International Journal of Information Technology and Language Studies (IJITLS)

Language Preservation and Semantization: Prototyping Automated Glossing of an Endangered Mixed Language Corpus

Karim Tharani
University of Saskatchewan, Saskatoon, Canada
karim.tharani@usask.ca

Article Information

Article type: Article

Article history:

Received: October 06, 2021
Revised: January 26, 2022
Accepted: January 26, 2022

Keywords:

Language semantization;
Endangered Language Preservation;
Online Glossing;
Ginans

Abstract

This article discusses the prototyping of an online vocabulary learning tool for the oral language of the ginans, a corpus of gnostic hymn-like poems of the Ismaili community. The language of the ginans is mixed and borrows vocabulary from various Indo-Aryan and Perso-Arabic dialects. The teachings encoded in the oral language of the ginans, therefore, remain foreign to the English-speaking community members living in the Western diaspora. This study is based on the premise that for the tradition and the teachings of ginans to be preserved in the diaspora, the successive English-speaking generations of the Ismaili community must learn and understand the vocabulary of the ginans. The process through which humans learn and understand the vocabulary of a language is called semantization. The glossing of foreign language (L2) materials with meanings in the native language (L1) of learners has proven to be an effective enabler of semantization. The prototype glossed ginan utilizing lexical resources, including a concordance and an English glossary to facilitate semantization of the ginan vocabulary. Using the design-based research (DBR) methodology, the prototype was implemented over two iterative design cycles. During the evaluation of the prototype by target learners, over 90% of the participants indicated that they would make use of the prototype when made available publicly.

1. Introduction

The heritage of ginans of the Ismaili community comprises over 1,000 individual hymn-like poems that were composed and transmitted orally. The ginans were composed as early as the eleventh century by several preacher-saints who are known as *pirs* and *sayyids* in the community. The teachings encoded in the oral language of the ginans guided the normative understanding of the community in India. The historical practice of composing ginans came to an end in the mid-nineteenth century but the texts and tunes of the ginans accompanied the members of the community to the West. While numerous ginans have been transcribed using the Latin script, the teachings of the ginans remain foreign to many English-speaking community members in the diaspora.

The cultural contexts and lived experiences of the English-speaking Ismaili youth born and raised in the West are very different from those of their elders who had to leave their native land and languages behind. As information technology permeates our everyday lives, it is natural for the younger members of the community to want to engage with their heritage

online and *on-demand*. This study designed and prototyped an online semantization tool for the mixed language of the ginans by incorporating Western best practices of language learning.[1]

2. RELATED WORKS

Mixed languages attract scholarly interest as they are uniquely formed through the phenomenon of "language contact," whereby people speaking two or more unrelated languages interact with each other at the same place and time (Bakker & Mous, 1994). For this reason, mixed languages are also referred to as *contact* languages. Initially, it was believed that Michif of the Métis people was the only known mixed language of the world, but other mixed languages have also been identified since then (Bakker, 2017). But even today, only a handful of over 7,000 documented languages are classified as contact languages (Lewis, Simons, & Fenning, 2020). Some of these mixed languages, like Michif, are also endangered and in dire need of preservation to remain alive (Barkwell, 2017; Petten, 2006).

The language of the ginans is the result of extreme mixing of multiple Indo-Aryan languages, including Gujarati, Sanskrit, Hindi, Urdu, Panjabi, with loanwords from Arabic and Persian (Kassam, 1995). This mixed language enabled the composers of the ginans to draw from the "bewildering thicket of Indian religions, mythologies and intellectual traditions" (Alibhai, 2020). Since the ginans were composed and transmitted orally, the language of the ginans remains without an established grammar or writing system of its own. However, the community devised a script called *Khojki* to transcribe ginan texts in manuscripts (Asani, 2002). Tracing its roots to mercantile communities of India, Khojki is not a language but a script that was utilized as a shorthand for transcribing the ginans.

With the advent of printing technology, the community founded the Khoja Press of India in 1903 to create an official canon of the authorized texts of the ginans in Khojki. The press used specialized German-made fonts for the Khojki script for printing the ginan corpus (Asani, 2011). Later, as the community embraced Gujarati as its *lingua franca*, the ginan texts were converted and printed in Gujarati for the community use. As the community members migrated away from the Indian subcontinent to Western countries in the 1970s, the ginan texts were also printed using the Latin or English script. Consequently, the ginan corpus today includes texts in various scripts, including Khojki, Gujarati, Urdu, and English (Latin).

While the instinctive adaptability of the community has been instrumental in preserving and transporting the corpus of ginans over time and geography, the vocabulary of the ginans remains inaccessible to the English-speaking community members in the Western diaspora. When it comes to learning the vocabulary of special or foreign languages, *semantization* is regarded as one of the most fundamental steps in mastering any language (Schmitt, 2008). Semantization is the process through which humans learn to ascertain multiple and contextual meanings and senses of the words of a language.

2.1 Language Learning and Semantization

In the literature on language learning, the terms L1 and L2 are commonly used to refer to the first and second languages of the learners. The rationale is that a learner's native language (L1) is the one that they first learn as a child. Any additional language that they learn during their life is regarded as their second (L2) or additional language. When it comes to L2, semantization is "a continuing process of getting acquainted with verbal forms in their polysemous diversity within varying contexts" (Beheydt, 1987, p. 55). For L2 learners to be successful, Beheydt also asserts that "it is essential that the learner be provided with a number of concrete representative usages of each word as a basis for the correct semantization of a word" (p. 61).

An important aspect of vocabulary semantization is recognizing that words are polysemous by nature, that is they are used in multiple senses. A word sense "is a discrete representation of one aspect of the meaning of a word" (Jurafsky & Martin, 2019, p. 354). Therefore, having access and exposure to language corpus is an important prerequisite for language learning. As Poole notes, "students exposed to authentic usages of vocabulary in multiple meaningful contexts may more efficiently semanticize and more completely acquire vocabulary than learners who are not exposed to the varied usages of lexical items" (2011, p. 3).

The emerging evidence in L2 literature suggests glossing, which is the practice of embedding brief definitions of content words in texts, to be an effective technique to facilitate vocabulary semantization for learners. Schmidt (2008) notes that "there are several reasons why glossing can be useful: more difficult texts can be read, glossing provides accurate meanings for words that might not be guessed correctly, it has minimal interruption to reading – especially compared to dictionary use—and it draws attention to words that should aid the acquisition process" (p. 351). Glossing can either be concordance-based (meaning-inferred) or glossary-based (meaning-given). In the concordance-based modality of glosses, learners are encouraged to infer meanings of new and unknown words from sample corpus texts. In contrast, the glossary-based glosses provide the meanings of the new and unknown words.

While studies in the L2 literature have generally confirmed the efficacy of glossing in enhancing learners' incidental vocabulary (Azari, 2012), the conclusions of the studies on comparing gloss modalities have been mixed. Such inconclusive

[1] The prototype of the Online Ginan Learning Tool is accessible at https://ginans.usask.ca/semantics/prototype.

results could be explained by the fact that such comparative studies have been undertaken on the pretext of proving the superiority of one gloss modality over the other. Rather than taking a comparative research stance, this study is based on the premise that with the use of information technology, it is possible to provide both gloss modalities to learners simultaneously. Thus, the prototype designed and developed in this study provides both gloss modalities, concordance-based and glossary-based, to learners simultaneously.

3. METHODOLOGY

A sizable collection of the ginan materials has already been gathered over the past 10 years at the University of Saskatchewan Library in collaboration with the Ismaili community. A sample of ginan texts in Latin script from this collection was enriched with glossing by utilizing lexical resources, including a concordance and an English glossary to facilitate vocabulary semantization. This study adopted design-based research (DBR) methodology, which is particularly well-suited for designing educational prototypes and products "through iterative analysis, design, development, and implementation based on collaboration among researchers and practitioners in real-world settings" (Wang & Hannafin, 2005, pp. 6-7).

When developing prototypes using DBR, the initial iteration begins with a theoretical design grounded in the literature, which is refined with subsequent iterations based on input from educators and learners in the field (Holmes, 2013). The design process for prototyping the tool in this study comprised two design iterations of analysis, design, and review (Figure 1).

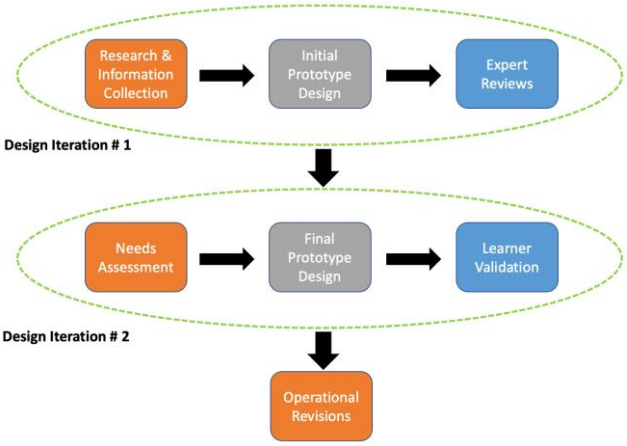

Figure 1. The Design-based research methodology used for prototyping the semantization tool

3.1 Initial Design Iteration

In the initial iteration, the initial design of the prototype was based on a multidisciplinary literature review and an in-depth semantic analysis of the available ginan materials. The focus of this iteration was to design an initial design or a proof-of-concept for the envisioned online ginan semantization tool. The design at this stage of the research was essentially a conjecture informed by the review of the existing semantization literature and the resulting semantic analysis of the sample corpus used for this study. The design wireframe gathered the necessary resources in one place for a given ginan. These resources included romanized ginan texts, English translation, audio recitations, as well as a glossary as illustrated in the figure below (Figure 2). The wireframe of the design went through several refinements as the desired functionalities and user-experience decisions were made. For instance, rather than listing all the recitals of the ginans, a specific recital conducive to learning was chosen for simplicity. A single link to other recitals was then presented for learners interested in listening to other renditions of the ginan. The initial design was used as a basis to solicit formative feedback from ginan reviewers who have experience in teaching ginans in the community.

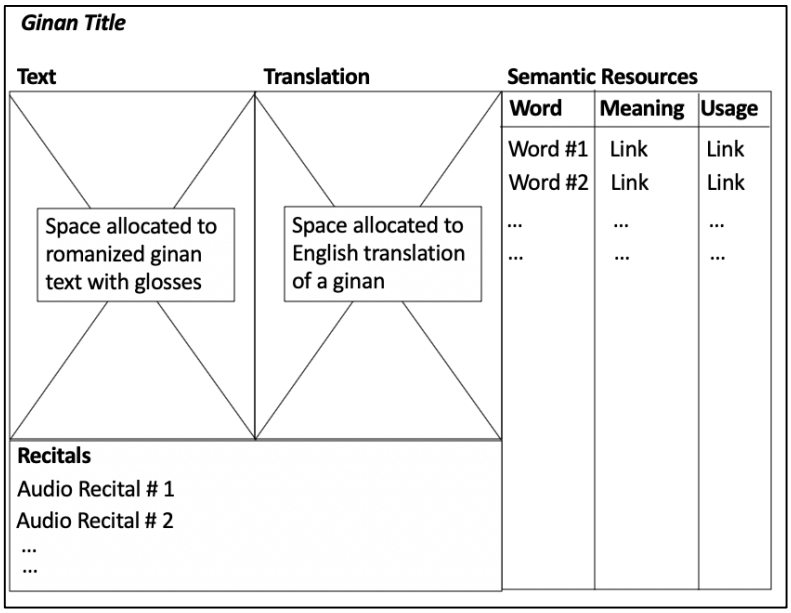

Figure 2. A wireframe of the initial prototype design

3.2 Final Design Iteration

In the second design iteration, the initial design of the prototype was evolved by adding several features to make it an operational learning tool (Table 1). For transforming the initial design into a functional prototype, both the needs of the learners as well as the best practices of community instructions from experts were synthesized and ranked to identify the feasibility of the desired features. The functional design considerations include aspects of placement and prominence of content, user interface, and navigational aspects (Figure 3).

Table 1. Prototype features in the final design

	Feature Type
A	Composer information
B	Ginan categories
C	Access to other resources
D	Ginan uncommonness
E	Extended audio player
F	Ability to print
G	Link to the pronunciation guide
H	Ability to download

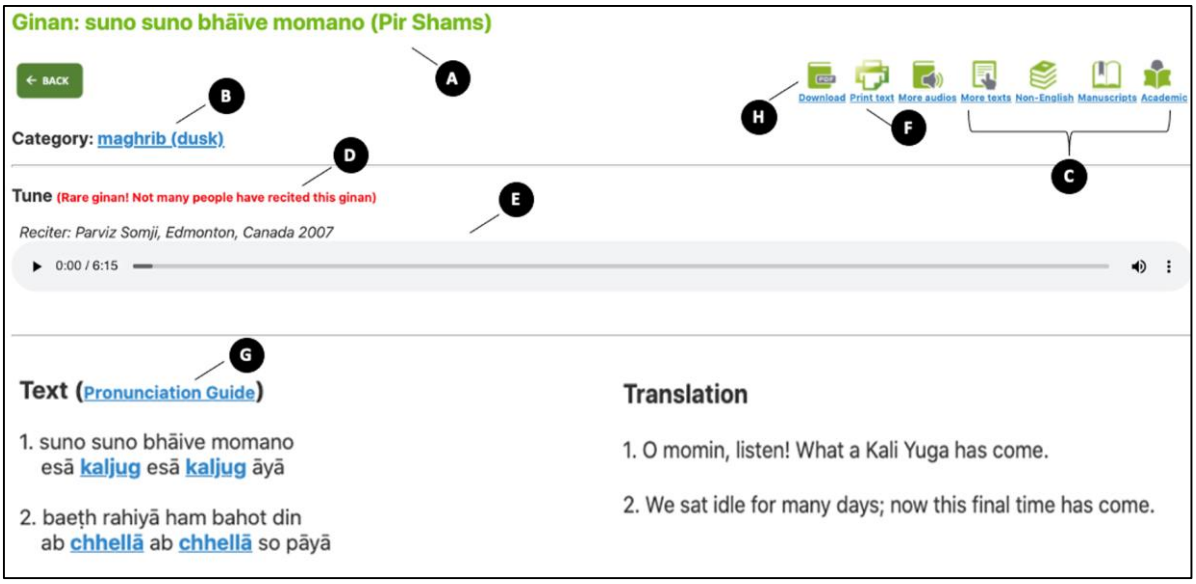

Figure 3. The final design of the prototype

4. RESULTS

The glossing process used in this research utilized an online lexicon, which was developed as part of the project by integrating the concordance and glossary of the ginans.[2] In today's age of full-text searching, where entire documents can be indexed efficiently and economically for full-text indexing, the need for a concordance may seem questionable. While full-text indexing may be efficient for searching terms in individual documents, it is unable to link documents together semantically. The ability to build semantic relationships amongst words and texts is crucial for developing semantic profiles to enhance semantization. The overall process followed to make these materials accessible for the online semantization prototype is illustrated in the figure below (Figure 4).

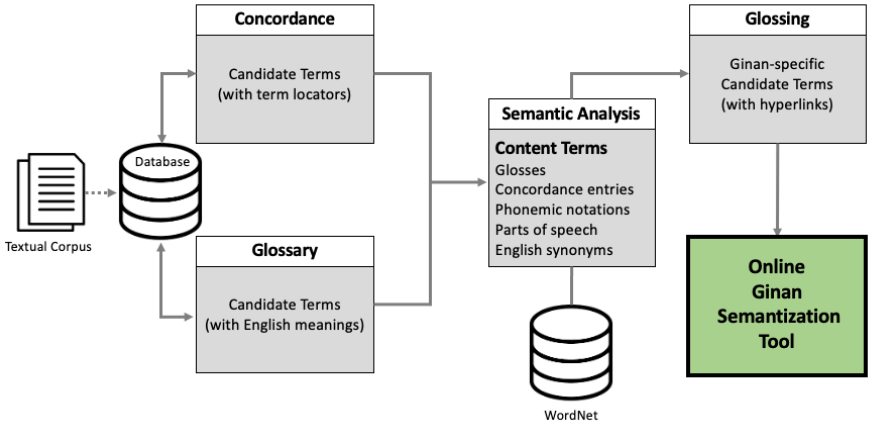

Figure 4. Glossing process overview

4.1 Identifying Candidate Terms

A candidate term is a basic meaning-bearing lexical unit of a language with one or more words that may potentially qualify for glossing. The sample ginan texts used in this project were enriched with glosses by integrating concordance and glossary of the ginans. The concordance in this project served the dual purpose of identifying candidate terms for semantization and collocating the terms within the selected corpus. Since glossaries typically do not provide the location of

[2] The online lexicon and semantic profiles can be accessed at https://ginans.usask.ca/semantics/analysis/content_terms_index.php.

terms in the corpus, the use of concordance provided the additional benefit of generating glossaries for specific ginans. There were 347 unique candidate terms, which were collectively used 479 times in the chosen sample of the ginan texts.

4.2 Selecting Content Terms

The content terms are words (nouns, verbs, adjectives, and adverbs) that are deemed pertinent for a learner to know and understand. The candidate terms obtained from the concordance were matched against the terms in the glossary to identify content terms for glossing using automation. A successful match between the concordance and glossary terms was used as the criterion for the selection of content terms. The automated process resulted in the selection of 119 content terms, which amounts to 34% of the candidate terms. Since back-of-the-book glossaries typically do not provide the location of terms in the corpus, the matching of concordance and glossary terms provided the additional benefit of generating glossaries for the selected ginans.

Once the glossary terms for a specific ginan were identified, the glossing algorithm was programmed to go through each line of the ginan text and insert hyperlinks for the terms in the line that matched any of the candidate terms. At the end of the process, all candidate terms appeared as hyperlinks in the ginan text that linked to their respective terms in the ginan-specific glossary presented at the end of the text. Additional links to pronunciation sound files and semantic profiles were also incorporated into the ginan glossary.

4.3 Identifying Word Senses

Since words are generally polysemous (having many meanings or senses), the fundamental principle of semantization is the ability for learners to identify and understand various senses of the terms. For this study, *word sense disambiguation* (WSD), which is the process of "determining which sense of a word is being used in a particular context" (Jurafsky & Martin, 2019, p. 354), was utilized. It entailed matching and retrieving word senses of the English synonyms of the candidate terms from the Online WordNet database.[3] The ginan glossary was used for matching and retrieving lexical information from WordNet. The WSD algorithm resulted in matching 73% of the content terms. Upon manual verification, the total number of correct matches was reduced to 26%. This sizable discrepancy confirms the polysemous nature of the ginan vocabulary and the need for developing richer ginan lexicons to improve automated word sense disambiguation.

4.4 Developing an Online Semantic Profile

The semantic profile or lexicon created for this study enumerates and describes pertinent attributes of the content terms. A semantic profile template was designed to capture these attributes consistently as illustrated in the table below (Table 2). Having a semantic profile template made it easier to store the data electronically to facilitate online semantization. The online lexicon data was stored using several tables in a MySQL database and was made accessible on the Web using HTML markups and PHP scripting.

Table 2. Semantic Profile Template for Candidate Terms

Attribute	Description
Content term ID	A unique numeric identifier assigned to a term.
Content term	Content word or phrase in romanized form.
Variant form(s)	Variant romanized transcriptions of the term, if any.
Phonetic notation	A pronunciation guide of the term using the Latin script.
Part of speech	The part of speech of the term e.g., noun, verb, etc.
Gloss	A short sentence describing the sense of the content term.
English synonyms(s)	Equivalent words in English for a given sense of the term.
Equivalent term(s)	Other ginan terms that share the same meaning and sense.
Usage example	Excerpts of ginan texts that use the content term
Related terms	List of semantically related words.
Related ginan(s)	List of ginans that use the content term.
Source(s)	List of ginan materials referenced.

The resulting prototype was shared with potential community users to validate if the prototype is an effective online ginan learning tool. On the 5-point Likert scale, the overall rating of the prototype was 4.64 (93%) based on the weighted average of the responses. The ratings of the three review categories (content, interface, and navigation) also received ratings of between 3.91 to 4.64 based on the weighted average. Over 90% indicated that they would make use of the proposed online ginan tool designed and developed as part of this study.

[3] Online WordNet® is an English-language lexical database accessible at https://wordnetweb.princeton.edu/perl/webwn.

5. SIGNIFICANCE

A language remains alive if there are people who continue to use the language in their everyday lives. For a language to remain relevant to people and to attract new learners, it must also have a growing base of popular, intellectual, and instructional materials in a variety of formats (text, audio, video, web, etc.) that language users can consume and contribute to in their professional and personal lives. With the proliferation of information technology, more and more materials, especially language resources, are being made available online and in multimedia formats. As a result, more learners are choosing to use commercial online tools to learn foreign languages. Due to the limited public interest, however, developing such tools for forgotten and endangered languages is not commercially profitable. In such cases, the responsibility of safeguarding endangered languages is often shouldered by researchers and scholars.

This study makes tangible contributions to identifying elements of an enabling online environment for learning oral and mixed languages that are on the verge of being forgotten. With deliberate and purposeful use of information technology, this research has implemented techniques and algorithms to leverage existing online lexical resources of popular and "resource-rich" languages such as English to build parallel (bilingual) online semantization tools for the "resource-poor" languages.

An equally important contribution of this study is the harmonious application of the contemporary Western best practices of language learning in a non-Western educational setting. The resulting design and prototype provide a concrete template for leveraging information technology to find common ground amongst competing interests, needs, philosophies, and practices for passing on traditional knowledge to successive generations. This study will be helpful for the Ismaili community stakeholders in developing strategies and programs to make the ginans more accessible and relevant to the younger generations living in the diaspora. In a broader context, this research benefits other ethnocultural and Indigenous communities with a rich heritage of traditional knowledge by enabling them to make informed decisions on how best to transmit and teach (and thereby safeguard) their oral traditions and languages in modern times.

6. CONCLUSION

We depend on languages to encode our culture, record our knowledge, preserve our history, express our feelings and connect with others. For many of the English-speaking Ismaili youth living in the Western diaspora, the language of the *ginans*, along with the encoded history and teachings, is largely foreign and is increasingly at risk of disappearing. Thus, learning the language and vocabulary of the ginans by successive generations is essential if the tradition and the teachings of ginans are to survive in the Western diaspora.

The purpose of this study was to investigate, design, and validate an online tool for learning and understanding the vocabulary of the ginans. This study used a design-based research (DBR) methodology to prototype an online semantization tool for the oral language of the ginans. Using two iterations of design and evaluation, the initial design of the tool evolved to a functional prototype that was validated by experts and learners.

In the initial design iteration, the theoretical conjectures of language learning were blended with traditional methods and materials of teaching ginans. Based on the literature review, glossing was identified as a crucial best practice for vocabulary semantization for L2 learners. As part of the final design iteration, volunteer respondents from the target group were invited to evaluate the prototype for its efficacy and ease of use. The results showed that the prototype developed in this study was accepted with a high degree of satisfaction in meeting their needs. Moreover, the design elements of the prototype, including content, interface and navigation were tested and found to be highly intuitive and effective by potential learners in the Ismaili community.

Acknowledgement

This article is an abridged version of the prototype development section of my unpublished doctoral dissertation, titled *Tradition and Technology: A Design-Based Prototype of An Online Ginan Semantization Tool* that was supervised by Professor Jay Wilson at the University of Saskatchewan in Saskatoon, Canada.

References

Alibhai, M. (2020, March 5). Tajbibi Abualy Aziz (1926-2019) Part one: A Satpanthi Sita. *The Olduvai Review*. https://theolduvaireview.com/tajbibi-abualy-aziz/

Asani, A. S. (2002). *Ecstasy and enlightenment: The Ismaili devotional literature of South Asia.* London: I.B. Tauris.

Asani, A. S. (2011). From Satpanthi to Ismaili Muslim: The articulation of Ismaili Khoja identity in South Asia. In F. Daftary (Ed.), *A modern history of the Ismailis: Continuity and change in a Muslim community* (pp. 95-128). London: I. B. Tauris Publishers.

Asani, A. S. (2021). *The Ginans: Betwixt Satpanthī Scripture and "Ismaili" Devotional Literature* [Manuscript submitted for publication]. University Department, Harvard University.

Azari, F. (2012). Review of effects of textual glosses on incidental vocabulary learning. *International Journal of Innovative Ideas, 12*(2), 13-24.

Bakker, P. J., & Mous, M. (1994). *Mixed languages. 15 case studies in language intertwining*. Amsterdam: IFOTT.

Barkwell, L. (2017). *A background paper on Michif: An overview of the last 35 years*. Unpublished manuscript. Downloaded from academic.edu. Retrieved February 20, 2021, from https://www.academia.edu/download/59510479/Background_on_Michif_and_Bibliography_final_Barkwell20190604-9486-1cac9a5.pdf

Beheydt, L. (1987). The semantization of vocabulary in foreign language learning. *System*, *15*(1), 55-67.

Holmes, W. (2013). *Level Up! A design-based investigation of a prototype digital game for children who are low-attaining in mathematics*. Doctoral thesis (D.Phil.) University of Oxford. Retrieved February 20, 2021, from https://solo.bodleian.ox.ac.uk.

Jimoyiannis, A., & Komis, V. (2006a). Exploring secondary education teachers' attitudes and beliefs towards ICT adoption in education, Themes in Education, 7(2), 181-204.

Jimoyiannis, A., & Komis, V. (2006b). Examining teachers' beliefs about ICT in education: implications of a teacher preparation programme, Teacher Development, 11(2), 149-173.

Jonassen, D. H. (2000). Computers as mind tools for schools. NJ: Prentice Hall.

Jonassen, D. H. (2003). Computers as mind tools for schools: engaging critical thinking. NJ: Prentice-Hall.

Jurafsky, D., & Martin, J. H. (2019). *Speech & language processing*. Unpublished manuscript. Draft of October 2, 2019. Retrieved from https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Kassam, T. R. (1995). *Songs of wisdom and circles of dance: hymns of the Satpanth Isma'ili Muslim saint, Pir Shams*. SUNY Press.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (2020). *Mixed languages*. Ethnologue: Languages of the world. Retrieved February 21, 2021, from https://www.ethnologue.com/subgroups/mixed-language

Petten, C. (2006). Michif speakers talk language preservation. *Ontario Birchbark*, *5*(4), 1772-1781. Retrieved February 20, 2021, from https://ammsa.com/publications/ontario-birchbark/michif-speakers-talk-language-preservation

Poole, R. E. (2011). *Concordance-based glosses for facilitating semantization and enhancing productive knowledge of academic vocabulary* [Doctoral dissertation, University of Alabama Libraries].

Russell, M., Bebell, D., O'Dwyer, L., & O'Connor, K. (2003). Examining teacher technology use. Implications for preservice and inservice teacher preparation, Journal of Teacher Education, 54(4), 297-310.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329-363.

Vrachnos, E. (2008). Factors determining teachers' beliefs and perceptions of ICT in education. In A. Cartelli & M. Palma (eds.), Encyclopedia of Information Communication Technology (pp. 321-334). Hershey, PA: IGI Global.

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational technology research and development*, *53*(4), 5-23.