Language Scent: Exploring Cross-Language Information Navigation

Jiawen Stefanie Zhu, Katharina Reinecke, Tanushree Mitra
University of Washington
Seattle, Washington, USA
jiawenz2@uw.edu,reinecke@cs.washington.edu,tmitra@uw.edu

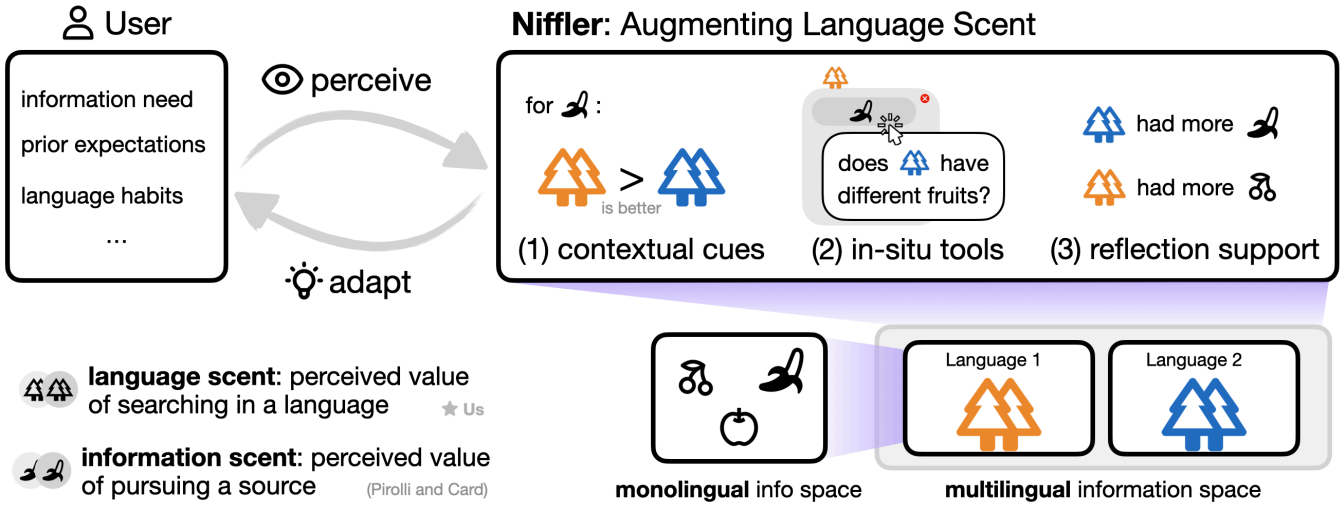

Figure 1: Overview of language scent and how Niffler augments it. To support understanding, we draw a parallel between the process of seeking information and that of scavenging for food. The information space can be viewed as a group of forests, each representing a monolingual subspace, with individual pieces of information represented as fruits. Seeking information is analogous to searching for fruit, where users follow scent cues in the environment, i.e. system signals, to locate desired information. In monolingual scenarios, users remain within a single forest, whereas in multilingual scenarios, users must also navigate between forests, moving across different languages' information spaces. Information scent, as defined by Pirolli and Card [47], captures the perceived utility of individual sources (fruits). Our notion of language scent, in contrast, focuses on meta-level strategy formation when navigating across different languages (forests). Niffler surfaces previously obscured cross-language information through (1) contextual cues about the utility of searching in different languages, (2) in-situ tools that facilitate quick overviews of information in other languages, and (3) reflection support that promote refinement of one's cross-language search strategies, thereby augmenting language scent.

Abstract

While multilingual users often switch between languages when seeking information, this process remains undersupported by current systems where information is typically siloed by language. Our formative study reveals that users' cross-language transitions are guided by their perceived value of switching to a language, a concept we formalize as **language scent**. Language scent extends Pirolli and Card's theory of information scent to multilingual scenarios by considering meta-level strategy formation when navigating between different languages. To support language scent, we designed Niffler, a search system that augments language scent and supports cross-language information navigation through contextual cues, in-situ tools, and reflection support. A lab study with 16 multilingual speakers showed that Niffler facilitated the formation and execution of exploratory and granular search strategies and leads to diverse information being gathered. Our findings establish language scent as a valuable lens on cross-language information seeking, highlighting language's role in enabling access to broader information and offering concrete implications for the design of multilingual search systems.

1 Introduction

Information sources across languages often contain distinct and non-overlapping content; consulting multiple languages can thus surface complementary perspectives and lead to a more comprehensive understanding [9, 31, 48]. Multilingual speakers therefore often draw on multiple languages to meet their information needs [8, 25, 39, 54]. However, current search systems tend to silo information by language, for example by showing only sources in the query language, making this broader information space difficult to navigate. Overall, seeking information in multiple languages and transitioning between them remains largely under-supported.

,

While a large body of prior work has examined multilingual speakers' information seeking behaviours—including their language preferences across topics [8, 25, 60], how they navigate different multilingual search result interfaces [52, 54], and how they (re)formulate queries when switching language [23, 24]—these studies tend to operate under the implicit assumption that users *inherently know* which language is best to use for a given search task. However, closer examination suggests that this assumption may not always hold. Empirical findings frequently highlight trial-and-error processes, wherein users recognize the need to switch to another language only after extensive, unproductive searches in their initial language [8, 65]. Despite this, the mechanisms by which users determine appropriate search languages and the challenges they face in doing so remain poorly understood. To bridge this gap, we investigate the following research questions:

RQ1 How do multilingual users strategize and navigate between languages during information seeking?
RQ2 How do we design search systems that facilitate this cross-language strategy formation and refinement process?
RQ3 How do such designs affect users' information-seeking processes and outcomes?

We focused on English-Chinese users as a case study, following Human-Computer Interaction (HCI) conventions of studying a specific language pair to be able to capture nuances [8, 10, 17, 64] (details in Section 2.4).

We began by conducting a formative study (**RQ1**) with 10 English-Chinese speakers to understand their information seeking strategies and pain points, especially with regard to cross-language information navigation. We found that although participants recognized the advantages of searching in multiple languages at a conceptual level, in practice they often avoided switching away from their initial search language unless it felt necessary because of the associated operational costs. They also noted that while they had intuitions about which language(s) would work best in different situations, these intuitions were not always reliable, at times leading them down unproductive paths. In particular, we found that participants selected their search language based on perceptions about the value of each language for fulfilling their information need. This perceived utility was influenced by several factors, including the informational value of sources in the language, alignment with the participant's language proficiency for the topic, and the ease of searching in or switching to that language within the system. We formalize this concept as **language scent**, or the *perceived value of using a particular language* for search, inspired and informed by the concept of information scent from information foraging theory [47]. While information scent focuses on choosing between individual information sources and implicitly assumes a monolingual context, language scent focuses on meta-level strategies for navigating between different languages, a unique need that arises in multilingual scenarios (Figure 1).

This newly defined mechanism of language scent helps explain why users often remain in a single language longer than is productive in current systems. For one, the perceived utility of searching in other languages (i.e., language scent) is low by default due to operational barriers, such as having to manually repeat searches to access information in different languages. For the other, the lack of system cues about what information is available in each language forces users to purely rely on their imperfect intuitions, further suppressing language scent. As a result, users are unable to form accurate views about which language to use during information seeking. To address this problem, we derived design guidelines from formative study results to develop Niffler (**RQ2**), a multilingual search system designed to support cross-language information navigation by augmenting language scent. Niffler consolidates and juxtaposes content from both languages to highlight the utility of searching in each. It provides information and linguistic cues at multiple levels to support cross-language awareness and encourage consideration of both languages throughout the information-seeking process. Niffler also includes in-situ tools for quick verification of hypotheses or intuitions about the most effective search language, minimizing context switching. Reflection support helps users consider and refine their mental models of the multilingual information space, potentially extending beyond the current session. Overall, Niffler augments language scent through contextual cues, in-situ tools and reflection support.

We examined how Niffler influences users' cross-language information seeking experience in a lab study with 16 English-Chinese speakers. Results showed that Niffler helped users develop more flexible and granular search strategies, enabling them to gather a more diverse set of information. Overall, our findings suggest that language scent is a valuable lens on cross-language information seeking and indicate how the concept can inform the design of multilingual search systems.

2 Related Work

2.1 Grounding in Information Foraging Theory

Information foraging theory is a fundamental theory of information seeking, positing that people navigate information spaces by following sources they perceive to be the most valuable for their needs, based on proximal cues [47]. This perception of the value of a source is called their information scent [47]. Information scent, in its original formulation [47], assumes a monolingual context and focuses on navigating between individual patches of information within a single information space. However, existing models do not account for the multilingual context, where users can navigate multiple information spaces. This introduces an additional decision layer in the information-seeking process: users must not only navigate within a given information space, but also decide when to switch and trade off between different information spaces. We study this additional dimension introduced by multilingual information seeking. In our formative study, we observed that in addition to the process of choosing between sources of information, as described by information scent, users also undergo a process of deciding when to use which language, which we named language scent. Language scent extends the theory of information scent by focusing on meta-level transitions between different language spaces, a dimension that becomes relevant as multiple languages are introduced.

Existing systems work has explored ways of enhancing users' (general) information scent. One line of work focuses on designing proximal cues that amplify the signal of individual sources, for example through enhanced thumbnails images [57, 63]. A recent line

of work surfaces more distant information patches by suggesting search queries as proximal cues based on users' past interactions [44, 45]. However, these were designed with monolingual contexts in mind and did not consider the additional layer of navigating between multiple information spaces introduced by multilingual information seeking. Niffler addresses this gap by designing for language scent to support the meta-level decision making process of switching between languages.

2.2 Understanding Multilingual Information Seeking

Information seeking is the conscious effort of acquiring information to fill a need or gap in one's knowledge [15], and is an important activity in daily life [7]. Existing work have investigated how multilingual users in particular seek information, and showed that they leverage different languages during this process [8, 25, 61]. One line of work investigates patterns used by people when employing certain languages. For example, users switch between their country-of-residence language and their native language during crisis information seeking to balance digestibility and authenticity [25]. Another line of work focuses on specific challenges and strategies that people encounter when switching languages, for example query (re)formulation [6, 23, 24]. However, none of the existing work has examined the process *before* users decide to use a certain language or switch languages, i.e. their search strategy formation stage. Rather, there is an implicit assumption that multilingual users always know when to use which language. This is not necessarily true, since existing work suggests that multilingual users often rely on trial and error, sticking with one language until they realize it is unable to satisfy their information needs [8]. Our formative study fills this gap by examining how multilingual users form and refine their information-seeking strategies, through the new lens of language scent.

2.3 Supporting Multilingual Information Seeking

While there is an abundance of work on underlying cross-lingual retrieval algorithms [32, 46], there is limited work on user-facing multilingual information seeking tools and systems. One such line of work primarily focuses on supporting query reformulation [23, 54], for example by automatically adapting imperfect user queries into more effective versions [56]. Another line examines how to design search result page UI to organize results from multiple languages effectively [19, 36, 53]. Neither supports the strategic navigation across languages. Our system, Niffler, fills this gap by designing around language scent, providing a tool that enables users to more consciously and meaningfully leverage information from multiple languages.

There are general information seeking tools that are related to some of our high-level design goals, like reducing information overload and facilitating search strategy formation. For example, systems like DiscipLink [67] and Selenite [37] consolidate and organize raw information, while CoNotate [44] and InterWeave [45] suggest relevant queries given user contexts. However, our contribution does not lie in general-purpose consolidation or suggestion mechanisms, but in manifesting the design concept of surfacing language scent through Niffler to better assist multilingual users.

2.4 Characterizing Multilingual Users

Multilingualism is an overloaded term with many definitions and interpretations [21, 38, 39]. Even in HCI alone, multilingualism carries two distinct connotations, one emphasizing the multi-competence of knowing multiple languages, e.g. [10], and the other emphasizing not being a native speaker of English, e.g. [34]. In this work, we define multilingual users as people who are fluent, i.e. can produce "complete meaningful utterances" [39], in two or more languages, and focus on the multi-competence aspect.

Furthermore, studying multilingual users as a whole is rare, given the diversity within this population. Rather, HCI research conventionally focuses on specific language pairs as case studies, enabling a more nuanced understanding [8, 10, 17, 64]. We focus on English and Chinese in this project, since they rank as the top two most spoken languages globally [4] and are spoken by some members of the research team [25].

3 Formative Study

We started with a formative study to investigate how multilingual users strategize and navigate between languages during information seeking.

3.1 Method: Formative Study

The study was conducted as an online interview study with 10 participants (**P#**, 7 women, 3 men; mean age = 25 ± 3), recruited through social media and snowball sampling. All participants were native speakers of Chinese and at least independent users of English according to the Common European Framework of Reference for Languages (CEFR) [3]. Specifically, one participant was at the B-level (independent users), nine were at the C-level (proficient/near-native users). All reported regularly seeking information online, on average about once a day (mean = 0.97 ± 0.09 times).

As a warm-up exercise, and to observe users' multilingual information seeking behaviours, we selected two tasks likely to induce language switching, based on prior work [53]: exploring public opinions on (1) the release of the AI model DeepSeek and (2) the practice of vegetarianism. Participants had 10 minutes for each task and were asked to think aloud [1, 16] during the process. We then concluded with a semi-structured interview on their general multilingual searching behaviours, challenges, and needs, which lasted about 30 - 40 minutes. In total, each study session took approximately 60 minutes and participants were remunerated US $15. This study was approved by our institution's ethics review board.

The task sessions and interview were screen- and audio-recorded. We conducted a thematic analysis by open-coding the interview transcripts [12].

3.2 Findings

From the formative study, we identified the mechanisms of language scent and the challenges of multilingual information seeking.

3.2.1 Language Scent. Our formative study revealed that multilingual users often leverage language as a heuristic to structure and facilitate their information seeking process. For example, while they may use English to *"[learn] a concept for school"* -P8, or due to the availability of a *"wide variety of sources"* -P3, they might decide to search in *"Chinese for creating travel plans for more practical tips"* -P9. Across participants, these choices were guided by expectations about the kinds of information each language tends to provide, which were refined through experience with those languages. We formalize this notion as **language scent**: the *perceived value of using a particular language during information seeking*. This extends the concept of information scent from information foraging theory [47] to a multilingual context. Language scent guides users' meta-level strategy for choosing between languages and works in conjunction with their information scent, which helps them navigate between individual sources once a language is selected (Figure 1).

3.2.2 Factors influencing Language Scent. Study results suggest that language scent is shaped by epistemic, interpretative, and practical factors. Additionally, current systems obscure language scent, leading users to face various challenges (**C#**) when seeking information across languages.

Epistemic Factors. We identified aspects of language scent related to the informational content available in a given language, including its availability, quality, and framing. Participants viewed language as a proxy for *"see[ing] other perspectives"* -P1, explaining that *"even if [they] ask the same question"* -P10, *"most likely the things [they] get from searching in Chinese and the things [they] get from searching in English are different"* -P10. Users also considered the applicability of information to their own context and positionality. For example, P4 explains that to *"know about the general visa application process"* -P4, it doesn't matter which language they use, but only *"the Chinese side is going to tell you to make your resume a little less sensitive to make sure it doesn't raise any eyebrows"* -P4.

Participants began with a prior mental model of the value of searching in each language to guide their search actions, refining this model *"if the acquired information is different from what [they] thought before"* -P1. However, based on their prior expectations alone, participants *"often not knowing which language is better at first"* -P3, and even if they did, found it difficult to predict *"if [they]'re getting the kind of information [they] want by searching in [a particular] language"* -P3. This difficulty is compounded by current systems, which silo information by language and provide few cues to help users develop more accurate intuitions about what each language contains (**C1 — Limited Multilingual Information Awareness**). A side effect is that users are often unaware of their language-related priors, only realizing by chance, after substantial trial and error, that *"over time this [a particular] language isn't helpful for [their] goal"* -P3 (**C2 — Difficulty Reflecting on Cross-Language Utility**).

Interpretative Factors. These are aspects of language scent related to the (perceived) ease of processing or digesting information using a particular language, shaped by the user's current language environment, past experiences, and language skills. Participants explained that the language they choose during information seeking is also *"about [their] thinking process"* -P9 and frequently *"not really a rational categorization"* -P2. Users would often *"just use the language that comes to mind first"* -P9, *"recall where [they] first encountered this problem and then habitually rely on that path"* -P2, or default to the language of greatest proficiency, rather than deliberately considering which language would best satisfy their information need.

The most cognitively natural path is not always the most effective for information seeking. Participants frequently encountered dead-ends when habitual strategies overshadowed their actual information needs (**C3 — Overreliance on Cognitive Shortcuts**). For instance, P1 would always *"first ask in Chinese [t heir native language] and see what kind of answers [they] get"*, since it was easiest for her to read and write. This strategy, however, often failed to satisfy her information needs, requiring additional effort to switch to another language such as English. The tacit *"conversion process"* -P10 of mapping concepts across languages can be mentally burdensome, hindering idea connection and discouraging switching to the most informative language when users' initial opportunistic language choice is insufficient.

Practical Factors. Our study also identified aspects of language scent related to the costs and effort required for cross-language interaction within the system or infrastructure. Currently, participants are often deterred from searching in multiple languages because *"information from different languages are captured in different silos"* -P9, making transitions between them expensive (**C4 - Inadequate Cross-Language Integration**). Part of the problem is that there is currently no *"unified entry point to search for information"* -P10 across languages. Users therefore need to *"manually add a middle step in between"* -P1 where they *"translate and redo [their search] to build up for seeking information in the other silos"* -P9, which can be tedious and time-consuming. This finding echoes prior work on query reformulation [6, 23, 24].

Even after successfully obtaining information in their desired language(s), participants found *"processing them [the information] tiring"* -P6 because *"there's too much information [...] and a lot of it is redundant and useless"* -P2 (**C5 – Information Overload**). The challenge of *"consolidating the information"* -P5 went beyond simple summarization and extended to the process of triangulation. Participants particularly desired *"seeing the similarities and differences between sources in different languages"* -P7 in order to *"make it clear the stance and positionality of each"* -P1. This requires extensive back-and-forth conversion and cross-referencing across languages, adding significant effort and compounding the sense of overload.

3.3 Design Guidelines

Overall, the formative study showed that during cross-language information seeking, participants had to rely solely on their imperfect mental models of information across languages due to the lack of system support. We derive four **d**esign **g**uidelines (**DG#**) to address the identified challenges:

- **[DG1]** Provide resources that support evidence-based awareness of information across different languages. (**C1**)
- **[DG2]** Support refinement of users' mental models regarding the utility of different languages for search. (**C2**)

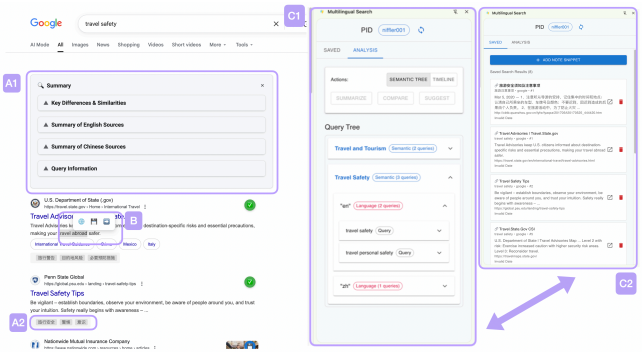

Figure 2: System Overview. NIFFLER consists of [A] a search results page enhancement of [A1] Comparative Summary and Query Information, and [A2] Cross-Lingual Keywords, [B] in-situ tools, and [C] a side panel with tabs [C1] analysis and [C2] saved content.

Figure 3: On the search results page, the [A] Comparative Summary consists of [A1] Cross-Lingual Comparison summarizing key similarities and differences between the two languages, and [A2] Summaries of English sources and Chinese sources, respectively. [B] Query Information is provided as background information to complement it. Each search result is decorated with [C] Cross-Lingual keywords.

[DG3] Expose users to multiple languages to reduce overreliance on cognitive shortcuts and activate thinking in both languages. (**C3**)

[DG4] Fulfill these guidelines while minimizing unnecessary effort and operational overhead for users. (**C4, C5**)

4 NIFFLER: Augmenting Language Scent

Guided by our design goals, we developed NIFFLER, a multilingual search system that augments language scent through contextual cues (**DG1**), reflection support (**DG2**), and in-situ tools (**DG3**). **DG4**, which focuses on minimizing practical barriers, informs the design of all features and is therefore integrated into the other goals rather than presented as a separate section.

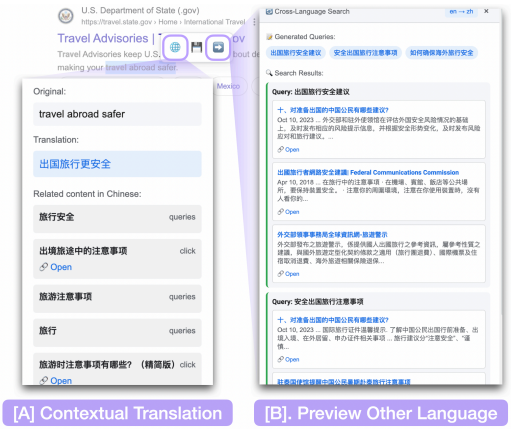

Figure 4: The tooltip opens when text is selected. Users can choose to see [A] Contextual Translation of the text, connecting to relevant user search history, or [B] Preview of Other Language, with suggested queries and sources in the other languages.

4.1 System Overview

NIFFLER comprises three main components (Fig. 2): (1) a search results page augmentation for forming search strategies (Fig. 2A1, A2), (2) tooltips for connecting across languages (Fig. 2B), and (3) an analysis side panel for reflecting on search strategies (Fig. 2[C1]) and note-taking (Fig. 2[C2]). We use the user's queries, clicks, saved content, and notes as a proxy of their search activity in the backend [33]. In particular, we treat queries as the smallest natural unit of search activity and organize clicks, saved content, and notes around them. Users can save queries and their corresponding search results or webpage snippets when using NIFFLER, or create custom **notes** in the side panel.

We implemented NIFFLER as a Google Chrome extension using the Chrome Extension API. The front-end was developed in TypeScript with React and Material UI, while the back-end was built in Python with FastAPI for handling API calls. Data were stored in Firebase and indexed and searched with TypeSense. We used the Google Search API to retrieve relevant sources and OpenAI's GPT-4o API for summarization and query generation.

4.2 Perceiving the Informational Value of Languages (DG1)

NIFFLER helps users gauge the informational value of languages, thereby surfacing language scent.

4.2.1 Always-On Overview. For every search, users can view a **Comparative Summary** (Figure 3A) of English and Chinese sources, helping to reduce information overload (**DG4**). It includes a Cross-Lingual Comparison, which highlights similarities and differences between the two languages and provides suggested queries to further explore the points of comparison (Figure 3[A1]). Summaries of Sources in each language is also provided, summarizing the key points with linked sources (Figure 3[A2]). To obtain information from both languages, we followed a pipeline of translating and

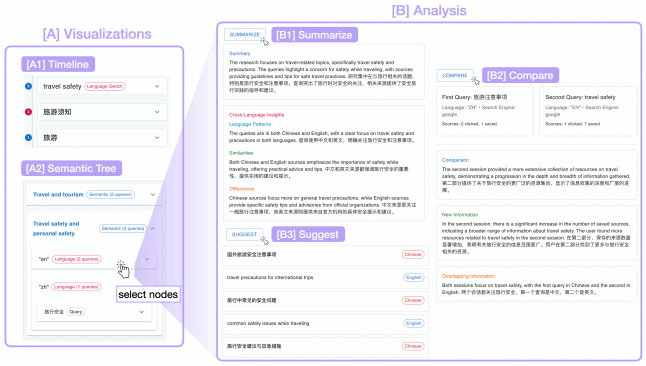

Figure 5: The side panel analysis tab provides [A] language-centred visualizations of user search activity and [B] analysis functions to help them consolidate information and evaluate past strategies. There are two visualizations, the [A1] Timeline highlighting language switching patterns and the [A2] Semantic Tree focusing on high-level concepts. Users can select nodes on the visualization for analysis, being able to [B1] Summarize, [B2] Compare, or [B3] Suggest queries.

rewriting the query in the other language [32, 51], retrieving relevant sources from search engine result pages [44, 45], clustering [59], and summarizing [18], following common information retrieval approaches [68]. The exact queries used are shown in the **Query Information** section for context (Figure 3B).

4.2.2 In-Situ Tool. Information seeking is a serendipitous process. To support this, the **Preview Other Language** function (Figure 4B) lets users view content from another language without leaving the current context by selecting text and clicking the tooltip button (Figure 4B). It provides suggested queries and relevant sources in the other language, allowing users to assess whether switching languages is worthwhile while cross-referencing with the current language, without needing to context-switch and do a full new search (**DG4**).

4.3 Reflecting on the Informational Value of Languages (DG2)

The side panel (Figure 5) helps users understand and refine their mental model of multilingual information spaces, by allowing them to view and analyze their search activity, organized by language use.

4.3.1 Visualizations. We introduced two language-centred visualizations of users' search activity (Figure 5A). The **Semantic Tree** (Figure 5[A2]) organizes past searches first by subject matter and then by language, allowing users to see their language use for different topics. The **Timeline** (Figure 5[A1]) presents searches chronologically, highlighting points of language switching. In both representations, users can click on query nodes to expand them and view the sources and notes related to it.

4.3.2 Analysis Functions. We also provide three complementary analysis functions (Figure 5B) to help users interpret their search history and the information gathered with less overhead (**DG4**). The **Summarize** function synthesizes the content of selected nodes, offering an overview and cross-language comparison of sources (Figure 5[B1]). The **Compare** function dives one step deeper by showing the marginal benefit of a later-selected query relative to an earlier one by identifying new versus overlapping information (Figure 5[B2]), for example allowing users to better evaluate the value of switching languages or remaining in a language. The **Suggest** function facilitates further exploration and expansion by recommending additional queries to extend the selected nodes in both languages (Figure 5[B3]).

4.4 Connecting Ideas across Languages (DG3)

To reduce over-reliance on cognitive shortcuts rooted in language proficiency and preferences, and to activate and connect thinking across languages, Niffler provides features to convert between the languages; as well as cues to prime users' latent knowledge in the other language and nudge them to connect them.

4.4.1 Always-On Cues. The comparative summary and search activity analysis functions are displayed in both languages, so that language is not a barrier. For search results, translating each of them entirely could create overload and cognitive stress (**DG4**), and is not space efficient. Instead, **Cross-lingual Keywords** (Figure 3C) summarize the content of individual sources in the other language.

4.4.2 In-Situ Tool. To address ad-hoc needs to connect languages, the **Contextual Translation** function (Figure 4A) translates selected text into the other language and shows relevant items from the user's search activity in the other language, i.e. queries, clicks, saved content, notes. The relevant items are retrieved using both query translation [32] and embedding-based [58] approaches. This helps users connect knowledge and intermediate search results across their two languages without having to manually sift through past records (**DG4**).

5 Lab Study of Niffler

We conducted a lab study of Niffler to understand how its features augmenting language scent influence users' multilingual information seeking.

5.1 Method: Niffler Lab Study

We followed a within-subjects design with two conditions (Niffler and Baseline) and two tasks. The conditions and tasks were fully Latin-square balanced to mitigate potential order effects.

5.1.1 Conditions. The Baseline condition (Figure 6) consisted of a parallel search interface and an AI chat panel, representing an enhanced version of status quo tools [41]. We did not directly compare against existing tools, as they operate in only one language, which would make the baseline inherently less information-dense than Niffler.

While no commercially available tool currently supports multilingual search, prior work has explored interfaces that display results in multiple languages [19, 52]. From this work, we adopted the panel design for our Baseline condition that was shown to be most preferred in Chu and Komlodi [19]. The AI chat panel was implemented using OpenAI's API to ensure consistency and avoid

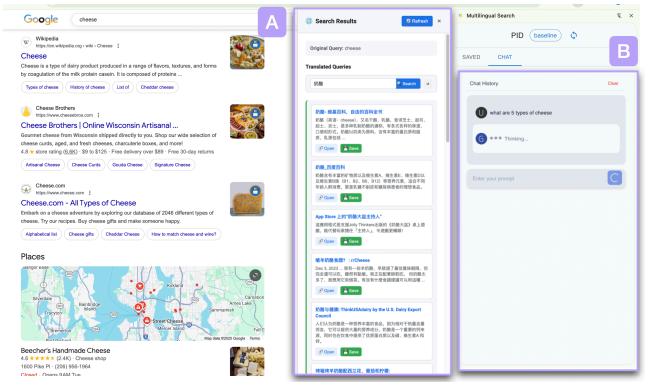

Figure 6: The Baseline consists of [A] Parallel Search Panel and [B] AI Chat Panel.

variability from participants using different models. To enable a fair comparison, Baseline also includes the same saving and notes functionalities as Niffler.

5.1.2 Task. We created two exploratory information seeking tasks (informed by [14]) that were open-ended, likely to occur in real-world settings, broadly applicable, and designed to minimize bias toward English or Chinese contexts. Participants were asked to collect diverse information on two topics (Appendix C.1): (1) Career Advice and (2) Food and Restaurant recommendation in Switzerland (chosen because its official languages do not include English or Chinese [2]). Participants were given 20 min for each task, based on prior research on the average duration for conducting exploratory online information seeking [7, 14]. They were also encouraged to take notes on the fly to stay engaged with the task.

5.1.3 Data Collection and Analysis. We collected Likert-scale ratings on ten questions, assessing participants' perception of Niffler's ease of use, and their experience of forming and reflecting on language scent (Appendix A). We also collected system logs and screen recordings from the sessions. For participants' navigational behaviour during information seeking, logs of user queries were used as a proxy [29, 40]. The measures we examined are: *number of queries*, number of *language switches*, number of consecutive queries with each language (*language span*), and distribution of queries across languages (*language balance*) through Shannon's entropy [35, 50] (details in Appendix C.2.1). For a proxy of the relevant information participants gathered, we used logs of the sources participants clicked [33], or saved or took notes on [67]. Two coders independently coded the topic coverage of these sources for each participant–task, blind to condition, following the iterative procedure in [49] (details in Appendix C.2.2). Finally, for qualitative feedback, we open-coded the interview transcripts to gain a systematic and structured understanding [12] of how users perceived and interacted with Niffler.

5.1.4 Participants. We recruited 16 participants (14 women, 2 men; mean age = 25 ± 2.18) via social media and snowball sampling. The sample size was determined based on an a-priori power analysis ($\alpha = 0.05$) for detecting a medium (Cohen's $d = 0.75$) effect size [20, 43]. All participants were native speakers of Chinese and at least independent users of English according to the Common European Framework of Reference for Languages (CEFR) [3]. All participants regularly engaged in online information seeking (mean = 2.76 ± 0.66 sessions a day). Participants self-reported frequently using both English (mean = 6.25±1.06) and Chinese (mean = 6.50 ± 1.03) for information seeking, on a 7-point Likert scale.

5.1.5 Procedure. We conducted the study remotely via zoom, starting by obtaining participant consent. Before proceeding to the tasks, participants were asked to install a Chrome extension containing both Niffler and Baseline. Before each task, we explained the system to participants and provided time for them to familiarize themselves with it as needed. Each task session lasted 20 minutes, and participants were encouraged to think aloud [1, 16]. After each task, participants completed a questionnaire with Likert-scale items assessing their experience. The study lasted approximately 90 minutes, and participants were compensated US $25. Task sessions and interviews were audio- and screen-recorded, then transcribed verbatim. Their interactions with the system were logged. The study protocol was approved by our institution's ethics review board.

5.2 Lab Study Findings

We report our qualitative (from participant interviews) and quantitative (from the questionnaire and system log data) findings below. For quantitative data, we applied the Wilcoxon signed-rank test to calculate statistical significance, as it is a non-parametric method that makes no assumptions about the underlying distribution [13]. We applied the Benjamini-Hochberg correction to the Likert-scale questionnaire to control for family-wise false discovery rate. Full statistics can be found in Table 1 for Likert-scale items and Table 2 for system logs (Appendix B). Here, we summarize the key results, reporting the mean (M), p-values, and effect size (r). For a technical evaluation for how Niffler performed in the study tasks, see Appendix D.

5.2.1 Contextual cues augment language scent. From the Likert-scale items, it was significantly easier to identify similarities and differences across languages in Niffler compared to in Baseline ($M_{\text{Niffler}} = 1.88 < 3.56 = M_{\text{Baseline}}$; $p = 0.044, r = 0.829$; lower is better). No significant difference was found in the other two Likert items on forming language scent (Appendix A.3).

Participants had amplified awareness of cross-language differences in Niffler:

> *"I feel Niffler is a bit beyond my expectations — it really exceeded what I imagined. I didn't expect that there would be such a big difference between the Chinese and English web."* -P5

The surprising insights tend to *"provide inspiration"* -P2 and *"trigger [participant's] interest in checking more"* -P1, even if they *"didn't think to search in this [a particular] language at first"* -P14. This motivated them *"to think about a problem from more diverse perspectives and to gather information from more diverse angles"* -P3. While Baseline had some nudging effect as well, participants *"didn't pay attention or notice as much"* -P2 because *"it doesn't clearly distinguish between results and [...] doesn't help understand general trends"* -P9 across languages.

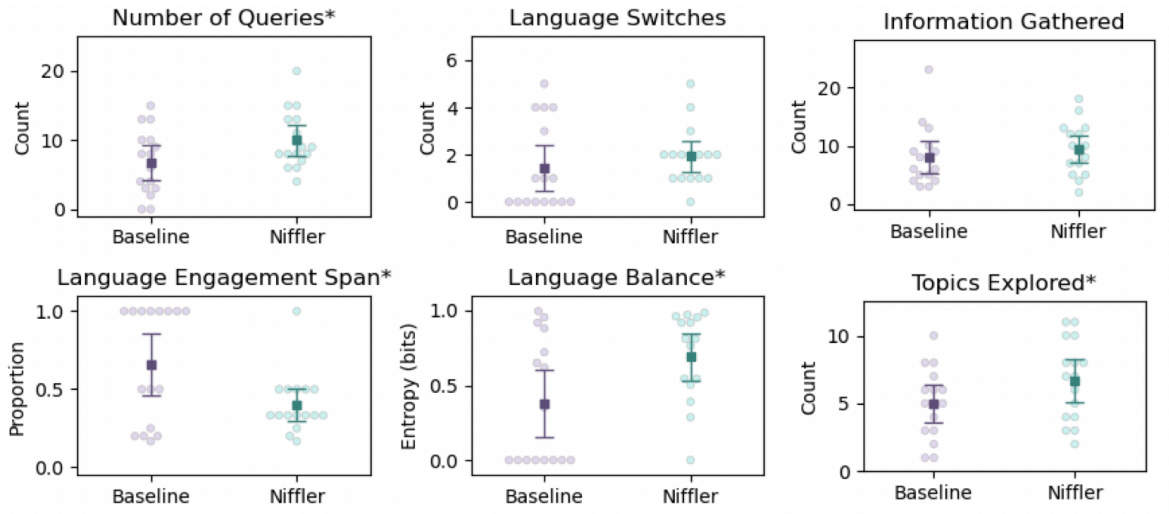

Figure 7: Swarm Plots for Analysis on Information Seeking Process and Outcomes. Square mark indicates mean and error bars are 95% confidence intervals. Asterisk (*) indicates statistical significance. Process: In Niffler, participants searched significantly more, stayed in the same language for significantly shorter duration, and the language distribution of their queries was significantly more balanced, compared to Baseline. There was no significant difference in the number of times they switched the query language. Outcome: There was no significant difference in the amount of information gathered, but participants explored significantly more topics with Niffler compared to Baseline.

On the **epistemic** side, participants drew on the similarities and differences across languages in the *Comparative Summary* to guide their search strategies, and to extract useful information for getting started. Similarities were perceived as *"more common and widely recognized, rather than culturally specific"* -P13, which participants either trusted more readily or treated as indicators of importance to investigate further, depending on their goals. Differences, on the other hand, were used to *"identify variations in cultural emphases"* -P12 and broaden perspectives. Participants scrutinized differences more carefully than similarities to assess their underlying causes and relevance before deciding whether to accept the information, particularly when the differences were *"unexpected or didn't align with [their] intuition"* -P4.

Niffler's design of showing summaries and content in both languages also helped amplify the interpretative factor of language scent by *"allowing [participants] to better understand the content"* -P8 through cross-referencing. In particular, the *cross-lingual keywords* were valuable since *"having tags in a different language felt like having two distinct pieces of information [...] if the tags were in the same language, it probably won't help as much"* -P2. *Cross-lingual keywords* also had a subtle nudging effect *"connecting to [participants'] knowledge and helps [them] gauge if checking what the other language says is better"* -P3. These features help participants recognize their preferences for working with specific topics or information through contextual cues.

5.2.2 Augmented language scent facilitates formation of search strategies. The system log of user queries (Figure 7) showed that users formulated significantly more queries ($M_{\text{Niffler}} = 10.0 > 6.75 = M_{\text{Baseline}}; p = 0.041, r = 0.549$) when using Niffler. Niffler's workflow was also more dynamic than Baseline, helping users to adjust their strategies flexibly and promptly in response to evolving information needs. There was no significant difference in the average number of times participants switched the query language across the two conditions. However, after switching, they stayed within the same language for significantly less consecutive queries than Baseline (normalized; $M_{\text{Niffler}} = 0.403 < 0.657 = M_{\text{Baseline}}$; $p = 0.017, r = 0.642$), i.e. switched more quickly. These align with participant observations that Niffler enabled them to form *"more fine-grained and effective strategies"* -P6, whereas in Baseline they tended to *"stick to one or completely switch languages mid-way"* -P6, in a more opportunistic manner. Perhaps a consequence of this, participants issued queries across the two languages significantly more evenly in Niffler than Baseline (Shannon entropy [35, 50]; $M_{\text{Niffler}} = 0.690 > 0.381 = M_{\text{Baseline}}; p = 0.020, r = 0.601$).

Participants thought that Niffler led to a more exploratory and flexible workflow compared to Baseline, where participants *"branched out"* -P11 more and *"digged deeper into details"* -P16 when using Niffler. Participants found whereas they had to primarily *"rely on [their] own intuition"* -P6 in Baseline, Niffler allowed them to be more *"targeted"* -P6 and *"intentional"* -P16 when switching languages, providing *"a clear idea and some inspiration for next steps"* -P4. For example, P9 described how Niffler help them efficiently form a search strategy when researching career advice:

> *"When searching for employment goals, the system [Niffler] helped me realize the results in Chinese and English were different. The Chinese results were mostly related to promotions or government statements, encouraging people to apply for specific jobs. For me, that kind of information wasn't very useful, so I didn't want to spend time searching in Chinese. Instead, I mainly*

> *looked at the English results, which helped me avoid spending extra time on repetitive searches."* -P9

Niffler also supported participants in flexibly adapting and refining their strategies. Continuing the previous example, Niffler later helped P9 realize that unlike for employment goals, for information on *"time management, [they] actually prefer Chinese because for English it's more websites from specific universities, with timelines targeted towards their own students, whereas in Chinese it's more general"* -P9.

5.2.3 Augmented language scent facilitates reflection on search strategies. While we did not find significance for the two Likert-scale items on facilitating reflection on language scent (Appendix A.3), this may be due to the nature and duration of the task. Participants reported that while *"the system [Niffler] supports reflection ... and [they] have some thoughts during the process"* -P7, they *"feel like [they] haven't reached the stage of reflecting yet"* -P7, and therefore many did not engage substantially with the reflection features. Participants did note that Niffler encouraged meta-level reflection on their language scent beyond the current session. In addition to evaluating their language scent through the search outcomes, participants were most interested about whether their engagement with different topics was skewed toward a single language. Participants mentioned that Niffler *"helped [them] realize [them] always subconsciously choosing a specific language, something that [they] might not otherwise have noticed"* -P6, for example using the conceptual groupings in the *Semantic Tree* to understand language distribution across topics. Participants did not regard uneven use of the two languages as intrinsically positive or negative. Instead, it prompted them to reflect on cross-language differences in context of the different topics, and, when deemed substantial, to *"adapt [their] strategies not just now but also in the future"* -P6.

5.2.4 Augmented language scent leads to more diverse information gathered. In terms of the information seeking outcomes (Figure 7), there was no significant difference between the number of relevant sources participants found across Baseline ($M_{\text{Baseline}} = 8.13$) and Niffler ($M_{\text{Niffler}} = 9.44$). However, with Niffler, participants were able to explore significantly more topics ($M_{\text{Niffler}} = 6.69 > 5.00 = M_{\text{Baseline}}; p = 0.032, r = 0.597$), compared to Baseline. Similarly, in their interviews, participants also found that searching in multiple languages *"gives [them] interesting perspectives"* -P16 and *"reveals things [they] wouldn't be aware of otherwise"* -P3. Since participants were able to gain more (diverse) information from accessing similar amounts of sources, this may triangulate previous findings that Niffler may lead to the formation of more effective strategies which allow participants to better leverage the multilingual information available to them.

5.2.5 Usability of Niffler. For the five self-reported Likert items on overall impression and ease of use (Appendix A.1-2), no significant differences were observed; these smaller-than-expected effects may partly reflect Niffler's *"steeper learning curve"* -P10. Indeed, system logs show that only 8 of 16 participants ever switched languages in Baseline, compared to 15 in Niffler, potentially suggesting that operational barriers of switching languages were lower in Niffler, even if participants did not perceive this subjectively. In interviews, participants noted that once they became familiar with the features, Niffler's *"tool assistance encouraged [them] to do multilingual information seeking"* -P6. Compared to their status quo and the Baseline workflow, participants described a *"trade-off between time and effort versus the result"* -P1, suggesting that Niffler lowered the effort required to explore multiple languages. Participants liked that *"it [Niffler] offers high-level insights"* -P1 such as *"the key differences between Chinese and English sources"* -P8, and does not *"require a lot of effort to compare or process the information by [oneself]"* -P1.

5.2.6 Envisioned Real-Life Use. Our findings indicate that Niffler may be the most helpful for exploratory tasks, where the goal is to *"hear as many voices and opinions as possible"* -P5, and less useful for transactional tasks, where a single clear answer can typically be obtained through one search. Exploration may be desirable *"when [they] care a lot about the truthfulness of the information"* -P4 (high-stakes) or when *"they are learning about a completely new topic"* -P3 (limited prior knowledge). There was no clear pattern in the topics where Niffler was expected to be helpful. Scenarios mentioned range from everyday tasks like *"buying cars"* -P5 and *"trip planning"* -P1, to more serious ones like *"medical advice"* -P10 and broader literature exposure in *"research"* -P1. Participants envisioned Niffler as an always-on support as *"it doesn't hurt to have more information, or someone comparing information from different sources for you"* -P1, especially since it's *"hard to predict when there are differences across languages"* -P13 and language scent *"evolves over time, adapting to [their] search outcomes"* -P6. Additionally, participants explained that they could easily disregard irrelevant information when it is not applicable.

6 Discussion

6.1 Mental Models of Multilingual Information

Language scent is shaped by users' prior mental models of multilingual information and by the environmental cues they encounter when interacting with the system. Niffler surfaces previously obscured language scent through explicit system signals, facilitating the refinement of users' priors and enabling us to observe patterns that were previously hidden due to the difficulty of detecting language scent. From the lab study results, we found that users displayed three main mental models regarding the role of information from other languages, which we explain below. Note that the user's mental model is not static but evolves based on their information seeking experience.

Multilingual Information as Fallback. In this view, the baseline information scent is low and users treat information in another language as a fallback, switching only when they are unable to find satisfactory results in their primary language. For example, P2 represents an extreme case, as they did not switch languages at all during the task, explaining that *"if Chinese [the other language] returns significantly better results than English [the primary language] [...] but [they] don't expect it to be the case"* -P2. Their existing mental model and low baseline language scent lead them to largely ignore the contextual cues surfacing language scent, leaving their information seeking was driven almost entirely by information scent. Only 3 out of 16 participants had this mental model (P2, P3, P11) at

some point, with P3 shifting towards the multilingual-information-as-safeguard view after using Niffler.

Multilingual Information as Safeguard. In this view, the baseline language scent is medium and users treat information in another language as a way to cross-reference and verify information. Although users also adopt a primary search language in this case, they are more willing to switch languages than in the multilingual-information-as-fallback case. When the accuracy or impartiality of information is especially important, participants with this mental model would *"check if the languages agree and if [they] missed any perspectives"* -P16. Otherwise, for easier tasks (e.g., transactional queries) or when they care less, they tend to remain in their primary language, mainly guided by their information scent. 10 out of 16 participants displayed this mental model (P3, P4, P5, P6, P7, P10, P12, P13, P14, P15), with three of them (P5, P6, P12) shifting towards the multilingual-information-as-complementary-resource model after interacting with Niffler.

Multilingual Information as Complementary Resource. In this view, the baseline language scent is high and users treat information in different languages as complementary resources that jointly form a complete understanding. Participants with this mental model switched languages freely and frequently, treating both languages as (mostly) equally valuable resources. Their information seeking drew on both language scent and information scent, with language scent playing an active and central role rather than remaining secondary. 7 out of 16 participants displayed this mental model (P1, P5, P6, P8, P9, P12, P16), with four them adopting this view after using Niffler.

6.2 Language as a Heuristic for Information Seeking

Our lab study showed that the language scent support in Niffler encouraged participants to explore more broadly and consider multiple perspectives by nudging them toward varied sources in different languages. This suggests that language itself can serve as a heuristic for accessing diverse information, which is beneficial in general, helping with forming informed opinions [11] and overcoming single-language filter bubbles [48]. As a heuristic, language has the advantage of being an inherent property of any piece of information, making it scalable and requiring less contextual learning than current heuristics based on platforms or media outlets [62]. While our work focuses on navigation, future work could more directly investigate how language can be leveraged as a mechanism for promoting exposure to diverse information. This may include studying how information from different languages can be used without requiring linguistic expertise.

6.3 Studying Multilingual Users in HCI

Multilingual users have long been studied in HCI, with most prior work focusing on issues of limited language competence and technological asymmetry across languages [17, 34]. In contrast, our work highlights a complementary and still underexplored approach: focusing on users' multi-competence. We show that multilingual users can leverage their proficiency across languages to achieve more effective information seeking, giving rise to new patterns of interaction and distinct cognitive processes. In this sense, multilingual users are not simply monolingual users replicated across multiple languages, but instead exhibit qualitatively different ways of engaging with information.

6.4 Limitations and Future Work

Our work has several limitations and opportunities for future work. First, we recruited people who are fluent in English and Chinese to investigate the broader multilingual population. While this is in line with conventional HCI practices [10, 17, 25] and certain aspects of multilingualism are considered universal regardless of the exact languages spoken [5, 26–28], our findings may not generalize to all language pairs. Future work can examine *language scent* and its support in other language contexts to validate and refine our insights.

Second, the backend of Niffler currently retrieves information from the web using the Google Search API. However, other localized search engines may work better for specific languages (e.g. Baidu for Chinese) and participants additionally noted that they sometimes also use social media platforms that may not be indexed by Google. While our focus is on the interaction design of language scent support, future work could explore incorporating such additional information sources to further optimize the backend. Niffler is also designed for two languages, and it may not support users who speak three or more languages. Future work could investigate how to design for multilingual users with more than two languages, for example examining whether they prefer to see information from all their languages simultaneously or selectively leverage specific subsets for different purposes.

Lastly, we only evaluated Niffler in a lab environment with two 20-minutes tasks. Although these tasks were designed to be broadly applicable, the limited task type and duration means they inevitably capture only a limited range of search behaviours. Future work could extend on our findings by deploying Niffler over a longer period of time to understand its utility in an organic setting and investigate long-term behavioural changes. Additionally, in this paper, we chose a baseline that surfaces information from multiple languages but does not explicitly design for language scent, allowing us to investigate which interface designs best support multilingual information seeking. For future work, monolingual and/or localized baselines (e.g., Baidu for Chinese) could further isolate the effect of access to multiple languages itself.

7 Conclusion

In this paper, we introduced the concept of language scent to describe how multilingual users assess the value of switching languages during information seeking. Our formative study with Chinese–English bilinguals revealed that users follow their language scent to access alternative perspectives, cross-validate information, and locate more relevant content. At the same time, the study highlighted challenges of multilingual search, including uncertainty about when to switch languages, barriers in transitioning between languages, and the cognitive burden of managing and triangulating information across linguistic contexts. To address these issues, we proposed design guidelines for supporting language scent and

instantiated them in Niffler. Results from a lab study with 16 multilingual participants demonstrated that Niffler increased awareness of information available across languages, promoted more frequent multilingual searching, and encouraged more systematic cross-lingual strategies. Together, these contributions establish language scent as a useful lens for understanding multilingual information seeking and suggest concrete directions for designing search systems that better support multilingual users.

References

[1] 2012. Thinking Aloud: The #1 Usability Tool. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/.

[2] 2024. Sprachen. https://www.aboutswitzerland.eda.admin.ch/de/sprachen.

[3] 2025. The CEFR Levels - Common European Framework of Reference for Languages (CEFR) - Www.Coe.Int. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.

[4] 2025. What Are the Top 200 Most Spoken Languages? https://www.ethnologue.com/insights/ethnologue200/.

[5] Christian Adjemian. 1976. On the Nature of Interlanguage Systems. Language Learning 26, 2 (1976), 297–320. doi:10.1111/j.1467-1770.1976.tb00279.x

[6] Frans Albarillo. 2018. Information Code-Switching: A Study of Language Preferences in Academic Libraries. College & Research Libraries 79, 5 (July 2018), 624. doi:10.5860/crl.79.5.624

[7] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is Exploratory Search Different? A Comparison of Information Search Behavior for Exploratory and Lookup Tasks. Journal of the Association for Information Science and Technology 67, 11 (2016), 2635–2651. doi:10.1002/asi.23617

[8] Anne Aula and Melanie Kellar. 2009. Multilingual Search Strategies. In CHI '09 Extended Abstracts on Human Factors in Computing Systems (New York, NY, USA, 2009-04-04) (CHI EA '09). Association for Computing Machinery, 3865–3870. doi:10.1145/1520340.1520585

[9] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: bridging the wikipedia language gap. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1075–1084.

[10] Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do Multilingual Users Prefer Chat-bots That Code-mix? Let's Nudge and Find Out! Proceedings of the ACM on Human-Computer Interaction 4, CSCW1 (May 2020), 1–23. doi:10.1145/3392846

[11] Md Momen Bhuiyan, Sang Won Lee, Nitesh Goyal, and Tanushree Mitra. 2023. NewsComp: Facilitating Diverse News Reading through Comparative Annotation. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581244

[12] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. Qualitative Research in Psychology 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[13] Patrick D Bridge and Shlomo S Sawilowsky. 1999. Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-Test and Wilcoxon Rank-Sum Test in Small Samples Applied Research. Journal of Clinical Epidemiology 52, 3 (March 1999), 229–235. doi:10.1016/S0895-4356(98)00168-1

[14] Zeljko Carevic, Maria Lusky, Wilko van Hoek, and Philipp Mayr. 2018. Investigating Exploratory Search Activities Based on the Stratagem Level in Digital Libraries. International Journal on Digital Libraries 19, 2-3 (Sept. 2018), 231–251. arXiv:1706.06410 [cs] doi:10.1007/s00799-017-0226-6

[15] Donald O. Case and Lisa M. Given. 2016. Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior. Emerald Group Publishing.

[16] Elizabeth Charters. 2003. The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. Brock Education Journal 12, 2 (July 2003). doi:10.26522/brocked.v12i2.38

[17] Yunjae J. Choi, Minha Lee, and Sangsu Lee. 2023. Toward a Multilingual Conversational Agent: Challenges and Expectations of Code-mixing Multilingual Users. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581445

[18] Janara Christensen, Stephen Soderland, Gagan Bansal, and Mausam. 2014. Hierarchical Summarization: Scaling Up Multi-Document Summarization. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland, 902–912. doi:10.3115/v1/P14-1085

[19] Peng Chu and Anita Komlodi. 2017. TranSearch: A Multilingual Search User Interface Accommodating User Interaction and Preference. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17). Association for Computing Machinery, New York, NY, USA, 2466–2472. doi:10.1145/3027063.3053262

[20] Jacob Cohen. 1992. Statistical Power Analysis. Current Directions in Psychological Science 1, 3 (June 1992), 98–101. doi:10.1111/1467-8721.ep10768783

[21] Florian Coulmas. 2018. An Introduction to Multilingualism: Language in a Changing World. Oxford University Press.

[22] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Nikolaos Aletras and Orphee De Clercq (Eds.). Association for Computational Linguistics, St. Julians, Malta, 150–158. doi:10.18653/v1/2024.eacl-demo.16

[23] Hengyi Fu. 2017. Query Reformulation Patterns of Mixed Language Queries in Different Search Intents. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17). Association for Computing Machinery, New York, NY, USA, 249–252. doi:10.1145/3020165.3022126

[24] Hengyi Fu. 2018. Mixed Language Queries in Online Searches: A Study of Intra-Sentential Code-Switching from a Qualitative Perspective. Aslib Journal of Information Management 71, 1 (Oct. 2018), 72–89. doi:10.1108/AJIM-04-2018-0091

[25] Ge Gao, Jian Zheng, Eun Kyoung Choe, and Naomi Yamashita. 2022. Taking a Language Detour: How International Migrants Speaking a Minority Language Seek COVID-Related Information in Their Host Countries. Proc. ACM Hum.-Comput. Interact. 6, CSCW2 (Nov. 2022), 542:1–542:32. doi:10.1145/3555600

[26] Kieran Green. 2023. Identification of Commonalities across Different Languages. Frontiers in Language Sciences 2 (Nov. 2023). doi:10.3389/flang.2023.1172925

[27] Francois Grosjean. 1997. The Bilingual Individual. Interpreting 2, 1-2 (Jan. 1997), 163–187. doi:10.1075/intp.2.1-2.07gro

[28] François Grosjean. 2012. Bilingual and Monolingual Language Modes. In The Encyclopedia of Applied Linguistics (1 ed.), Carol A. Chapelle (Ed.). Wiley. doi:10.1002/9781405198431.wbeal0090

[29] Jacek Gwizdka and Ian Spence. 2006. What Can Searching Behavior Tell Us About the Difficulty of Information Tasks? A Study of Web Navigation. Proceedings of the American Society for Information Science and Technology 43, 1 (2006), 1–22. doi:10.1002/meet.14504301167

[30] Sean N. Halpin. 2024. Inter-Coder Agreement in Qualitative Coding: Considerations for Its Use. American Journal of Qualitative Research 8, 3 (July 2024), 23–43. doi:10.29333/ajqr/14887

[31] Brent Hecht and Darren Gergle. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In Proceedings of the SIGCHI conference on human factors in computing systems. 291–300.

[32] David A. Hull and Gregory Grefenstette. 1996. Querying across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '96. ACM Press, Zurich, Switzerland, 49–57. doi:10.1145/243199.243212

[33] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately Interpreting Clickthrough Data as Implicit Feedback. ACM SIGIR Forum 51, 1 (Aug. 2017), 4–11. doi:10.1145/3130332.3130334

[34] Bo Young Kim, Qingyan Ma, and Lisa Diamond. 2024. "It's in My Language": A Case Study on Multilingual mHealth Application for Immigrant Populations With Limited English Proficiency. In Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24). Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3613905.3637125

[35] J. Lin. 1991. Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory 37, 1 (Jan. 1991), 145–151. doi:10.1109/18.61115

[36] Chenjun Ling, Ben Steichen, and Alexander G. Choulos. 2018. A Comparative User Study of Interactive Multilingual Search Interfaces. In Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18). Association for Computing Machinery, New York, NY, USA, 211–220. doi:10.1145/3176349.3176383

[37] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–26. doi:10.1145/3613904.3642149

[38] William F. Mackey. 1962. The Description of Bilingualism. Canadian Journal of Linguistics/Revue canadienne de linguistique 7, 2 (April 1962), 51–85. doi:10.1017/S0008413100019393

[39] Parviz Maftoon and Masoume Shakibafar. 2011. Who Is a Bilingual? Journal of English studies 1, 2 (2011), 79–85.

[40] Mazlita Mat-Hassan and Mark Levene. 2005. Associating Search and Navigation Behavior through Log Analysis. Journal of the American Society for Information Science and Technology 56, 9 (2005), 913–934. doi:10.1002/asi.20185

[41] Kerstin Mayerhofer, Rob Capra, and David Elsweiler. 2025. Blending Queries and Conversations: Understanding Trust, Verification, and System Choice in Search and Chat Interactions. In Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval. ACM, Melbourne Australia, 168–178. doi:10.1145/3698204.3716454

[42] Cliodhna O'Connor and Helene Joffe. 2020. Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. International Journal of Qualitative Methods 19 (Jan. 2020), 1609406919899220. doi:10.1177/1609406919899220

[43] Anna-Marie Ortloff, Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas, and Matthew Smith. 2025. Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. ACM, Yokohama Japan, 1–28. doi:10.1145/3706598.3713671

[44] Srishti Palani, Zijian Ding, Austin Nguyen, Andrew Chuang, Stephen MacNeil, and Steven P. Dow. 2021. CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3411764.3445618

[45] Srishti Palani, Yingyi Zhou, Sheldon Zhu, and Steven P. Dow. 2022. InterWeave: Presenting Search Suggestions in Context Scaffolds Information Search and Synthesis. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. ACM, Bend OR USA, 1–16. doi:10.1145/3526113.3545696

[46] Carol Peters, Martin Braschler, and Paul Clough. 2012. Multilingual Information Retrieval: From Research To Practice. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-23008-0

[47] Peter Pirolli and Stuart Card. 1999. Information Foraging. Psychological Review 106, 4 (Oct. 1999), 643–675. doi:10.1037/0033-295x.106.4.643

[48] Dorian Quelle, Calvin Cheng, Alexandre Bovet, and Scott A. Hale. 2023. Lost in Translation – Multilingual Misinformation and Its Evolution. arXiv:2310.18089 [cs] doi:10.48550/arXiv.2310.18089

[49] K. Andrew R. Richards and Michael A. Hemphill. 2018. A Practical Guide to Collaborative Qualitative Data Analysis. Journal of Teaching in Physical Education 37, 2 (April 2018), 225–231. doi:10.1123/jtpe.2017-0084

[50] C. E. Shannon. 1948. A Mathematical Theory of Communication. The Bell System Technical Journal 27, 3 (July 1948), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

[51] Páraic Sheridan and Jean Paul Ballerini. 1996. Experiments in Multilingual Information Retrieval Using the SPIDER System. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '96. ACM Press, Zurich, Switzerland, 58–65. doi:10.1145/243199.243213

[52] Ben Steichen and Luanne Freund. 2015. Supporting the Modern Polyglot: A Comparison of Multilingual Search Interfaces. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). Association for Computing Machinery, New York, NY, USA, 3483–3492. doi:10.1145/2702123.2702541

[53] Ben Steichen, Chenjun Ling, and Silvia Figueira. 2023. Multilingual News Search—A Comparative User Study of Desktop and Mobile Interfaces. International Journal of Human–Computer Interaction 0, 0 (2023), 1–16. doi:10.1080/10447318.2023.2238978

[54] Ben Steichen and Ryan Lowe. 2021. How Do Multilingual Users Search? An Investigation of Query and Result List Language Choices. 72, 6 (2021), 759–776. doi:10.1002/asi.24443

[55] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. ACM, San Francisco CA USA, 1–18. doi:10.1145/3586183.3606756

[56] Zhongkai Sun, Zhengyang Zhao, Sixing Lu, Chengyuan Ma, Xiaohu Liu, Xing Fan, Wei Shen, and Chenlei Guo. 2023. CL-QR: Cross-Lingual Enhanced Query Reformulation for Multi-lingual Conversational AI Agents. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, Mingxuan Wang and Imed Zitouni (Eds.). Association for Computational Linguistics, Singapore, 423–431. doi:10.18653/v1/2023.emnlp-industry.40

[57] Meirav Taieb-Maimon. 2025. Enhancing Snippet Visualizations to Improve Web Search. International Journal of Human–Computer Interaction (Jan. 2025), 1–20. doi:10.1080/10447318.2024.2443267

[58] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15). Association for Computing Machinery, New York, NY, USA, 363–372. doi:10.1145/2766462.2767752

[59] Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, and Yihong Gong. 2011. Integrating Document Clustering and Multidocument Summarization. ACM Trans. Knowl. Discov. Data 5, 3 (Aug. 2011), 14:1–14:26. doi:10.1145/1993077.1993078

[60] Jieyu Wang and Anita Komlodi. 2018. Switching Languages in Online Searching: A Qualitative Study of Web Users' Code-Switching Search Behaviors. In Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - CHIIR '18. ACM Press, New Brunswick, NJ, USA, 201–210. doi:10.1145/3176349.3176396

[61] Jieyu Wang, Anita Komlodi, and Omar Ka. 2018. Understanding Multilingual Web Users' Code-Switching Behaviors in Online Searching. Proceedings of the Association for Information Science and Technology 55, 1 (2018), 534–543. doi:10.1002/pra2.2018.14505501058

[62] Jenny S. Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J. Watts. 2025. Media Bias Detector: Designing and Implementing a Tool for Real-Time Selection and Framing Bias Analysis in News Coverage. arXiv:2502.06009 [cs] doi:10.1145/3706598.3713716

[63] Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrsion, and Peter Pirolli. 2001. Using Thumbnails to Search the Web. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, Seattle Washington USA, 198–205. doi:10.1145/365024.365098

[64] Yimin Xiao, Cartor Hancock, Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, Marine Carpuat, and Ge Gao. 2025. Sustaining Human Agency, Attending to Its Cost: An Investigation into Generative AI Design for Non-Native Speakers' Language Use. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3706598.3713626

[65] Shutian Zhang and Dan Wu. 2024. How to Bridge the Gap? Information Asymmetry in Tibetan-Chinese Bilingual Search Behavior. Library & Information Science Research 46, 4 (Oct. 2024), 101329. doi:10.1016/j.lisr.2024.101329

[66] Yao Zhang and Chang Liu. 2020. Users' Knowledge Use and Change during Information Searching Process: A Perspective of Vocabulary Usage. 47–56. doi:10.1145/3383583.3398532

[67] Chengbo Zheng, Yuanhao Zhang, Zeyu Huang, Chuhan Shi, Minrui Xu, and Xiaojuan Ma. 2024. DiscipLink: Unfolding Interdisciplinary Information Seeking Process via Human-AI Co-Exploration. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24). Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3654777.3676366

[68] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2024. Large Language Models for Information Retrieval: A Survey. arXiv:2308.07107 [cs] doi:10.48550/arXiv.2308.07107

A Likert Items in the Lab Study

A.1 Overall Impression

- The multilingual aspect of this system did not provide any additional benefit compared to searching in a single language.

A.2 Ease of Use

- I could search in multiple languages efficiently.
- I had to deal with too much information.
- Switching between languages during the information seeking process felt smooth.
- It was difficult to connect my intermediary search results, prior knowledge, and thoughts across different languages.

A.3 Forming Language Scent

- The similarities and differences in information across languages were difficult to identify.
- It was easy to gauge the kinds of information available in each language.
- It was clear when searching in another language would be useful.

A.4 Reflecting on Language Scent

- This system helped me evaluate the effectiveness of my multilingual information seeking strategies.
- This system supports reflection on my intuitions about the availability of information across languages.

B Full statistical tests in Lab Study

B.1 Likert-scale Questionnaire

See Table 1 and Figure 8.

B.2 System Logs

See Table 2.

C Lab Study Method Details

C.1 Task Description

1. **Career.** You are teaching a career coaching course for university students who are about to graduate. In preparation, you want to gather a diverse set of career advice. Consider things like time management, goal setting, planning, choosing a career path, etc.
2. **Food.** You are travelling to Switzerland for a month with a group of friends from different nationalities. You want to find as many foods and restaurants to try as possible, considering local specialties, the diverse tastes and dietary habits of your group, etc.

C.2 Measures

C.2.1 Information Seeking Process. We analyzed logs of user queries as a proxy for participants' navigational behaviour during information seeking [29, 40]. Our focus was on high-level patterns of information seeking and language switching, rather than low-level details. Specifically, we examined the following metrics:

- **Number of Queries**: The total number of queries conducted by a user during the session.
- **Language Switches**: The total number of times a user changed the query language during the session.
- **Language Engagement Span**: The average number of consecutive queries a participant conducted in the same language (n_i for segment i), normalized by the total number of queries (n) in the session, i.e. $\frac{n_i}{n}$ for segment i. The greater the span, the longer users stayed in the same language without switching, on average.
- **Language Balance**: Shannon entropy [35, 50] was used to quantify the balance of queries across languages. Higher values indicate a more even distribution of queries across languages, while lower values indicate skewness towards one language.

C.2.2 Information Seeking Outcome. We analyzed logs of user queries as a proxy for participants' navigational behaviour during information seeking [29, 40]. For a proxy of the relevant information participants gathered, we used the sources participants clicked [33], or saved or took notes on [67]. In alignment with existing work, we used the number of topics as a measure of the diversity, or range of information covered [55, 66]. The topics were derived through systematic coding by two researchers, following the procedure in [49]. Each source was assigned the topic that most comprehensively describes its contents. All coding was done blind to condition. The coding rules were established by open coding and discussing 30% of the data. As a pilot test, the two researchers independently coded 20% of the data, achieving an initial percentage agreement [30] of 69.6%. We judged percentage agreement based on whether the clustering aligned. For example, if both coders grouped the same three items together, we counted it as an agreement even if the cluster

Table 1: Baseline and Niffler Likert-scale statistics, based on Wilcoxon signed-rank test [13]. We applied the Benjamini-Hochberg correction to control for family-wise false discovery rate. We report the mean (M), adjusted and original p-values, and effect size (r). * and bolding indicates statistical significance. For italicized items, the smaller the better, for the rest, the greater the better.

Likert Item	M_{Baseline}	M_{Niffler}	Adjusted p	Original p	r
The multilingual aspect of this system did not provide any additional benefit compared to searching in a single language.	2.75	1.94	0.303	0.061	0.583
I could search in multiple languages efficiently.	5.62	6.06	0.319	0.159	0.528
I had to deal with too much information.	3.69	3.56	0.860	0.860	0.049
Switching between languages during the information seeking process felt smooth.	5.38	6.19	0.319	0.150	0.526
It was difficult to connect my intermediary search results, prior knowledge, and thoughts across different languages.	3.75	2.88	0.319	0.140	0.432
****The similarities and differences in information across languages were difficult to identify.***	3.56	1.88	0.044	0.004	0.829
It was easy to gauge the kinds of information available in each language.	5.31	5.56	0.634	0.571	0.180
It was clear when searching in another language would be useful.	5.06	5.38	0.634	0.569	0.216
This system helped me evaluate the effectiveness of my multilingual information seeking strategies.	4.81	5.38	0.459	0.303	0.362
This system supports reflection on my intuitions about the availability of information across languages.	4.88	5.50	0.459	0.321	0.313

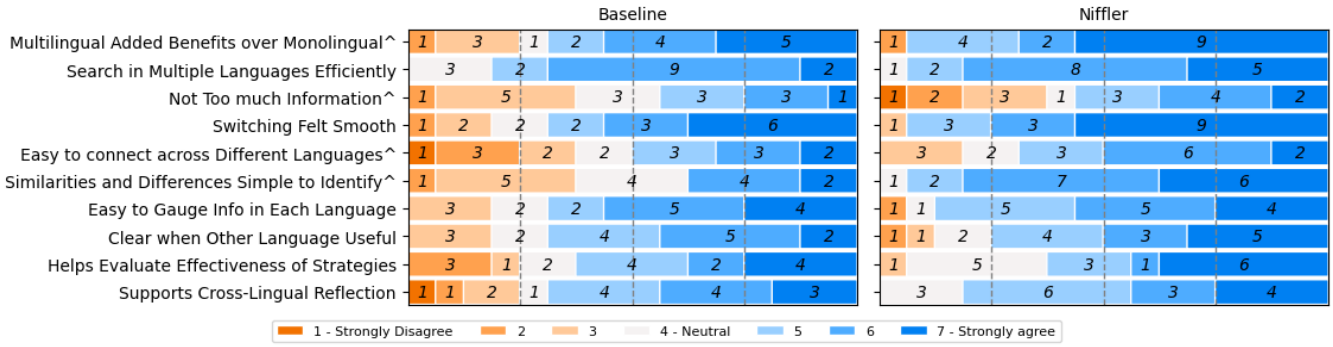

Figure 8: Distribution of Likert-scale Items from the Lab Study. Statements marked with ^ are flipped for reporting. Original and full statements can be found in Appendix A

labels differed slightly (e.g., "cheese dishes" vs. "cheese dish"). After refining the coding rules and reaching agreement on the pilot test, the remaining 50% of data were independently coded, this time reaching a percentage agreement of 86.5%, above the conventional threshold of 80% [42]. The points of disagreement were discussed until resolved. The top 5 topics for each task is presented in Table 3 for reference.

D Technical Evaluation

To further contextualize our study findings in terms of system capabilities, we conducted a technical evaluation of the generated comparative summaries. From the original 268 queries created across participants during the tasks, duplicates were removed, resulting in 230 unique queries. We then employed stratified sampling, randomly selecting 10 queries from each task (career, food) × language (English, Chinese) stratum, yielding a total sample of 40 queries

Table 2: Statistical comparison between Niffler and Baseline system logs. We report the mean (M), p-value, and effect size (r) for each measure. * and bolding indicates statistical significance.

Measure	M_{Baseline}	M_{Niffler}	p	r
***Number of queries**	6.75	10.0	0.041	0.549
Language switches	1.44	1.94	0.401	0.221
***Language Engagement Span**	0.657	0.403	0.019	0.601
***Language Balance**	0.381	0.690	0.017	0.642
Number of Sources Gathered	8.13	9.44	0.404	0.207
***Number of topics explored**	5.00	6.69	0.032	0.597

Table 3: The top 5 topics participants covered for each task, based on system logs.

Food	**Career**
food recommendation	career planning
swiss cuisine	career advice
restaurant recommendation	LinkedIn tips
Rösti[1]	networking
international restaurants	curriculum vitae (CV)

along with the corresponding generated comparative summaries. We recruited 2 experts who are fluent in both English and Chinese, and regularly seek information online. They were asked to rate and comment on the components of the comparative summaries — Cross-Lingual Comparison, Summary of sources in the original query language (L1), and Summary of Sources in the other language (L2), in relation to the query context. Ratings were based on three dimensions: **accuracy** (binary; "The summary is accurate."), **answer relevance** (7-point Likert; "Relevant information is provided.", and **context relevance** (7-point Likert; "No irrelevant information is provided."), adopted from [22].

On average, the comparison components had an accuracy of 87.5%, summaries of sources in the original query language (L1) had an accuracy of 100% and summaries of sources in the other language (L2) had an accuracy of 97.5%. Answer relevance ratings were 6.25 (out of 7) for the comparisons, 6.4 for L1 summaries, and 6.15 for L2 summaries. Context relevance ratings were 6.18 for the comparisons, 6.33 for L1 summaries, and 6.03 for L2 summaries. Overall, the comparative summaries appeared reasonably accurate and relevant. Issues mainly occurred due to (1) meaning lost in translation, particularly in query interpretation, and (2) the information provided or the point of comparison being overly generic or overly specific.