
- 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: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: the highlighted buttons "useful
- G16 Convey consequences of user actions: ☐ Yes ☐ Partial ☐ No **X** N.A. Evidence: ____
- G17 Provide global controls: ☐ Yes ☐ Partial ☐ No **X** N.A. Evidence: ________
- G18 Notify users about changes: ☐ Yes ☐ Partial ☐ No **X** N.A. Evidence: ________

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

Persona impact tagging rule

For every issue logged in section F, fill:

- Persona impacted: **P1 / P2 / P3 / All**
- Optional rationale: 1 line

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

- P1: (Anna Ferrando) this tool will be very useful for her, but since she doesn't trust MT probably it will be difficult for her to approach a tool like this. Surely the fact that it gives short and precise information about the language and its use, there's still a possibility for her to learn how to use it.
- P2: (Carlos Sanchez) Since he has a good level of digital literacy and works with GenAI and MT, it will be easy for him to discover and use a tool like this. It should be very useful, since his main goal is to learn English and how to use the language based on the context.
- P3: (Marta Soler) Since she doesn't trust MT translation, but she uses GenAI translation, this tool will be a great help for her, because it fulfill all of her needs. Plus, she could use it, not only to translate documents, but also to learn how to use the language, why there are differences in translation and how to deal with cultural aspects connected to the language and its use. The only limit I see, and I would see it with any AI tool, is the fact that she wants to make realistic and true content, and a lot of times this is a very big limit for this kind of tools, you have to train them properly before you can rely on them.