Glossed Data in Northern Interior Salish

Anna Stacey

University of British Columbia
Vancouver, Canada
anna.stacey@ubc.ca

Abstract

The Northern Interior subgroup of the Salish language family, spoken in the Pacific Northwest of North America, comprises three languages: Sťát'imcets, nɬeʔkepmxcín, and Secwepemctsín. Each has a small number of first-language (L1) speakers remaining due to the effects of colonization, though language revitalization efforts are ongoing. This work introduces the first compiled and cleaned language datasets in these languages, useable in natural language processing (NLP) projects. This data is in glossed format, with transcriptions in the language, translations into English, and linguistic segmentations and glosses that provide a detailed breakdown of meaning. In order to achieve consistently formatted data within and across each language, extensive data cleaning was conducted. This paper provides the glossed data standards that were developed and recounts the cleaning process. Scripts that help to automate parts of the data preparation processes are included. Finally, this work strives to keep the interconnectedness of language and community as a central consideration.

Keywords: low-resource languages, Salish languages, glossed data, data cleaning, data ethics

1. Introduction

The Salish language family consists of 23 languages traditionally spoken in the Pacific Northwest of North America (Davis, 2019). Due to the effects of colonization including residential schools, only 14 of these languages have first-language (L1) speakers remaining, many of whom are Elders. Three of these languages are Sťát'imcets (ISO: lil), nɬeʔkepmxcín (thp), and Secwepemctsín (shs), which together make up the Northern Interior subfamily. While the small number of L1 speakers of these languages (less than 160 for each; Gessner, Herbert, and Parker, 2022) remain, revitalization efforts are underway, such as university-level language degree programs teaching second-language (L2) speakers.

The publication of glossed data is a priority for fieldworkers working to support the revitalization of these languages. In the Salish context, this typically takes the form of a four-line gloss: transcription in the target language, segmentation of that transcription into individual morphemes, gloss labels assigning a meaning to each morpheme, and translation into English, as a more widely-spoken language. An example is given in (1). This format is used in many texts, such as recorded conversations and stories (see Section 2.1 for examples). Including all this information for each sentence creates rich resources for a variety of audiences: fluent readers can read the transcription line alone, whereas learners can read the transcription and translation lines in tandem, referring to the segmentation and gloss lines for extra detail on how the sentence is composed. Linguists can also make use of the glosses for analysis.

(1) ƛ̓uʔ xʷúy̓ xéʔ ʔiƛ̓ʔiƛ̓m[1]
ƛ̓uʔ xʷúy̓ xéʔ ʔiƛ̓~ʔiƛ̓-m
until PROSP DEM PL~sing-CTR.MID
'They are going to sing.'
(Hannon, Stacey, & Steiner, 2023:125; speaker KBG)

This paper presents the first cleaned, glossed dataset in each of the three Northern Interior Salish languages, useable for NLP work in these languages. Given the urgent language revitalization underway, developments in language technology could be of great use to the community. However, given some of the experiences these communities have had with their language and outsiders, it is essential that those working with the language data do so in a respectful manner that prioritizes transparency with and benefit to the speech community. Therefore, the following sections cover the acquisition of permissions in addition to the gathering and preparation of the datasets. Those

[1] Some notes on glossed examples in this paper:
- I have elected not use the formatting typical of presenting glossed text in papers (e.g., aligning segmented and glossed words) because the format of the data is part of the content of this work.
- All examples are in nɬeʔkepmxcín.
- Non-Leipzig glossed abbreviations used are as follows: CTR.MID = control middle, D/C = determiner/complementizer, PROSP = prospective.
- Speaker initials refer to full names in the Acknowledgements section.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3490–3495
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

interested in working with this data should consult Section 6 for more details.

2. Gathering Glossed Data

2.1 Data Sources

A number of published books of glossed narratives are available in Sťát'imcets (Matthewson, 2005; Alexander, 2016; Edwards, LaRochelle & Mitchell, 2017; Mitchell, 2022; Alexander, 2025) with more under development (Edwards et al., [in prep.]; Jackson, [in prep.]). Most of these are co-published by the Upper Sťát'imc Language, Culture, and Education Society (USLCES) and the Pacific Northwest Languages and Literatures (PNWLL) Press, and are freely accessible in PDF format online. The Sťát'imcets data used in this project comes from one such volume (Alexander, 2016). This book comprises 18 stories told by a single speaker. The data is written in the van Eijk orthography, which is very similar to and indeed based on the Kuipers orthography used for Secwepemctsín (Davis, P.C.). The data was copied from a LaTeX file provided by the publishers into .txt files.

One book of glossed text is published in nɬeʔkepmxcín (Thompson, Egesdal, & Jimmie, 2011). With the necessary permissions for its digitization not able to be secured, it could not be used in this project. Instead, a dataset was compiled by extracting the glossed sentences from all papers covering nɬeʔkepmxcín from the 2023 International Conference on Salish and Neighbouring Languages (ICSNL): Givens (2023); Givens & Hall (2023); Hall (2023); Hannon & Smith (2023); Hannon, Stacey and Steiner (2023); Matthewson (2023); and Reid (2023). Due to the many sources, the data is somewhat varied: it consists of stories, conversations, and sentences from targeted elicitation sessions, and comes from three speakers. This data is written in an Interior Salish version of the North American Phonetic Alphabet (NAPA), which is fairly similar to the IPA. The data was copied from each PDF into .txt files.

In Secwepemctsín, no glossed texts have been published. A small dataset was accessible via an unpublished gloss (Oliver, 2024) of a single narrative found in the language's grammar (Kuipers, 1974). This data is written in the Kuipers orthography and was copied out of a .docx file into a .txt file.

With some minor omissions due to cleaning, the final dataset sizes are as follows:

Language	Sentences	Words
Sťát'imcets	976	8744
nɬeʔkepmxcín	335	2179
Secwepemctsín	101	1621

Table 1: Size of each language dataset

2.2 Data Permissions

Permission to use the Sťát'imcets data in this project was granted by submitting a proposal to and receiving approval from the Upper Sťát'imc Language Authority (USLA). For nɬeʔkepmxcín, permission was granted directly by the speakers of the data. Unfortunately, for Secwepemctsín, neither of these approaches were possible as there is no known language authority in the community and the speaker has passed away.

As these different experiences suggest, there is no one-size-fits-all approach for getting permission for a language data project. Nonetheless, a good faith effort on the part of computational linguists to communicate with the speech community (keeping them in the loop on project progress and results, informing them of potential risks and benefits, etc.) is essential for establishing respectful relationships that everyone benefits from. Getting community input early in a project's development ensures that their priorities, concerns, etc. can be adequately taken into account.

3. Cleaning Glossed Data

3.1 Philosophy

Some cleaning of glossed data can be considered useful to the human reader – e.g., fixing typos and improving the consistency of gloss labels used. When preparing data for use in technological developments (e.g., training machine learning models), however, one must be even more stringent and have very clearly-defined expectations for the formatting of the data. Glossed data is typically 'semi-structured', i.e., only *fairly* rigid in its format, and permits some freedom on the part of the fieldworker (e.g., they may note that there was singing or laughter during a sentence, or have their own practice for noting false starts). It is therefore not trivial to convert glossed data to a truly structured format.

While the goal of using the data in technological developments thus creates a need for consistency, what this consistent format looks like is more open-ended. Knowing that the existing glossed data used here has been created by fieldworkers with expert knowledge in language documentation and in community preference for writing their language, making as few changes as possible to the data seems preferable. This is also in the interest of the efficiency of standardizing the glossed data. In particular, edits to the transcription line were dispreferred over edits to the segmentation and gloss lines, as this represents the language as written and read by its speakers, while the other lines represent an analysis that has been added.

3.2 Data Format

To give a brief overview, each sentence must include the four lines described in Section 1. The transcription line is fairly freeform, consisting of

words in the target language. The segmentation line contains the same words as the transcription line, but with a) morpheme boundaries added and b) normalization processes undone such that the underlying form is captured. Normalization is quite common in the glossed data in these languages: for example, in (2), there is an underlying transitivizer *t* which is often unrealized due to the surrounding sounds. Thus, it is unwritten in the transcription line in such cases, but included in the segmentation line to give the reader a full picture of the underlying morphology. Similarly, the null 3rd-person object marker is explicitly included in the segmentation line.

(2) ʔex xʔe ʔúpis nsqáczeʔ
ʔex xʔe ʔúpi**[-t]-ø**-s n-s-qáczeʔ
IPFV DEM eat-TR-3OBJ-3ERG 1SG.POSS-NMLZ-father
'My father used to eat those.'
(Hannon, Stacey, & Steiner, 2023:137; speaker KBG)

The gloss line has the same words and morphemes as the segmentation line, but instead provides a label for the meaning of each morpheme. Finally, the English translation line is freeform.

With the transcription and translation lines having minimal expectations, most rules are focussed on the segmentation and gloss lines. The formatting for these largely follows the Leipzig Glossing Rules (Comrie, Haspelmath, & Bickel, 2008), with some additional specifications. The morpheme boundaries used in these datasets are:

- hyphen (-) by default
- equals sign (=) for clitics
- tilde (~) for reduplication
- angle brackets (<>) for infixes
- curly brackets ({}) for infixing reduplication[2]

Made more explicit, the data formatting expectations are noted below. Some of the rules are necessary to have a consistent data format that is easily machine-processable:

1. The transcription, segmentation and gloss lines have the same number of words.
2. The segmentation and gloss lines have the same number of morphemes.
3. Each morpheme boundary symbol matches between the segmentation and gloss lines.
4. Infix boundaries (regular or reduplicating) are always used in pairs, with the infix morpheme in between. In the segmentation line this morpheme and boundaries are written within another morpheme, while in the gloss line they immediately follow that morpheme (e.g., segmented *ƛ̓é<ƛ̓>z-m* 'be easy' is glossed as easy<DIM>-CTR.MID).
5. Each morpheme boundary has a morpheme on either side of it, with the exception of infix boundaries in the gloss line, which are either at the end of the word or followed immediately by another morpheme boundary.
6. Brackets indicating underlying morphology are always square brackets and appear only in the segmentation line. Brackets may go around all of or part of a morpheme, but not a morpheme boundary. The bracketed *content* is absent in the transcription line, and present unbracketed in the gloss line. For example, in (2), the transitivizer *t* is absent in the transcription line, bracketed in the segmentation line, and its gloss (TR) appears unbracketed in the gloss line.

The remaining rules are not needed to permit machine processing of the data but nonetheless promote consistency in its content:

7. Out-of-language (OOL) tokens (e.g., English words) are kept in the data but marked with an initial asterisk in the transcription, segmentation, and gloss lines (e.g., *Vancouver* in (3)).

(3) keʔ ks y̓es tək swʔex ʔe ***Vancouver**
keʔ k=s=y̓e=s tə=k=s=wʔex ʔe ***Vancouver**
Q D/C=NMLZ=good=3POSS OBL=D/C=NMLZ=live DET ***Vancouver**
'Is it good to live in Vancouver?
(Matthewson, 2023:302, speaker BP)

8. Primary stress marking never occurs more than once in a word and is consistently marked across the transcription and segmentation lines.
9. Punctuation is absent, aside from:
   - characters that are a part of the language's orthography, e.g., diacritics, in the transcription and segmentation lines
   - characters that are a part of the linguistic analysis, e.g., morpheme boundary symbols, in the segmentation and gloss lines
   - any characters in the translation line

3.3 Example Challenge: Clitics and Word Boundaries

Though many of the rules outlined above may seem to be innately followed by glossed data, a

[2] The curly brackets were added during the cleaning stage – in its original form, the data did not use morpheme boundary symbols to distinguish reduplicating infixes from non-reduplicating infixes. The use of curly brackets means that *all* reduplicating morphology, whether infixing or not, is clearly identified, aligning with the goal of consistency in glossed data formatting.

considerable number of complications were encountered. In such cases, a formatting decision must still be made to ensure consistency, though there may be multiple choices with reasonable justification.

One such example involves Rule 1 (that word counts be consistent). The issue is that where word boundaries lie may not be clear-cut (see Bickel & Zúñiga, 2017). Writing practices may vary because of prosodic considerations, avoidance of very long words, etc. In these languages, this kind of inconsistency appears with clitics, with some typically written attached to their host (as a single word), and others typically written separately. For example, in (4), there are two clitics, nominalizer *s* and 3rd-person possessive *s*, that are written attached to their host in the transcription line, and a third clitic, determiner *ʔe*, that is written as its own word.

(4) kéʔe xʷuy̓ **s**cwuwm**s ʔe** scmém̓iʔt
kéʔe xʷuy̓ **s=**cwuw-m**=s ʔe=**scmém̓iʔt
Q PROSP NMLZ=work-CTR.MID=3POSS DET=children
'Should children work?'
(Matthewson, 2023:302, speaker KBG)

To follow Rule 1, either the transcription line could be modified to always have clitics attached (in keeping with the segmentation line), or the segmentation line could be modified to sometimes write clitics as their own words (in keeping with the transcription line). Guided by the principle of dispreferring modifications to the transcription line, the latter approach was selected. However, comments from the fieldworkers have argued that the approach in the segmentation line is more consistent and linguistically-grounded. Therefore, this decision of how to make the data abide by Rule 1 is still under discussion and may be reversed in future.

Regardless of which choice is ultimately settled upon, this provides an illustrative example of the ways glossed data deviates from seemingly straightforward formatting expectations. In order to modify the data to follow these expectations, we must choose between approaches that will each have their own drawbacks.

3.4 Methodology

Ideally, the cleaning process would be as automated as possible, performed by scripts for data processing. However, since one of the goals of cleaning is to output machine-processable data, the pre-cleaned data is naturally not well-suited to automated cleaning.

In order to have *some* of the cleaning be automated (e.g., fairly simple changes like removing sentence-final punctuation), I strived to first conduct as little manual cleaning as possible to make the data machine-processable. This was a large undertaking, and a Python script[3] was used to help flag issues in the data. As the more basic issues are manually corrected, the data becomes more machine-processable, allowing the script to flag more complicated issues. For example, the script can first help flag sentences that do not have exactly four lines. After those are resolved, it is able to identify which lines are the transcription, segmentation and gloss lines and flag sentences that do not have the same number of words between these.

Some flagged issues were systemic practices, such as the challenge involving clitics discussed in the previous section. Many others were simply small errors, such as a word with three morphemes in the segmentation line having only two in the gloss line. To determine the correct resolution of such issues, I used (in descending order of preference) my own language knowledge, similar sentences elsewhere in the data, consultation with the fieldworker, or removal of the sentence from the dataset altogether. A careful record of manual changes was kept.

Once Rules 1–7 were met and the data was thus easily parsable by code, other scripts were run for further cleaning. Some of these are language-generic, whereas others attend to the specifics of each language.

At this stage, the data is in a consistent format. However, consistency in the glossing practices within this format (e.g., what gloss label does each morpheme get? which morphemes are segmented as clitics versus affixes?) is an outstanding goal. An additional script was thus used to generate a summary of these glossing practices, making it easy to notice, e.g., that a morpheme is being glossed as 'happy' sometimes and 'glad' other times. This tool was used predominantly for the nɬeʔkepmxcín data, where the variety of fieldworkers involved made this kind of inconsistency particularly common. The summary is also of interest to fieldworkers – sharing it with the Secwepemctsín fieldworker allowed us to collaboratively make some labels more consistent using his language knowledge. Thankfully, many such unifications of glossing labels could be made automatically since the data is easily handled by scripts at this stage.

Finally, some effort was made to improve the consistency of glossing practices *across* the three languages, so that crosslinguistic developments are possible. Firstly, orthographic conversion functionality is included in the language-specific cleaning scripts, so that all the data can appear in one uniform orthography. Secondly, there were a few clear instances of different gloss abbreviations being used for the same gloss

[3] https://github.com/anna-stacey/NorthernInteriorGlossing

labels across the languages, which were unified so that only one abbreviation is used.

4. Conclusion

This works presents the first machine-readable glossed datasets in the three Northern Interior Salish languages: St'át'imcets, nɬeʔkepmxcín, and Secwepemctsín. It also introduces a set of standards for glossed data formatting used to ensure consistency in this data, as well as a description of the process undertaken to clean data to this standard. Finally, scripts are shared that can automate aspects of the preparation of future glossed data into this format, including flagging issues to clean, the data cleaning itself, and summarizing gloss labels to flag areas where consistency can be improved. While the data itself can be used to create technological resources for the revitalization of these languages, the data preparation process (both the description and scripts themselves) can be used to create more datasets of this kind for these and other languages.

5. Acknowledgments

I would like to thank the Northern Interior Salish speakers whose sentences are the heart of this project:

- St'át'imcets: *Qwa7yán'ak* (Carl Alexander)
- nɬeʔkepmxcín: Bev Phillips, *c̓úʔsinek* (Marty Aspinall), and *kʷaɬtəzétkʷu* (Bernice Garcia)
- Secwepemctsín: Seymour Petel

Bernice wishes it to be acknowledged that she is a Kamloops Indian Residential School speaker, who is re-learning her language. She introduces herself thus: *ʔes ʔúməcms kʷaɬtəzétkʷu təw ɬe c̓əɬétkʷu wéʔe ncitxʷ. ʌ̓uʔ wéʔec ʔex netíyxs scwew̓xmx, ʌ̓uʔ tékm xéʔe ne nɬeʔképmx e tmíxʷs*, 'My traditional name is kʷaɬtəzétkʷu, my home is in Coldwater of 'Nicola' of Nlaka'pamux lands.'

Many thanks also to the Upper St'át'imc Language Authority (USLA) for their support of this work.

For their guidance on this project, I would like to thank my thesis committee: Henry Davis, Garrett Nicolai, and supervisor Miikka Silfverberg. As well, my gratitude goes to the nɬab (nɬeʔkepmxcín lab) at UBC, especially Lisa Matthewson and Bruce Oliver. This work was funded by a scholarship (CGS-M) from the Social Sciences and Humanities Research Council (SSHRC) of Canada.

6. Ethical Considerations

As alluded to in the main paper, any public release of the data involved in this project (and any expectations accompanying its use) are dependent on speaker and/or community consent and wishes. Because these can evolve with time, they are described in the linked repository to make updates straightforward. Thus, please see the repository page for the most up-to-date statement on data permissions for those interested in working with these datasets.

7. Bibliographical References

Alexander, C. (2016). Sqwéqwel' múta7 sptakwlh: St̓át̓imcets Narratives by Qwa7yán'ak (Carl Alexander). Recorded, transcribed, translated and edited by Elliott Callahan, Henry Davis, John Lyon & Lisa Matthewson. Vancouver and Lillooet, Canada: University of British Columbia Occasional Papers in Linguistics and the Upper St'át'imc Language, Culture and Education Society (USLCES).

Alexander, C. (2025). Sqwéqwel' múta7 sptakwlh: St̓át̓imcets Narratives by Qwa7yán'ak (Carl Alexander) Volume II. Transcribed, translated and edited by Matt Andrew, Henry Davis & John Lyon. Vancouver and Lillooet, Canada: Pacific Northwest Languages and Literatures Press (PNWLL) and the Upper St'át'imc Language, Culture and Education Society (USLCES).

Bickel, B. & Zúñiga, F. (2017). The 'Word' in Polysynthetic Languages: Phonological and Syntactic Challenges. The Oxford Handbook of Polysynthesis,. Oxford, UK: Oxford University Press, pp. 158-185.

Comrie, B., Haspelmath, M. & Bickel, B. (2008). The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Max Planck Institute for Evolutional Anthropology.

Davis, H. (2019). Salish Languages. The Routledge Handbook of North American Languages.

Edwards, A., Edwards F., Hickson & Tom. In preparation.

Edwards, B., LaRochelle, M., & Mitchell, S. (2017). Sqwéqwel's Nelh Skelkekla7lhkálha (Tales of Our Elders). Recorded, transcribed, translated and edited by Davis H., Lyon J., van Eijk, J., & Whitley, R. Vancouver and Lillooet, Canada: University of British Columbia Occasional Papers in Linguistics and the Upper St'át'imc Language, Culture and Education Society (USLCES).

Egesdal, S. M., Thompson, T. M., & Jimmie, M. N. (2011). nɬeʔképmxcín: Thompson River Salish Speech. Bellingham, Washington: Whatcom Museum Publications.

Gessner, S., Herbert, T., & Parker, A. (2022). Report on the Status of B.C. First Nations Languages. Online (https://fpcc.ca/resource/language-status-report-2022/): First Peoples' Cultural Council.

Givens, K. (2023). Degree Comparison in Nɬeʔkepmxcín. In Reisinger, D. K. E., Griffin, L., Mellesmoen, G., Nederveen, S., Schillo, J., &

Trotter, B. (eds.), Proceedings of the 58th International Conference on Salish and Neighbouring Languages (ICSNL). Vancouver, Canada: University of British Columbia Working Papers in Linguistics, pp. 66-77

Givens, K., & Hall, B. (2023). The Moon and the Birchbark Canoe (ɬ máʔxetn pe ɬ qwɬinéwɬ). In Reisinger, D. K. E., Griffin, L., Mellesmoen, G., Nederveen, S., Schillo, J., & Trotter, B. (eds.), Proceedings of the 58th International Conference on Salish and Neighbouring Languages (ICSNL). Vancouver, Canada: University of British Columbia Working Papers in Linguistics, pp. 78-84.

Hall, B. (2023). A Brief Look at Infinitives in Nɬeʔkepmxcín. In Reisinger, D. K. E., Griffin, L., Mellesmoen, G., Nederveen, S., Schillo, J., & Trotter, B. (eds.), Proceedings of the 58th International Conference on Salish and Neighbouring Languages (ICSNL). Vancouver, Canada: University of British Columbia Working Papers in Linguistics, pp. 85-93.

Hannon, E., & Smith, C. (2023). A Brief Comparison of Two Nɬeʔkepmxcín Evidentials. In Reisinger, D. K. E., Griffin, L., Mellesmoen, G., Nederveen, S., Schillo, J., & Trotter, B. (eds.), Proceedings of the 58th International Conference on Salish and Neighbouring Languages (ICSNL). Vancouver, Canada: University of British Columbia Working Papers in Linguistics, pp. 94-116.

Hannon, E., Stacey, A., & Steiner, R. (2023). Glossed Conversational Data in Nɬeʔkepmxcín. In Reisinger, D. K. E., Griffin, L., Mellesmoen, G., Nederveen, S., Schillo, J., & Trotter, B. (eds.), Proceedings of the 58th International Conference on Salish and Neighbouring Languages (ICSNL). Vancouver, Canada: University of British Columbia Working Papers in Linguistics, pp. 117-158.

Jackson, S. (in preparation).

Kuipers, A. H. (1974). The Shuswap Language. The Hague, The Netherlands.

Matthewson, L. (2005). When I Was Small – I Wan Kwikws. Vancouver, Canada: UBC Press.

Matthewson, L. (2023). Two Types of Polar Question in Nɬeʔkepmxcín. In Reisinger, D. K. E., Griffin, L., Mellesmoen, G., Nederveen, S., Schillo, J., & Trotter, B. (eds.), Proceedings of the 58th International Conference on Salish and Neighbouring Languages (ICSNL). Vancouver, Canada: University of British Columbia Working Papers in Linguistics, pp. 291-332.

Mitchell, S. (2022). Wa7 Sqwéqwel' sSam: St'át'imcets Stories from Sam Mitchell. Volume I of The Bouchard Tapes, recorded by Bouchard, R., transcribed, translated and edited by Lyon, J., & Davis, H., with an introduction by Davis, H. Vancouver and Lillooet, Canada: Pacific Northwest Languages and Literatures Press (PNWLL) and the Upper St'át'imc Language, Culture and Education Society (USLCES).

Oliver, B. (2024). Gloss of The Gambler's Son and Red-Cap by Seymour Petel. Unpublished.

Reid, D. (2023). An Acoustic Analysis of Aspiration in Nɬeʔkepmxcín. In Reisinger, D. K. E., Griffin, L., Mellesmoen, G., Nederveen, S., Schillo, J., & Trotter, B. (eds.), Proceedings of the 58th International Conference on Salish and Neighbouring Languages (ICSNL). Vancouver, Canada: University of British Columbia Working Papers in Linguistics, pp. 368-383.