A) Session info

- **Expert:** (ChatGPT as evaluator)
- **Expertise:** HCI / UX + Human–AI interaction

B) Task outcomes (quick status)

T1 (Settings): Issues

Severity (0–4): 2

Notes (incl. Persona impact):

- The only visible "setting" is **Direction: IT → EN** in the header. There's **no obvious onboarding or settings affordance** (e.g., user profile/proficiency, highlight intensity, turning learning cues on/off).
- Persistence can't be verified from static screenshots, but the UI doesn't communicate what is "configured" vs "default."
- **Persona impact:**
    - **P1 (Anna) & P2 (Carlos):** need reliability and control in professional contexts; lack of explicit settings/control reduces confidence.
    - **P3 (Marta):** may want tone/verbosity controls for publishable output; none visible.

T2 (Highlighting / noticing): Issues

Severity (0–4): 3

Notes (incl. Persona impact):

- Highlights appear on both source and target (blue/orange), but there is **no legend or explanation** of:
    - what *highlighting means* (error? learning opportunity? idiom? low-confidence?)
    - what the *colors mean*
- This forces **guesswork**, especially on first use, and risks misunderstanding the purpose of the tool.
- **Persona impact:** All; strongest for **P1** and **P2** (high-stakes correctness), and **P3** (time pressure, wants quick clarity).

T3 (AI-generated cues + expansions): Issues

Severity (0–4): 2

Notes (incl. Persona impact):

- The **Language Insight** panel is readable and structured (selected text, short explanation, image, expansion buttons).
- However, it lacks:
  - **trust calibration** (how reliable is this explanation? when could it be wrong?)
  - **why this cue is shown** (why this segment is highlighted / why now)
  - **source/grounding** (especially important for professional users)
- **Persona impact:**
  - **P2 (Carlos):** wants standards/terminology reliability; needs citations or confidence indicators.
  - **P1 (Anna):** needs clear, dependable guidance for administrative texts.

T4 (Expansion + history/trace): Issues (Expansion = Issues; History/trace = N.A.)

Severity (0–4): 2

Notes (incl. Persona impact):

- Expansion affordances are mostly consistent (buttons like "GROW UP: more examples", "UP: similar uses", "UP: more uses").
- Issues:
  - The expansions **shift concept level** (from "grow up" to particle "UP") without clear framing; may confuse why "UP" is now the focus.
  - As you scroll, you can lose the **anchoring context** (selected span isn't persistently visible).
- **History/trace:** not implemented → **N.A.**
- **Persona impact:**
  - **P3 (Marta):** may like examples/variety but could get lost in long content; wants quick, skimmable relevance.
  - **P1/P2:** want efficiency; expansions should be easier to navigate/exit.

T5 (Useful/Not useful + history/trace): Issues (explicitly not active yet)

Severity (0–4): 3

Notes (incl. Persona impact):

- The UI shows a **"Useful" button** (and partial "Not useful" visible), but protocol states it's **not active yet**.
- This creates a **high risk of expectation break** if users click and nothing happens (or unclear state), which can undermine trust quickly.
- Also: even if it were active, there's no visible explanation of **what feedback does** (local vs global learning, future personalization, etc.).
- **Persona impact:** All; especially **P1/P2** (trust + accountability).

C) Nielsen heuristics checklist (violations + evidence)

I'm marking as "violated" where screenshots show clear risk.

- ☑ **H1 Visibility of system status**
  Evidence: no clear indicator of what highlights mean, and feedback controls appear without confirming effect/state.
- ☑ **H2 Match between system and the real world**
  Evidence: mixed labeling like "Language Insight / Selected text" alongside Italian explanatory text; can feel inconsistent for IT-native users.
- ☑ **H3 User control and freedom**
  Evidence: panel can be closed, but there's no obvious "quick dismiss" workflow for time-pressured users; also no controls to disable/adjust cues.
- ☑ **H4 Consistency and standards**
  Evidence: inconsistent language (Italian vs English UI labels); highlight color semantics not standardized/communicated.
- ☑ **H5 Error prevention**
  Evidence: users could misinterpret highlights as "errors to fix" or "recommended phrases" with no guardrails; feedback UI shown though inactive.
- ☑ **H6 Recognition rather than recall**
  Evidence: users must remember what colors mean (no legend); expansions can lose anchoring context while scrolling.
- ☑ **H7 Flexibility and efficiency of use**
  Evidence: no visible shortcuts or mode controls (e.g., "translation-only mode" vs "learning mode", verbosity).
- ☑ **H8 Aesthetic and minimalist design**
  Evidence: the right panel can become content-heavy (images + long lists), competing with the primary translation task.
- ☐ **H9 Help users recognize, diagnose, and recover from errors**
  (Not enough evidence of error states in screenshots, but "expectation breaks" are likely if clicking inactive feedback.)
- ☑ **H10 Help and documentation**
  Evidence: no onboarding/help affordance visible to explain highlighting, cue purpose, or how the AI should be used.

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

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

Expectations & timing

- **G1 Make clear what the system can do: Partial**
  Evidence: "LearnTranslate Beta" + Language Insight exists, but no explicit framing of features (why cues appear, what they cover).

- **G2 Make clear how well it can do it: No**
  Evidence: no confidence, limitations, or fallibility cues; no provenance.
- **G3 Time services based on context: Partial**
  Evidence: highlights are inline and non-modal (good), but they are always present and may distract; no "quiet mode."

Control, transparency, efficiency

- **G4 Show contextually relevant info: Partial**
  Evidence: panel shows "Selected text" and is linked to a highlight; but long expansions can drift from the exact span and context.
- **G7 Support efficient invocation: Yes**
  Evidence: clicking a highlighted segment appears to open the panel directly.
- **G8 Support efficient dismissal: Partial**
  Evidence: close "X" exists; however, panel occupies significant space and there's no compact/minimize or keyboard hint.
- **G9 Support efficient correction: No**
  Evidence: no mechanism visible to correct AI cue, request alternative explanation, or correct a translation directly from the cue.
- **G10 Scope services to users' needs: Partial**
  Evidence: content is helpful but "one-size-fits-all"; no visible tailoring controls (verbosity, proficiency, goal: speed vs learning).
- **G11 Make actions/outcomes understandable: Partial**
  Evidence: not clear **why** a segment is highlighted; outcomes of clicking are visible but rationale is missing.
- **G12 Maintain short-term memory: N.A. (not implemented)**
  Evidence: protocol explicitly notes history not implemented.

Adaptation

- **G13 Learn from user behavior: N.A. (not implemented)**
- **G14 Update and adapt cautiously: N.A. (not implemented)**

Failure & recourse

- **G15 Encourage granular feedback: N.A. / Partial risk**
  Evidence: feedback buttons are visible, but protocol says not active. If inactive, it's N.A.; if shown, it risks misleading users.
- **G16 Convey consequences of user actions: N.A. (not active)**
  Evidence: no explanation of what Useful/Not useful would do.
- **G17 Provide global controls: No**
  Evidence: no global toggle to disable cues, adjust frequency, or set learning vs translation mode.
- **G18 Notify users about changes: N.A.**
  Evidence: no change events shown in screenshots.

E) Persona impact tagging + light walkthrough notes

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

P1 (Anna Ferrando — administrative clerk, clarity/reliability, time pressure):
Highlights could help spot risky phrases quickly, but the lack of a clear legend and trust cues makes the AI feel unsafe in high-accountability work. Needs "translation-first" flow with fast dismiss and clearer guarantees/limits.

P2 (Carlos Sanchez — technical PM, standards/terminology accuracy):
Would use expansions/examples, but needs provenance (sources, standards references) and correction controls. Without trust calibration and the ability to steer/correct, the AI guidance may be ignored or distrusted.

P3 (Marta Soler — marketing/content, quality/appropriateness):
Expansions (examples/similar uses) are valuable for style variation, but the panel can become heavy and distract from primary editing. Needs scannable relevance cues, and controls for tone/verbosity and "publishable" style.