

SELF & FEIL: Emotion Lexicons for Finnish

Emily S. Öhman¹

¹Waseda University, Nishiwaseda 1-6-1, Shinjuku, Tokyo, Japan

Abstract

This paper introduces a Sentiment and Emotion Lexicon for Finnish (SELF) and a Finnish Emotion Intensity Lexicon (FEIL). Sentiment analysis and emotion detection require annotated data regardless of the chosen approach, but most existing resources are for the English language. To overcome this, the SELF and FEIL lexicons use projected annotations from existing resources with carefully edited translations and domain adaptations. In this paper the creation process and translation issues are explained in detail to allow others to create similar lexicons for other languages. The usefulness of SELF and FEIL are demonstrated via several interdisciplinary affect-related projects. To our best knowledge, this is the first comprehensive sentiment and emotion lexicon for Finnish.

Keywords

sentiment analysis, lexicon creation, emotion detection, lexicon validation

1. Introduction

There are three main approaches to sentiment analysis and emotion detection: machine learning, lexicon-based, and hybrid methods that combine the first two approaches. The one thing all of these approaches have in common is the need for annotated data. This annotated data can consist of labeled datasets for training and testing classifiers, or lexicons to be used with different types of word-matching approaches. Usually, the annotation process requires human annotators in an iterative validation process to confirm the validity of each and every label and is therefore labor-intensive and expensive [1].

The benefit of lexicon-based methods is in the re-usability of them for multiple domains, and therefore also the lower cost as there is no need to re-annotate the lexicons for each project. Although both machine learning datasets and lexicons are somewhat context-dependent, lexicons are slightly less so [2], and they are typically easier and cheaper to edit for a new domain than machine learning datasets. Emotion and sentiment lexicons can be used “as is” in purely lexicon-based approaches or as a feature extraction tool for data-driven classifiers, increasing their context-sensitivity (see e.g. Schmidt et al. [3]). However, most emotion lexicons have been created for the English language and there are very few quality emotion lexicons for other languages, including Finnish.

This paper introduces the SELF (Sentiment and Emotion Lexicon for Finnish) and FEIL (Finnish Emotion Intensity Lexicon) lexicons. These are to the author’s best knowledge the


The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.

✉ ohman@waseda.jp (E. S. Öhman)

ORCID 0000-0003-1363-7361 (E. S. Öhman)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

first and only emotion and sentiment dictionaries that have been made widely available for Finnish and for which translations have been manually verified by independent annotators. In this paper the focus is on the manual creation of lexicons, specifically the augmentation of existing ones that were originally created manually and later automatically translated. The goal of this pilot project is to help other researchers create similar lexicons for their languages, give step-by-step directions on how to increase domain-specificity for their projects, and to demonstrate the usability of these lexicons in interdisciplinary research projects. In the next section, lexicon-based sentiment analysis and lexicon creation in general is explained, after which an overview of the lexicon creation process is presented. Section 4 illustrates how these lexicons can be used in interdisciplinary projects and serves as a real-world evaluation of the quality of the lexicons followed by a concluding discussion in section 5.

The SELF (Sentiment and Emotion Lexicon for Finnish) and FEIL (Finnish Emotion Intensity Lexicon) lexicons are available on GitHub¹. Domain- and project-specific versions of the lexicons will be made available after the publication of the respective papers they were used in.

2. Background

Emotion lexicons can be created automatically or manually, but both approaches often, at least partially, utilize dictionaries that list emotion categories in some way. For automatic creation no manual verification is applied. Instead, other features are used to determine the sentiment or emotion. For example, Kimura and Katsurai [4] automatically created an emoji-lexicon by leveraging cosine similarity values and co-occurrence from WordNet Affect [5]. The manual creation of lexicons involves using human annotators and asking them to determine how specific words relate to different sentiments and/or emotions. This process needs to be repeated and each data point annotated by multiple independent annotators as humans rarely agree with each other on any labeling task, but even less so with emotion labeling [6, 7]. The agreement between annotators is calculated using inter-rater agreement scores such as Krippendorff's α .

One of the most widely-used and well-known emotion lexicons is the NRC (National Research Council Canada) Emotion Lexicon [8, 9]. The first version of the NRC Emotion Lexicon (a.k.a. EmoLex) [8] was created using Mechanical Turk, a platform for crowd-sourcing annotations from humans. The lexicon was later significantly augmented to its current 14,182 lexical items also using Mechanical Turk [9]. For the first version, the annotators were asked to annotate for emotions **evoked** by the words, and for the later version, they were asked to annotate for emotions **associated** with the words as the developers of the lexicon discovered this led to better agreement scores. Although, it too was created based on English data, the English words have been translated using Google Translate and there are now 14,182 entries in total translated into 104 languages. Even though very few of these translations have been manually verified, some of the multilingual versions of the NRC Emotion Lexicon have been tested previously on at least Spanish, Portuguese, and Arabic [10, 11] to evaluate emotion preservation in translation, and the translations as well as the original English lexicon are also a built-in part of the *syuzhet* package for R [12].

¹<https://github.com/Helsinki-NLP/SELF-FEIL>

The annotations for the NRC Emotion Intensity Lexicon [13] were compiled using best-worst scaling (BWS). BWS is an efficient method to collect massive amounts of scaled annotations and has been proven to beat rating scales and other methods in both quality and cost² [14]. The main difference between the intensity lexicon and the emotion lexicon is that the intensity lexicon does not include sentiments (positive, negative), and that instead of a Boolean option for each emotion (0 for not associated with a particular emotion, 1 for being associated with a particular emotion) the intensity lexicon gives a score as to how intense the emotion associated with a word is (a score between 0 and 1, with 0 for no association and therefore no intensity, to 1 for the highest intensity).

There are many considerations that need to be taken into account when using resources created for one language with another language. Emotion words are closely linked with culture and emotions and feelings are expressed quite differently in different cultures and languages. Although research shows the universality of affect categories [15, 16], Mohammad et al. [11] list several error types when translating emotion words. These are, e.g., mistranslation, cultural differences and different word connotations, as well as different sense distributions. When projecting annotations, particularly emotion annotations, from one language to another, it is important to consider all of these aspects.

3. Lexicon Creation and Description

The automatic Finnish translations of the NRC Lexicons were re-translated with the most current version of Google Translate. First, within these translations, duplicates were marked and all translations were carefully evaluated to match the English word's meaning as well as the associated emotions and in the case of the FEIL lexicon, intensity too. The problem with Google Translate is that it chooses the most common translation, especially when translating single words out of context rather than words in context. This process means that with an emotion dictionary with many synonyms and near-synonyms, Google Translate provides the same common word for all as the default translation.

Naturally, this leads to issues with representativity as the Finnish lexicon, if left as is, would only represent the most common words in Finnish, disregarding any synonyms or near-synonyms leading to a loss of nuances and insufficient coverage of emotion words. Hence, those duplicates were carefully examined in order to find alternative translations that matched the original word better in terms of meaning and connotation, both by human experts and synonym dictionaries and thesauruses (see table 1 for an example). The original lexicon also included several alternative spelling options, which were translated as the same word. Such duplicates (e.g. tumor/tumour) were also removed. Another type of over-generalization was also discovered: connotative-generalization, i.e. the meaning of the original word was over-generalized so that the translation had lost all original connotations and many emotion associations (see the example of 'emaciated' in table 2 in the Appendix).

The reverse was also found with some instances where English has a more fine-grained separation of some concepts where Finnish does not. E.g. Finnish does not differentiate between *poison-venom-toxin*. It is all the same word *myrkkyy*. Some English words can be both nouns and

²Cost because other methods take exponentially more time to collect the same amount of annotations.

Table 1
Example of lexicon editing: the case of *hurskas*

Original English	Google Translate	Edited Translation
pious	hurskas	hurskas
devout	hurskas	harras
saintly	hurskas	pyhimysmäinen
godly	hurskas	jumalinen

verbs, and therefore, if not specified it is impossible to tell which one is being evaluated³ or if both are. One such word was *rape*. In this case the verb form was added with the same emotion and intensity associations as the noun form. There were also cases of clear mistranslations, some of which occurred with polysemic and homonymous words where the Finnish translation was clearly not the one meant based on the associated emotions and/or intensities. For example *birch* had been translated as ‘koivu’, a birch tree, but the negative associations suggested that what had been meant was the act of flogging. In many cases, the least contentious solution was to remove ambiguous entries; a recommendation if there is only one annotator. With multiple annotators, agreement scores can be calculated and contested annotations can be solved by various means, including, but not limited to removal. See table 2 for an overview of common corrections.

Table 2
Examples of fixes to the SELF and FEIL lexicons

Original English	Automatic Translation	Corrected Form(s)
birch	koivu	piiskata
emaciated	laihtunut (having lost weight)	riutunut
rabble	lauma (herd, flock)	roskaväki
corroborate		-entry removed-
strengthen	vahvistaa	-entry kept-
cede		-entry kept-
relinquish	luovuttaa (to give up)	-entry removed-
rape	raiskaus (N)	-entry kept-
	raiskata (V)	-entry added-

For *emaciated*, the automatic translation was too general. The corrected form is less common but almost identical in meaning and connotation to the original English. As the term *rabble* has such negative connotations in English, the automatic translation was much too neutral and was changed to a word with similar connotations. As for *corroborate* the issue was with there not existing a one-word translation for *corroborate* in Finnish and the meaning of the automatically translated word being much closer to *strengthen*. The meaning of the original English words *cede*, *relinquish* are somewhat synonymous so it makes sense to have both words in the English lexicon with nearly identical emotional intensities, but there is only one one-word translation

³In the original task for English, the annotators were given a test for each word to check that they understood the meaning of the word as intended. This information is not available in the published dictionary.

for Finnish so the duplicate was removed. Another example is *furor* which in English had anger intensities of 0.9 and in Finnish had been translated into *villitys - fad*, but the best translation was already paired with *rage* and as no suitable alternative translation was found, the entry was deleted.

The best translation is not always necessarily the best choice for a lexicon entry. If one translated word truly is the best translation for several English words, an evaluation of the best match needs to be made. If possible, the duplicate is not removed, but altered to still match the English word, but using a different word, even if not as accurately translated to increase the diversity of the lexicon while remaining true to the original emotion.

Cultural differences were also evident in the lexicon. North-Americans are on average more religious than the Nordic people [17], which means that many religious words seem to have much more positive connotations in the US and Canada than they do in Northern Europe, and Scandinavia in particular (see the example of *hurskas* in table 1). This was most evident with religious words, but other cultural connotations seemed to be present as well. In these cases though, unless the association was glaringly wrong, they were kept as is: It is a slippery slope to try and push one's own judgment onto the intensity scores and emotion-word associations. Everyone is inherently biased and relies on their own experience when making judgments. Only when enough people agree on a judgment, can it be seen as culturally representative. Since in this case only one person did most of the corrections, words were rather deleted than letting a lone annotator's subjective judgment overly influence the lexicon⁴.

The final distribution of the SELF and FEIL lexicons can be seen in table 3. All the adjustments mentioned in this section resulted in an overall reduction in lexicon size by 12.2% from 14182 word-emotion association pairs to 12448 entries in Finnish. The final distribution of the FEIL lexicon with all the adjustments mentioned in this section resulted in an overall reduction in lexicon size by 10.5% from 8149 intensities to 7291 entries for Finnish.

Table 3

Distribution of Emotions in SELF & FEIL after corrections.

positive	negative	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	Lexicon
2117	2938	1084	783	919	1309	636	1059	473	1130	SELF
		1304	805	946	1554	1145	183		1354	FEIL

The distribution of the types of corrections is presented in table 4. It is easier to detect sense and specificity mistranslations in the intensity lexicon when the intensities of the associated emotions are clear. This is likely the reason for the higher relative occurrence of such corrections in the intensity lexicon. However, these types of errors are also the hardest to detect overall, and therefore, this is the category that is most likely to increase the most when the lexicon is updated and revised further.

⁴Newer versions of the lexicon and domain adaptations have been evaluated by four annotators including additions and intensity changes.

Table 4

Percentage of different types of corrections in SELF and FEIL

% of corrections	SELF	FEIL	Ex.
Duplicate removal	46.1%	41.3%	identical target words with no alternative translation
Duplicate replacement	36.8%	33.4%	identical target words with alternative translation
Mistranslation -sense	2.0%	4.9%	<i>birch</i> to birch tree instead of flogging
Mistranslation -specificity	3.7%	6.3%	<i>emaciated</i> to <i>laihtunut</i> instead of <i>riutunut</i>
Grammatical difference	1.5%	1.2%	part-of-speech difference
Cultural difference	0.9%	1.7%	overly positive connotations of religious words
Other, undefined	9.0%	11.2%	

4. Lexicon Evaluation

SELF was first evaluated against the original EmoLex by using the R package *syuzhet* [12] on the Finnish novel *Rautatie* by Juhani Aho. The results can be seen in figure 1 where on the left is the original EmoLex in Finnish and on the right SELF. The two plots show similar overall patterns, but are decidedly different in detail, with SELF providing the more accurate plot when evaluated by a literary expert on Aho [18].

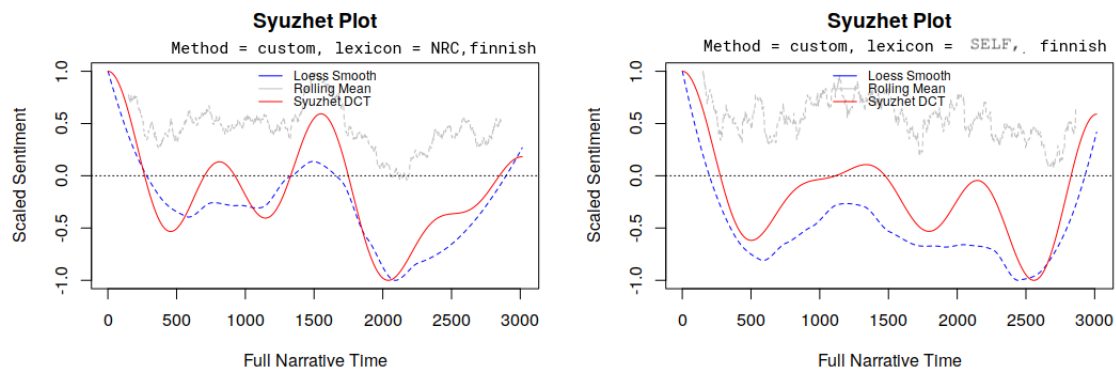


Figure 1: *syuzhet* results for *Rautatie* using different lexicons

The emotion word distributions (see table 5) reveal that in general the SELF lexicon finds more emotion words in the novel despite having a reduction in lexical items of 10%, but that *anger*, *disgust*, and *sadness*, and therefore also *negative* were over-represented in the original lexicon, likely by duplicates.

Both SELF and FEIL were used on a project where Finnish political party manifestos between the years of 1945-2019 were analyzed for patterns of emotion [19]. In this case, FEIL proved to be the more useful lexicon, as it allowed for us to examine more detailed expressions of emotions. The output using the lexicon-based approach was evaluated against manual annotations as well as by checking for statistical significance of the results. The validity of this evaluation process has been discussed in detail in Ohman [20]. The most interesting finding was perhaps that

Table 5
Comparison of lexicon matches for *Rautatie* by Juhani Aho

positive	negative	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	
1153	1258	299	482	303	506	345	368	248	584	SELF
828	1363	481	299	381	433	241	451	162	321	NRC

populist parties use the same amount of emotion words as other parties, but that the intensity of the emotion words they use is significantly higher.

Finally, FEIL has also successfully been used in conjunction with structural topic modeling (STM) to measure change in attitudes towards specific entities during the COVID-19 pandemic [21]. In this project we also made sure that the most common words in the massive dataset we collected were included in FEIL. This resulted in some additions to the lexicon, the emotional intensity of which were carefully evaluated by 4 humans. This COVID-19 enhanced version of FEIL will also be made public after the publication of the original article.

5. Concluding Discussion

The SELF and FEIL lexicons have proved to be valuable augmentations of the NRC Emotion and Intensity Lexicons. The lexicons have shown that they can match more words in real-world texts than the original translations of EmoLex/NRC lexicon despite being approximately 10% smaller in size. The problems with the use of the NRC lexicons' automatically translated versions do not seem to be caused by any issues with the original annotations (also attested by the thousands of projects that have successfully used them for English), nor with the annotation projection as such, but mostly by Google Translate's algorithms which are not optimal for translating single words out of context. With these issues fixed, SELF and FEIL have proven to be useful in several interdisciplinary tasks and likely many more in the future.

Both SELF and FEIL have been evaluated against the original automatically translated NRC EmoLex as well as had their results validated against human annotations of texts. In all cases SELF and FEIL produced reliable, real-world congruent results that closely aligned with human perceptions of emotions expressed in those texts. It is the recommendation of the author to use FEIL over SELF whenever methodologically possible as it showed closely matching results to human impressions of longer texts [20] and if simply comparing emotion presence, is more real-world accurate than SELF. It is also recommended to confirm SELF and FEIL results with statistical significance testing or similar. If the method makes it possible to validate against human annotations, it is of course always prudent to do so.

Both lexicons work especially well on longer texts with semi-informal registers. To increase the domain-specific coverage of the lexicon, it is also recommended to compile a list of the most common tokens in the dataset that is being examined and check to see that all words that are associated with emotions are added to the lexicon. Their emotion associations should be confirmed by at least 3 annotators to reduce the risk of a single annotator's biases influencing the results.

As for future work, it would be interesting to simultaneously use multiple emotion lexicons to

examine if they could be used together to automate some of the manual verification required for translated data. The lexicon could also be used in conjunction with machine learning methods, particularly with literary works and similar, where manually annotating large texts to create fine-grained training and testing data is usually infeasible.

Acknowledgments

I would like to thank Dr. Saif Mohammad, the creator of the NRC emotion lexicons, for taking the time to discuss the SELF and FEIL lexicons with me, and for his helpful suggestions in improving the draft version of this paper.

I would also like to thank the anonymous reviewers for their helpful comments that helped improve this paper.

This work was in part supported by JSPS KAKENHI Grant Number 22K18154.

References

- [1] E. Öhman, Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task, in: *Digital Humanities in the Nordic Countries 2020*, CEUR Workshop Proceedings, 2020.
- [2] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, *Computational linguistics* 37 (2011) 267–307.
- [3] T. Schmidt, K. Dennerlein, C. Wolff, Using deep learning for emotion analysis of 18th and 19th century german plays (2021).
- [4] M. Kimura, M. Katsurai, Automatic construction of an emoji sentiment lexicon, in: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, ASONAM '17*, Association for Computing Machinery, 2017, p. 1033–1036.
- [5] C. Strapparava, A. Valitutti, et al., Wordnet affect: an affective extension of wordnet., in: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, volume 4, 2004, p. 40.
- [6] P. S. Bayerl, K. I. Paul, What determines inter-coder agreement in manual annotations? A meta-analytic investigation, *Computational Linguistics* 37 (2011) 699–725.
- [7] S. Mohammad, A practical guide to sentiment annotation: Challenges and solutions., in: *WASSA@ NAACL-HLT*, 2016, pp. 174–179.
- [8] S. Mohammad, P. Turney, Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon, in: *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, 2010, pp. 26–34.
- [9] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, *Computational Intelligence* 29 (2013) 436–465.
- [10] M. Salameh, S. Mohammad, S. Kiritchenko, Sentiment after translation: A case-study on Arabic social media posts, in: *Proceedings of the 2015 conference of the North American*

chapter of the association for computational linguistics: Human language technologies, 2015, pp. 767–777.

- [11] S. M. Mohammad, M. Salameh, S. Kiritchenko, How translation alters sentiment, *Journal of Artificial Intelligence Research* 55 (2016) 95–130.
- [12] M. Jockers, *Syuzhet 1.0.4 now on CRAN*, Matthew L. Jockers (2017).
- [13] S. M. Mohammad, Word affect intensities, in: *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, 2018.
- [14] S. Kiritchenko, S. Mohammad, Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, 2017, pp. 465–470.
- [15] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, D. Keltner, The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures, *Nature human behaviour* 3 (2019) 369 – 382.
- [16] K. R. Scherer, H. G. Wallbott, Evidence for universality and cultural variation of differential emotion response patterning., *Journal of personality and social psychology* 66 2 (1994) 310–28.
- [17] V. Skirbekk, P. Connor, M. Stonawski, C. P. Hackett, *The future of world religions: Population growth projections, 2010-2050*, Pew Research Center, 2015.
- [18] E. Öhman, R. Rossi, *Affect and Emotions in Finnish Literature: Combining Qualitative and Quantitative Approaches*, forthcoming).
- [19] J. Koljonen, E. Öhman, P. Ahonen, M. Mattila, Strategic sentiments and emotions in post-Second World War party manifestos in Finland, *Journal of Computational Social Science* (2022-forthcoming).
- [20] E. Ohman, The validity of lexicon-based emotion analysis in interdisciplinary research, in: *Proceedings of the ICON 2021 workshop NLP4DH*, 2021.
- [21] J. Paakkonen, E. Ohman, S.-M. Laaksonen, Unconventional communicators in the covid-19 crisis. (Forthcoming).