

Direct Use of Corpora in Japanese Language Education: Implementing DDL at the Upper-Intermediate Level

Giuseppe PAPPALARDO

Ca' Foscari University of Venice, Italy
giuseppe.pappalardo@unive.it

Abstract

Large-scale corpora offer access to authentic language and represent a valuable resource for Japanese language education. In current practice, they are mostly used indirectly by instructors, while direct learner engagement with corpora (data-driven learning, DDL) remains relatively limited in Japanese courses, partly due to time constraints and the complexity of corpus tools. This article reports on a four-year project conducted at Ca' Foscari University of Venice. The course integrated direct corpus consultation through NINJAL corpora accessed via Chūnagon, with the dual aim of supporting language development and fostering research-oriented skills for working with Japanese primary sources. The study provides a qualitative evaluation based on students' corpus-based presentation projects and anonymous student comments, highlighting both perceived benefits and critical issues, thus offering practical insights into the feasibility of DDL in Japanese language education.

Keywords: data-driven learning (DDL), Japanese language education, corpus literacy, NINJAL corpora, graduate-level language teaching

Povzetek

Obsežni jezikovni korpusi omogočajo dostop do dejanskega jezika in predstavljajo pomemben vir pri poučevanju japonščine. V obstoječi praksi jih učitelji večinoma uporabljajo posredno, medtem ko je neposredno delo študentov s korpusi (podatkovno vodeno učenje, DDL) pri pouku japonščine še vedno razmeroma redko, deloma zaradi časovnih omejitev in zahtevnosti korpusnih orodij. Članek predstavlja štiriletni projekt, izveden na Univerzi Ca' Foscari v Benetkah. V okviru predmeta je bilo neposredno delo s korpusi NINJAL prek sistema Chūnagon vključeno z dvojnimi cilji: podpreti razvoj jezikovnih kompetenc ter spodbujati raziskovalno usmerjene spretnosti pri delu z japonskimi primarnimi viri. Študija prinaša kvalitativno evalvacijo, ki temelji na študentskih predstavitvenih projektih, zasnovanih na korpusnih analizah, in anonimnih komentarjih študentov. Rezultati osvetljujejo tako zaznane prednosti kot tudi ključne izzive pristopa ter ponujajo praktični vpogled v izvedljivost podatkovno vodenega učenja pri poučevanju japonščine.

Ključne besede: podatkovno vodeno učenje (DDL), izobraževanje na področju japonščine, korpusna pismenost, korpusi NINJAL, poučevanje japonščine na podiplomski ravni



1 Introduction

Over the past three decades, corpus linguistics has progressively entered language education, ranging from corpus-informed materials to the more radical option of allowing learners to consult corpora directly as part of their learning process (Bernardini, 2004; Chambers, 2007; McEnery & Xiao, 2011). This latter option—often labeled Data-Driven Learning (DDL)—invites learners to observe authentic language data, notice patterns, form hypotheses and test them against corpus evidence, with the teacher acting primarily as a mediator and facilitator rather than a provider of rules (Johns, 1991, 1994; Gavioli, 1997).

DDL has been studied extensively in English-language contexts, where research and meta-analyses generally point to positive effects on vocabulary learning, collocation awareness, and aspects of grammatical development, while also highlighting the importance of task design and guided support (Mizumoto & Chujo, 2015; Ueno & Takeuchi, 2023; Lusta et al., 2023). In recent years, scholars have also called for broader methodological and theoretical perspectives on DDL, including richer qualitative data and closer engagement with learning theory (O’Keeffe, 2021).

By contrast, empirical work on DDL in Japanese language education remains comparatively limited and often tool-specific or teacher-mediated. A recurrent pattern is the use of corpora indirectly to produce teaching materials (e.g., frequency lists and example collections), while direct learner access is less frequently reported and tends to occur in highly scaffolded settings (Tanabe et al., 2012; Hori, 2019, 2021; Yovkova-Shii, 2023). This imbalance is striking given the maturity of Japanese corpus resources and the strong relevance of corpus skills to students engaging with Japanese primary sources in advanced academic and professional contexts.

This article contributes to this gap by presenting a qualitative analysis of a four-year experiment that implemented direct corpus consultation in two master’s-level Japanese language courses at Ca’ Foscari University of Venice. The study addresses two research questions: (1) What kinds of corpus-based investigations did students conduct after a guided training phase, and what do these types of investigation reveal about the affordances of direct corpus consultation at the upper-intermediate level? (2) How did students experience the intervention, as reflected in their open-ended feedback, and what design implications follow for Japanese DDL at the graduate level?

2 Previous studies on DDL

DDL is commonly traced back to Tim Johns' early work on classroom concordancing, where he framed learners as "detectives" who uncover linguistic regularities by interrogating data rather than memorizing rules. In what has become the field's most frequently cited slogan—"Every learner is a Sherlock Holmes!"—learners are invited to observe patterns in concordance lines, test hunches against evidence, and revise interpretations as new examples complicate or confirm initial assumptions (Johns, 1991). The pedagogical value lies not only in exposure to authentic usage but also in inquiry processes—noticing, comparison, hypothesis formation, and iterative refinement—which can foster autonomy and metalinguistic awareness (Johns, 1994).

Subsequent research has examined a wide spectrum of DDL configurations, from hands-on concordancing to guided, needs-driven corpus tasks (Braun, 2007; Leńko-Szymańska & Boulton, 2015). A consistent conclusion is that corpus consultation is rarely "self-explanatory": learners typically require mediation, training in query formulation and interpretation, and tasks that connect corpus patterns to meaningful communicative or academic goals (Chambers, 2007; Crosthwaite, 2020). Evidence syntheses support the overall effectiveness of corpus use and DDL, while also underscoring variability and the role of learner level, task type, duration, and instructional support (Mizumoto & Chujo, 2015; Ueno & Takeuchi, 2023). A recent systematic review similarly highlights persistent implementation barriers—technical constraints, limited corpus literacy, and the risk that learners experience DDL as cognitively demanding or peripheral—unless interventions are carefully scaffolded and aligned with assessment-relevant course aims (Lusta et al., 2023). Complementing these syntheses, de Carvalho (2025) maps influential work and shows how DDL research has diversified into multiple strands, including pedagogical applications and technological tools and resources.

Although DDL is now a well-established paradigm, the empirical literature remains dominated by English-language contexts, a pattern also noted in reports from non-English settings (Yovkova-Shii, 2023). At the same time, DDL has been adapted to other languages, including Italian, where learner-friendly corpora and scaffolded activities have supported classroom integration without assuming advanced technical expertise (Corino, 2019). Against this backdrop, Japanese remains comparatively under-documented: published studies are fewer and often rely on guided tools or teacher-mediated materials rather than fully learner-driven corpus querying, raising questions about the sustainability of direct consultation in Japanese courses.

In Japanese language education, corpus-informed pedagogy has a comparatively long trajectory, particularly in corpus-based teaching materials and grammar resources (Sunakawa, 2016; Mori, 2018; Nakamata, 2021). It is useful to distinguish corpus-informed language teaching—where corpus evidence shapes the teacher’s choices—from corpus-based language teaching, where corpus data become part of learners’ inquiry (often aligned with DDL). However, direct learner-driven DDL remains underrepresented in the published literature. Tanabe et al. (2012) describe a DDL-labeled practice built around a newspaper corpus and teacher-prepared, corpus-derived materials targeting collocational behavior of two-kanji Sino-Japanese items, exemplifying a guided configuration with restricted learner agency. Terajima (2011) reports an intensive upper-level course in which corpus consultation supports collocation learning and kanji vocabulary development through a structured workflow and the Kabosu tool (an Excel-based learning support tool for recording corpus findings and generating practice materials). Another strand centers on *Hagoromo*, a web-based database/interface that provides corpus-attested examples of function words for grammar learning: Hori (2019) reports classroom implementation with learner consultation and sentence production, while Hori (2021) links metacognitive self-ratings and performance and emphasizes the need for explicit guidance. Yovkova-Shii (2023) likewise uses *Hagoromo* in iterative cycles of production, consultation, and feedback and notes constraints typical of hybrid/remote settings. Overall, these studies suggest feasibility and generally positive reception, but also convergence on the need for clear task framing, tool-specific training, and sustained mediation.

Against this background, the present study examines an under-documented configuration: direct corpus consultation via a full-featured concordancer (Chūnagon) embedded in an upper-intermediate master’s curriculum, where language-learning aims intersect with academic research skills. By tracing learner searches and learner feedback longitudinally, the study responds to calls to widen the empirical lens on how DDL is enacted in real courses (O’Keeffe, 2021) and to make corpus literacy development visible as a pedagogical object rather than an assumed prerequisite. This emphasis is consistent with recent calls to treat corpus literacy as a foundational condition for sustainable DDL adoption, requiring explicit instruction in query formulation, interpretation, and critical evaluation of corpus evidence (Ma, 2025).

3 Implementing DDL in upper-intermediate Japanese language courses

3.1 Context and resources

The experiment was conducted in two upper-intermediate Japanese language courses (approximately at the JLPT N2–N1 transition) offered within two master's degree programs at Ca' Foscari University of Venice. Each class typically includes around 40 attending students and consists primarily of first-year master's students, i.e., learners in their fourth year of Japanese study. The courses run in the second semester, marking the final stage of the intermediate level and functioning as a bridge: they consolidate upper-intermediate competence while equipping students with the tools needed to progress into advanced-level coursework. During the same semester, Japanese instruction comprises seven weekly 90-minute classes: one taught by a non-native instructor and six taught by native-speaker instructors. The six native-speaker classes are explicitly aimed at consolidating the four language skills (written and spoken comprehension and production). The DDL component is implemented in the one weekly class taught by the non-native instructor, where the focus can extend beyond skills practice to methodological training, drawing on the instructor's expertise not only in language teaching but also in metalinguistic analysis of Japanese.

The first program, *Language and Civilization of Asia and Mediterranean Africa*, typically attracts a humanities-oriented cohort with interests ranging from literature and cultural history to translation and discourse studies. The second program, *Languages, Economics and Institutions of Asia and North Africa*, tends to attract students oriented toward social sciences and specialized registers (e.g., legal, economic, political, and institutional discourse). Across both cohorts, students share a practical need: reading, analyzing, and citing Japanese primary sources for academic purposes. This profile is not unique to the present setting. In Italy—and, more broadly, in many European contexts—upper-intermediate Japanese instruction is concentrated in university settings, particularly at the master's level, where language study is tightly intertwined with disciplinary content and research-oriented academic tasks.

For this reason, at the master's level, the goal cannot be framed solely in terms of developing the four language skills. Students are expected to engage with Japanese critically and to work from authentic usage rather than from isolated rules. This involves observing patterns, formulating and testing hypotheses, and explaining variation across register, medium

(written vs. spoken), genre, domain, and—where relevant—time period, while building procedures for empirical inquiry. In this sense, master's-level study requires more than learning Japanese: it requires learning *through* Japanese and learning *about* Japanese. This institutional and epistemic orientation motivates the decision to integrate corpus consultation not as an optional add-on but as a core component of the course, bridging language development, digital literacy, and research training. The corpus-literacy skills practiced through DDL are also directly reusable for thesis work, especially for students pursuing linguistics-oriented projects: synchronic and diachronic language research increasingly presupposes corpus-based evidence and the ability to operationalize research questions through systematic querying and interpretation.

In terms of resources, the corpora used were primarily those developed and hosted by NINJAL and accessed through Chūnagon, a web-based concordancer. Students were not restricted to a single dataset: they could consult any of the corpora available through the Chūnagon portal, selecting resources that matched their linguistic questions and research interests. From a pedagogical perspective, this environment encourages learners to treat Japanese as a variable system rather than a fixed set of textbook rules: depending on the chosen corpus, learners can explore how usage varies across registers, genres, communicative settings, and—in the case of historical resources—across periods.

3.2 Intervention design

The intervention followed a two-stage design repeated across four academic years (AY 2021/2022–2024/2025). Stage 1 consisted of a guided training phase focused on corpus literacy and query skills. Stage 2 required students to complete an individual mini-research project based on corpus evidence and present it orally with slides.

Stage 1 was conceived as systematic familiarization with both Chūnagon and corpus-based linguistic reasoning. To anchor this training in content that is both level-appropriate and analytically fruitful, activities used JLPT N1 grammatical expressions as shared objects of inquiry. These items are well-suited to corpus exploration because they often show morphosyntactic variation, sensitivity to register and medium (written vs. spoken), and recurrent co-occurrence patterns. Working with them helped students see—through evidence rather than assertion—that grammar is probabilistic, and usage is shaped by distribution, collocation, and discourse function, while also developing explicit awareness of collocation.

Instruction in Stage 1 proceeded predominantly deductively. Students began from a rule description in a grammatical-patterns reference and then tested it against corpus output, refining both their understanding and their searches. This routine made visible how attested usage can diverge from concise rule statements, revealing additional variants, distributional restrictions that emerge only in larger samples, and recurrent lexical environments not captured by standard examples. Through iterative cycles of hypothesis formation, search design, and interpretation, students acquired procedures that later supported independent inquiry.

A central goal was to develop an option-aware command of Chūnagon. Search steps were modeled through live demonstrations and guided worksheets, with collective troubleshooting to minimize unproductive trial and error. Students learned to treat “word” as an analytical unit in a morphologically annotated corpus, to navigate segmentation settings (e.g., Short Unit Word vs. Long Unit Word), and to switch flexibly between surface-form and lemma/lexeme searches. They combined lexical and structural constraints, used part-of-speech and morphological filters to handle inflectional variation, narrowed or compared subcorpora where relevant (e.g., by genre or text category), expanded context when needed, and refined searches iteratively to improve precision and interpretability.

Stage 2 shifted from guided training to student-led inquiry. Each student designed and carried out an individual corpus-based investigation and presented the results orally with slides (a written paper was required in the first two years but was later dropped due to perceived workload). Topic selection was explicitly connected to students’ research interests and to the specialized registers and genres associated with their degree programs. Projects therefore ranged from phenomena linked to literary language and stylistic choices to patterns relevant to sectoral and institutional discourse (e.g., legal, economic, political, policy-related, and administrative language), as well as broader questions concerning collocation, morphosyntactic constructions, discourse markers, writing-system variation, and diachronic change—provided that the phenomenon could be operationalized through queries in Chūnagon-accessible corpora. Depending on the question and the feasibility of extraction, students could pursue deductive or inductive inquiry and could adopt qualitative or quantitative approaches.

Teacher mediation continued in Stage 2, but with a different emphasis than in Stage 1. Rather than leading students through fixed tasks, the instructor provided consultations on research design, query construction, interpretation of concordance evidence, and basic reporting practices (including, when quantitative comparisons were attempted, the need to

normalize counts when corpora differ in size). Overall, Stage 2 served as a form of research apprenticeship in which language learning, linguistic analysis, and methodological reasoning were deliberately intertwined, and corpus literacy was positioned not only as a learning outcome but also as a transferable competence for empirically oriented thesis work.

3.3 Data and methods

This study adopts a qualitative design to examine the effectiveness and reception of a four-year DDL-based instructional experiment at the upper-intermediate level. Across the four academic years, the intervention involved two cohorts per year, with approximately 40 attending students per class (around 80 students per year). Given the exploratory and classroom-based nature of the project, the analysis draws on two complementary sources of evidence.

First, it examines student research outputs, namely the slide-based presentations in which students reported the results of an original, corpus-informed investigation. Across the four-year period, approximately 320 student projects were included. Projects were treated as a complete set rather than a sample and were organized analytically by type of analysis (i.e., recurring analytical moves), allowing the study to describe what learners were able to do with direct corpus consultation in practice. The analysis focuses on the linguistic phenomena selected for investigation and on the analytical approach adopted, treating projects as indicators of emergent methodological competence rather than as direct measures of language gains.

Second, the study analyses anonymous end-of-course student comments collected through institutional evaluation channels. Not all attending students provided written feedback. All available comments were included and were analyzed as naturally occurring reception data. For analytic clarity, comments were first classified by overall stance (broadly positive vs. critical) and then examined thematically to identify perceived benefits, difficulties, and suggested improvements.

Because the primary purpose of the project was to evaluate the feasibility and perceived instructional value of an innovative approach in a real university setting, the study is not presented as a controlled empirical evaluation of learning outcomes. Nonetheless, the combined dataset enables systematic qualitative analysis and supports grounded conclusions about (i) what forms of corpus-based inquiry are achievable at this level and

(ii) the main pedagogical and practical conditions shaping students' reception of direct DDL.

3.4 Results

3.4.1 Typologies of student corpus-based research

The slide-based student presentations document a wide range of investigations, illustrating that direct access to the corpora enabled learners to move beyond isolated “grammar points” toward broader questions about variation, discourse, and writing practices. A key outcome visible across projects is that students did not treat the corpus environment as a single default dataset. While the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) and the Corpus of Everyday Japanese Conversation (CEJC) (Koiso et al., 2022) were frequently used as baseline references for written and spoken data, students selected corpora strategically and with discernment, aligning resource choice with the logic of the research question. At the same time, these methodological choices did not occur in a vacuum: corpus selection and research design were supported by targeted guidance from both the course instructor (acting as a facilitator) and a designated tutor attached to these courses, who helped students assess dataset suitability and refine the feasibility of their queries. Projects concerned institutional and political language, for instance, tended to draw on corpora containing parliamentary or public discourse materials; projects on historical change exploited historical corpora accessible through Chūnagon, such as the Corpus of Historical Japanese (CHJ); studies on dialect or regional variation turned to the Corpus of Japanese Dialects (COJADS); and projects dealing with learner language or interlanguage contrasts sought out corpora where learner-produced data were available or comparable. This pattern of corpus selection is itself an indicator of emerging corpus literacy: students were not simply retrieving examples, but making reasoned decisions about representativeness, register coverage, time depth, and comparability before interpreting concordance evidence.

Across the four academic years, the investigations clustered around a set of analytical approaches. One prominent strand of work centered on morphosyntactic and constructional analysis. In these projects, students defined a grammatical target, translated it into explicit search parameters, and used concordance evidence to examine variation and constraints in usage. Suitable objects for this approach included, for example, *ra-nuki* forms (*ra-nuki kotoba* ‘r-deleted forms’) and alternations in Sino-Japanese verbs that can pattern both transitively and intransitively. Even when the

focus remained synchronic, this work encouraged students to connect formal patterning to distributional tendencies across genres or registers, rather than treating a construction as a context-free rule.

Closely related investigations adopted a more explicitly historical lens, where the research question required time depth and period-sensitive interpretation. Students selected historical corpora available through Chūnagon, operationalized a target phenomenon, and compared distributions across periods. This was also where basic methodological issues became especially salient: learners had to consider corpus comparability and to normalize counts when datasets differed in size, learning to interpret apparent increases or decreases cautiously rather than as straightforward proof of “language change”.

A different set of projects approached usage from the perspective of lexical meaning and phraseology, using collocation and distributional profiling to connect semantics to recurring co-text. Students examined near-synonym distinctions in domain-specific discourse and explored construction-like lexical patterns such as *-mamire* ‘covered, smeared with’ and *-zukume* ‘nothing but, entirely’. Although the lexical items varied across students, the analytic procedure was consistent: meaning was inferred from typical collocates, distributional environments, and evaluative tendencies that become visible only when one inspects larger sets of attested contexts.

Other presentations foregrounded discourse practices and specialized registers, often aligned with students’ disciplinary interests. In these investigations, corpora were used to examine patterned stance and evaluation in context, frequently via structural searches that captured the environments of target expressions. Constructions comparable to *-zaru o enai* ‘cannot avoid, have no choice but’ and modality expressed through *beki* ‘should, ought to’ served not simply as “advanced grammar” but as entry points for analyzing pragmatic positioning in institutional discourse.

Variation across register and text type was another recurrent analytical move. Here, students examined how competing expressions distribute across genres and why certain choices cluster in particular written domains, combining subcorpus filtering with close reading of extended contexts. Items comparable to *izure ni shite mo / izure ni seyo / izure ni shiro* ‘in any case, anyway’ were useful precisely because they push learners to explain stylistic choice empirically, in terms of genre conventions and discourse function rather than assuming interchangeability.

When students turned to spoken data, the focus often shifted to sociolinguistic and interactional variation. Learners selected spoken

datasets whose interaction types and participant profiles aligned with the research question and examined features such as sentence-final particles and their indexical effects.

Some projects were contrastive in orientation and used corpora to compare learner language (interlanguage) with native-speaker reference usage, where suitable datasets made such comparisons feasible. In these investigations, students operationalized a target form or alternation and then examined how its distribution differed across learner and native-speaker data. Contrasts comparable to *-te shimau* 'end up, unfortunately do' versus the contracted *-chau* 'end up doing (colloquial)' were used to explore how contraction, formality, and stylistic choice pattern differently in learner production and in native usage.

A distinct group of projects focused on writing-system and orthographic variation, using corpora to trace how representations shift across periods and genres. Rather than treating spelling or script choice as incidental, these investigations framed orthography as an empirical object: students tracked alternative written forms over time, examined genre-conditioned preferences, and used concordance contexts to relate orthographic choices to discourse conventions and historical change.

Separately, one presentation focused explicitly on Japanese dialect varieties, using a dialect corpus to investigate regionally anchored grammatical systems rather than treating dialectal features as sporadic "deviations" from the standard. The project examined nominative case-marking particles in northwestern Kyūshū varieties, tracing how alternative nominative markers are distributed across localities and contexts and what this reveals about dialect grammar as a structured, internally coherent system. Interpretation did not rely only on contrasts with present-day standard Japanese: drawing on students' background in Classical Japanese and the history of the language, the analysis also ventured informed parallels with earlier diachronic stages of Japanese, using historical knowledge to frame dialectal patterns as potential continuities or reanalyses within longer-term trajectories of change.

The range of student presentations can be summarized by grouping projects into broad types of corpus-based investigation that reflect shared analytical moves rather than recurring topics. To provide an overview of this repertoire and illustrate it with concrete examples, Table 1 lists the main types of investigation alongside selected titles of student research projects.

Table 1: Types of corpus-based investigations, with selected titles of student research projects

Types of corpus-based investigations	Selected titles of student research projects
Grammatical analyses	- Quantitative study of Sino-Japanese transitive/intransitive verbs - The spread of <i>ra-nuki kotoba</i> in written and spoken Japanese: a quantitative analysis
Diachronic variation	- A comparison of <i>kango</i> usage in Late Middle Japanese written and spoken production - Variation in the use of copular and adjectival forms in Meiji-period textbooks
Lexical analyses	- Collocations with the expressions <i>-mamire</i> and <i>-zukume</i> - Differences in usage between <i>kaisoku</i> (改革) and <i>kaisei</i> (改正) in Japanese political language
Political discourse analyses	- Verbs preceding the expression <i>-zaru o enai</i> in parliamentary debates - Diachronic variation of verbs preceding <i>-beki</i> in parliamentary debates from the 1970s to the 2000s
Text-type variation (including written–spoken differences)	- Usage variation of <i>izure ni shite mo / izure ni seyo / izure ni shiro</i> across different source types - A comparison of the use of the temporal expressions <i>toki</i> and <i>sai</i> across text types
Sociolinguistic analyses of spoken language	- Sociolinguistic variation in the use of sentence-final particles following <i>wa</i> - An analysis of the use of the masculine personal pronouns <i>boku</i> and <i>ore</i>
Native vs. non-native comparison	- Variation in the choice of obligation expressions in native and non-native Japanese - Quantitative analysis of <i>-te shimau</i> and <i>-chau</i>: a comparison between native and non-native speakers
Writing system and orthographic variation	- Changes in the written representation of Western country names from the Meiji period to the present - A quantitative analysis of differences in the written representation of the word <i>tabako</i>
Grammatical analyses of dialectal varieties	- Variation in the use of nominative case particles in northwestern Kyūshū dialects

Seen through a learning-as-participation lens, the projects can be read as evidence of developing research agency: learners were not only practicing forms but also participating in simplified versions of disciplinary inquiry (question–evidence–claim). This may be particularly relevant at master’s level, where DDL can function as a research apprenticeship rather than solely as a technique for accuracy gains.

3.4.2 Student feedback

Student comments provide a complementary perspective, capturing both positive and critical views of the intervention. Many respondents reported that the course was engaging and methodologically valuable, yet their feedback also reveals strong expectations about what a “language course” should prioritize at this level. Overall, the comments repeatedly touched on perceived usefulness/value, course organization/assessment, the need for more guidance/training, a desire for more traditional language practice, workload/time pressure, initial difficulty/overwhelm, and issues/logistics, which help clarify both the feasibility of direct DDL and the conditions under which it is experienced as legitimate and worthwhile.

A substantial portion of the feedback highlights perceived usefulness and the sense of gaining tools that extend beyond the immediate course. Some students described corpus-based inquiry as the most valuable part of their language training—*the most useful and valuable part of all the language courses I took*—and explicitly linked it to post-graduation applicability (*I believe it will be very useful even after graduation*). Others framed the main benefit in terms of methodological growth: *the course was very useful for acquiring new metalinguistic and methodological knowledge, or it made me autonomous in studying Japanese by providing advanced tools I wouldn’t otherwise have obtained*. Such perceptions resonate with DDL’s foundational claim that discovery-oriented work can cultivate independent learning strategies and metalinguistic awareness (Johns, 1991; Bernardini, 2004).

At the same time, the feedback stresses that corpus consultation is not self-explanatory and requires explicit training, step-by-step modeling, and sustained opportunities to practice. Some comments describe an initial affective barrier that later shifted into appreciation—*after my initial resistance (terror) toward Chūnagon, I learned a lot... I’m leaving enriched and glad I took it*—suggesting a learning curve rather than a fixed obstacle. Even for students with advanced proficiency, Chūnagon can be intimidating as a

research interface because it introduces Japanese technical terminology,¹ multiple layers of settings, and analytic constructs that are not part of everyday language study. This is consistent with the broader DDL literature, which emphasizes the role of mediation and careful task sequencing, especially when corpus work is not a minor supplement but a central pillar of the course (Braun, 2007; Chambers, 2007; Crosthwaite, 2020; Lusta et al., 2023). Comments also make visible that learning to use corpora extends beyond the concordancer itself: students requested support for handling and reporting results—such as organizing datasets, managing counts, and presenting findings—often mentioning spreadsheet skills as a practical need. This points to an “ecosystem” view of corpus work in pedagogy and aligns with accounts that conceptualize these procedural and interpretive competences as corpus literacy (Ma, 2025).

Several comments focused on course organization, coordination, and assessment expectations, calling for greater transparency about how training activities, independent project work, and evaluation fit together. Others expressed a desire for stronger integration with more traditional forms of language work and interaction. One student, for example, expressed a preference to revisit the grammatical structures already covered in the native-speaker classes with the non-native instructor, in Italian: *I would have preferred to discuss the grammatical structures covered with the native-speaking instructors in Italian. I believe that explanations given by an Italian instructor are necessary to grasp the subtleties of meaning.* Such remarks illustrate a recurring expectation that advanced language courses should prioritize guided interpretation and discussion of linguistic nuances. They also point to a broader design implication: corpus work is more readily accepted when it is explicitly connected to interpretive and communicative practice (Crosthwaite, 2020; Lusta et al., 2023).

Critical comments also reveal the limits of perceived relevance for some learners. While several students appreciated the approach, others described corpus work as *somewhat self-contained, with little practical*

¹ Here, ‘technical terminology’ refers to Japanese grammatical metalanguage used in Chūnagon’s interface and search settings, including terms for inflectional types (*katsuyōgata* 活用型) and the associated conjugation-class labels, as well as base forms (*katsuyōkei* 活用形), such as *mizenkei* 未然形, *ren’yōkei* 連用形, etc., with which not all students are equally familiar. It also includes analytic notions that are central to Chūnagon searches but are not intuitive for learners, such as lexeme (*goiso* 語彙素). In practice, students need to understand not only what a lexeme represents, but also that lexemes in the corpus are registered with a specific canonical written form (i.e., a particular character representation), which affects how searches should be formulated and interpreted.

usefulness for their Japanese, or felt that it did not lead to meaningful learning (*The Chūnagon activities did not increase my knowledge; I didn't acquire meaningful skills*). A related critique questioned the curricular centrality of the approach: *Basing the entire program on corpus research felt of limited usefulness... not essential for all students of Japanese*. These responses underscore a fault line in graduate language education: corpus-based inquiry can support language awareness and research competence, but it may be experienced as only indirectly related to communicative development unless its relevance is made visible through task design and assessment alignment.

Workload and time pressure also appear as constraints. Some comments highlight the cognitive and temporal demands of DDL, especially when students must simultaneously learn a complex interface, acquire procedural competence in searching, and develop interpretive confidence in handling authentic usage data. This aligns with evidence syntheses indicating that DDL can be experienced as demanding and time-consuming, particularly in early phases when tool proficiency is still developing (Chambers, 2007; Ueno & Takeuchi, 2023). Finally, a smaller subset of feedback concerns pragmatic conditions of access and learning logistics (reliable online access for working students, classroom arrangements for real-time corpus consultation). While not specific to DDL, these remarks matter because corpus pedagogy presupposes stable access to devices, connectivity, and working conditions that enable sustained interaction with concordance output.

Overall, student feedback suggests that direct DDL can be perceived as highly valuable at master's level—especially when learners recognize its relevance to independent study and research-oriented competence—but that its success depends on explicit framing, careful scaffolding of corpus literacy (including ancillary data-handling skills), and transparent alignment between course objectives and assessment.

3.5 Discussion and implications

The four-year implementation suggests that direct corpus consultation can be integrated meaningfully into upper-intermediate Japanese courses when the curriculum explicitly values methodological training alongside proficiency development. In line with DDL scholarship, the intervention treats corpora not as repositories of examples but as environments for inquiry that can cultivate evidence-based reasoning and analytical habits of mind (Bernardini, 2004; O'Keeffe, 2021). From a second-language learning

perspective, this inquiry orientation can be read as a structured way of promoting noticing and hypothesis testing: learners are repeatedly prompted to connect form, distribution, and context, and to revise interpretations in light of evidence rather than relying on single-rule generalizations.

Student feedback indicates that feasibility alone is not enough: “successful integration” depends on alignment with learners’ expectations. Many students welcomed the research dimension and the autonomy it enabled. In contrast, others questioned its relevance when corpus work was perceived as self-contained or as shifting the course toward “linguistics” rather than Japanese. This tension echoes the wider literature, which shows that perceived relevance and usability are decisive for uptake even when DDL is methodologically robust (Crosthwaite, 2020; Lusta et al., 2023). A useful way to interpret these mixed responses is through learner engagement: the same tasks can generate high engagement when learners see a clear pathway from corpus evidence to communicative or academic payoff, but lower engagement when the activity is experienced primarily as “tool use” without an obvious link to language use.

A central practical constraint is that corpus interfaces are often inherently complex. Developing corpus literacy takes time, and in early phases learners may experience the main task not as analyzing corpus evidence but as learning to operate the concordancer—especially when the interface relies on specialized terminology and layered settings that directly affect what counts as evidence. In this study, some comments suggest that the perceived “course object” can temporarily shift toward Chūnagon itself, with substantial cognitive effort required to overcome the interface before attention can move to linguistic interpretation, a pattern also highlighted in qualitative syntheses of DDL barriers (Sun & Mizumoto, 2025). This observation supports the view that corpus literacy should be treated as a foundational competence rather than an assumed prerequisite (Ma, 2025). It also helps articulate learner engagement with corpus data along three complementary dimensions: behavioral engagement (iterative querying, selective use of filters and subcorpora, and refinement of search parameters), cognitive engagement (moving from example-hunting to hypothesis-driven searches and cautious interpretation), and affective engagement (initial anxiety or overwhelm shifting toward increased confidence as routines develop). When the interface absorbs most of the cognitive and affective bandwidth, engagement may remain procedural; when the interface becomes “transparent enough,” engagement can shift toward interpretation and language-related reasoning.

These considerations motivate three practical implications. First, corpus literacy should be front-loaded in ways that reduce extraneous cognitive load: staged introduction of key settings, repeated modeling, reusable query templates, and explicit support for ancillary skills such as basic data organization and counting in spreadsheets (Ma, 2025). Second, the curriculum benefits from visible bridges between corpus inquiry and language use, so that learners experience corpus work as more precise production and interpretation rather than as a parallel “method course” (Bernardini, 2004; O’Keeffe, 2021). Third, the initial interface barrier could be lowered by making the tool more user-friendly without reducing functionality—for example, through a multilingual layer for interface labels and help prompts, and a progressively disclosed “beginner view” that keeps advanced options available but introduces them step by step. This issue gains further salience as DDL is increasingly discussed in relation to AI-assisted research environments, from early corpus-based tools to large language models such as ChatGPT (Flowerdew, 2025).

Japanese-specific constraints further explain why interface and literacy issues are so salient. Morphological complexity and the absence of spaces mean that segmentation choices shape what counts as a searchable unit, while orthographic variation (kanji/kana alternation, variant characters, historical spellings) can complicate recall and precision. At the same time, the richness of annotation and coverage in Japanese corpora available through Chūnagon supports advanced questions about register, genre, and diachrony—goals that are highly compatible with graduate-level study. Making these constraints explicit, and at times treating them as objects of inquiry, can turn early difficulties into linguistic insight and can redirect engagement from interface frustration to linguistic interpretation.

Finally, the findings underscore the centrality of mediation. Across stages, the instructor’s role shifts from facilitating access and protecting learners from overwhelming complexity to coaching methodological decision-making during projects (corpus choice, parameter justification, and responsible interpretation). This aligns with Japanese-focused studies where guided consultation and instructor feedback are pivotal (Hori, 2019; Yovkova-Shii, 2023), and with broader DDL arguments that outcomes depend less on corpus access per se than on the quality of scaffolding and feedback (Crosthwaite, 2020; Lusta et al., 2023). In this view, mediation is not simply support for “using the tool” but a mechanism for sustaining engagement with corpus evidence—helping learners shift from procedural engagement to interpretive engagement as the course progresses.

4 Conclusion

This qualitative study examined a sustained four-year implementation of direct DDL in upper-intermediate Japanese courses at master's level, drawing on student presentation work and anonymous course evaluation comments. The analysis shows that, once students had developed basic corpus literacy, they were able—under guided support—to select corpora strategically, operationalize linguistic questions through explicit search parameters, and interpret concordance evidence with attention to register, medium, genre, and, where relevant, time depth. In this sense, direct corpus consultation supported not only language-focused noticing but also research-oriented practices aligned with graduate-level study (Bernardini, 2004; O'Keeffe, 2021).

At the same time, the study highlights that the main obstacles are rarely conceptual. Rather, they are practical and curricular: the time required to develop corpus literacy, the complexity of real concordancer interfaces, workload calibration, and—most centrally—the need for explicit alignment between inquiry-based work and students' expectations of a language course. Student comments suggest that when this alignment is not made visible, some learners experience the course as “about using Chūnagon” rather than as an analysis of Japanese usage through corpus evidence, despite acknowledging the methodological value of the approach. Addressing these issues through front-loaded corpus-literacy support, explicit bridges to language use, and (where possible) more user-friendly interface design could broaden the sustainability of direct DDL in Japanese language education (Crosthwaite, 2020; Lusta et al., 2023; Ma, 2025).

Future work could operationalize learner engagement with corpus data more explicitly (e.g., trace data from query logs, targeted assessments, and interviews) to link engagement profiles to learning outcomes and longer-term uptake. This agenda is likely to become even more pressing as corpus-mediated inquiry increasingly coexists with AI-assisted tools, making corpus literacy and interpretive judgment central skills for sustainable DDL.

References

- Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J. McH. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15–36). John Benjamins Publishing Company.
<https://doi.org/10.1075/scl.12.05ber>

- Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL*, 19(3), 307–328.
<https://doi.org/10.1017/S0958344007000535>
- Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom: Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6), University of Granada, Spain, 4–7 July 2004* (pp. 3–16). Rodopi.
- Corino, E. (2019). Data-driven learning: Tra lingue straniere e CLIL, tra ricerca e didattica. *EL.LE*, 8(2), 271–286.
<https://edizionicafofscari.unive.it/media/pdf/article/elle/2019/2/art-10.14277-ELLE-2280-6792-2019-02-001.pdf>
- Crosthwaite, P. (Ed.). (2020). *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.
<https://doi.org/10.4324/9780429425899>
- de Carvalho, C. T. (2025). Notable publications in data-driven learning. In L. McCallum & D. Tafazoli (Eds.), *The Palgrave encyclopedia of computer-assisted language learning* (pp. 1–7). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-51447-0_151-1
- Flowerdew, J. (2025). Data-driven learning: From Collins Cobuild Dictionary to ChatGPT. *Language Teaching*, 58(3), 359–376.
<https://doi.org/10.1017/S0261444824000144>
- Gavioli, L. (1997). Exploring texts through the concordancer: Guiding the learner. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 83–99). Longman.
- Hori, K. 堀恵子. (2019). Kinōgo webu tsūru o tsukatta jiritsuteki bunpō gakushū no kōka 機能語ウェブツールを使った自律的文法学習の効果 [Effects of autonomous grammar learning using a web-based function-word search system]. In *Proceedings of the 23rd Japanese Language Symposium in Europe* (pp. 568–580).
- Hori, K. 堀恵子. (2021). Jiritsuteki dēta kudō-gata gakushū ni okeru bunpō kōmoku no ninchido to bun sakusei no kankei 自律的データ駆動型学習における文法項目の認知度と文作成の関係 [Relationship between recognition of grammatical items and sentence production in autonomous data-driven learning]. In *Proceedings of the 24th Japanese Language Symposium in Europe* (pp. 174–185).
- Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom concordancing (English Language Research Journal, No. 4, pp. 1–16)*. University of Birmingham.

- Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 293–313). Cambridge University Press.
<https://doi.org/10.1017/CBO9781139524605.014>
- Koiso, H., Amatani, H., Den, Y., Iseki, Y., Ishimoto, Y., Kashino, W., Kawabata, Y., Nishikawa, K., Tanaka, Y., Usuda, Y., & Watanabe, Y. (2022). Design and evaluation of the Corpus of Everyday Japanese Conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5587–5594). European Language Resources Association.
<https://aclanthology.org/2022.lrec-1.599/>
- Leńko-Szymańska, A., & Boulton, A. (Eds.). (2015). *Multiple affordances of language corpora for data-driven learning*. John Benjamins Publishing Company.
<https://doi.org/10.1075/scl.69>
- Lusta, A., Demirel, Ö., & Mohammadzadeh, B. (2023). Language corpus and data-driven learning (DDL) in language classrooms: A systematic review. *Heliyon*, 9(12), Article e22731. <https://doi.org/10.1016/j.heliyon.2023.e22731>
- Ma, Q. (2025). Corpus literacy and data-driven learning. In L. McCallum & D. Tafazoli (Eds.), *The Palgrave encyclopedia of computer-assisted language learning* (pp. 1–7). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-51447-0_88-1
- Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., & Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2), 345–371. <https://doi.org/10.1007/s10579-013-9261-0>
- McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 364–380). Routledge.
- Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approaches in the Japanese EFL classroom. *English Corpus Studies*, 22, 1–18.
- Mori, A. 森篤嗣 (Ed.). (2018). *Kōpasu de manabu gengogaku: Nihongo kyōiku e no ōyō* コーパスで学ぶ言語学—日本語教育への応用 [Learning linguistics through corpora: Applications to Japanese-language education]. Asakura Shoten 朝倉書店.
- Nakamata, N. 中俣尚己. (2021). “Chūnagon” o katsuyō shita kōpasu Nihongo kenkyū nyūmon 「中納言」を活用したコーパス日本語研究入門 [An introduction to Japanese-language research using the Chūnagon corpus]. Hitsuji Shobō ひつじ書房.
- O’Keeffe, A. (2021). Data-driven learning—A call for a broader research gaze. *Language Teaching*, 54(2), 259–272.
<https://doi.org/10.1017/S0261444820000245>

- Sun, A. X., & Mizumoto, A. (2025). Exploring the barriers to data-driven learning in the classroom: A systematic qualitative synthesis. *Applied Corpus Linguistics*, 5(2), Article 100126. <https://doi.org/10.1016/j.acorp.2025.100126>
- Sunakawa, Y. 砂川有里子 (Ed.). (2016). *Kōpasu to Nihongo kyōiku コーパスと日本語教育* [Corpora and Japanese-language education]. Asakura Shoten 朝倉書店.
- Tanabe, K. 田辺和子, Chujo, K. 中條清美, Ito, S. 伊藤誓子, & Nishigaki, C. 西垣智佳子. (2012). Shinbun kōpasu o katsuyō shita Nihongo DDL kyōzai to shidō jitsurei 新聞コーパスを活用した日本語 DDL 教材と指導実例 [Japanese DDL teaching materials and examples of instruction using a newspaper corpus]. *Nihon Daigaku Seisan Kōgakubu Kenkyū Hōkoku 日本大学生産工学部研究報告*, 45, 73–82.
- Terajima, H. 寺嶋弘道. (2011). Nihongo kyōiku ni okeru kōpasu ōyō: Dēta kudō-gata no gakushū to sono jissen hōhō no kōsatsu 日本語教育におけるコーパス応用—データ駆動型の学習とその実践方法の考察 [Corpus applications in Japanese-language education: A study of data-driven learning and its practical methods]. *Polyglossia*, 20, 91–103.
- Ueno, S., & Takeuchi, O. (2023). Effective corpus use in second language learning: A meta-analytic approach. *Applied Corpus Linguistics*, 3(3), Article 100076. <https://doi.org/10.1016/j.acorp.2023.100076>
- Yovkova-Shii, E. ヨフコバ四位・エレオノラ. (2023). Dēta kudō-gata gakushū o toriireta bunnō kyōiku: “Nihongo riterashi” ni okeru jissen no kokoromi データ駆動型学習を取り入れた文法教育—「日本語リテラシー」における実践の試み— [Grammar instruction incorporating data-driven learning: An attempt at implementation in “Japanese literacy”]. *Toyama Daigaku Kyōyō Kyōikuin Kiyō 富山大学教養教育院紀要*, 4, 49–60.