The Loreen Corpus

User Manual

Compilation

Introduction

This corpus was compiled by Xander Johnson, a graduate student in Linguistics at the University of Kentucky. The corpus was completed on December 7, 2024, and consists of eighty-two text files. Each text file contains the lyrics of one of the forty-one songs released by Swedish artist Loreen since the start of her career. Forty-one text files contain only the text and metadata; the remaining forty-one have been tagged for part of speech and lemma with SpaCy tags in TagAnt (Anthony, 2024). Both English and Swedish lyrics have been tagged using their respective tag sets. In total, forty-one texts are represented across the eighty-two text files. Tagged files are located within the subfolder 'tagged' and denoted with the suffix '_tagged'. The corpus and its metadata are compatible with AntConc version 4.3.1.[1]

Project Intent

This corpus was created to examine the lyrics of Swedish artist Loreen diachronically across her career, allowing users to explore her use of Swedish and English in her songs, as well as draw comparisons between her music across different time periods. The corpus can explore questions of structure, topic, and themes present in Loreen's lyrics. It also allows users to

[1] The metadata for songwriters, tagged "Writers" is not compatible with AntConc, but can be recovered by searching for appropriate text strings in the files.

compare by songwriter, year of release, mode of release (album vs. single), and the album (if applicable) to draw comparisons across several points of her musical career.

Text Sampling

I elected to include all songs listed on Loreen's Spotify page as of December 7, 2024, where she is listed as the primary artist. She need not be the only artist, but any songs where she was listed as a featured artist were excluded, as these do not necessarily represent the music that she chooses to create under her own brand. This includes any covers of previously existing songs; as long as she released the track under her own name and brand, it is treated as a part of the repertoire that she claims for artistic purposes. For songs or albums that received multiple releases or versions, only the first album release was included, if existent. For those tracks not released on albums, the first released version was the one included in the corpus; however, there is not significant variation between the lyrics of these versions when compared to the album releases. This was done to avoid heavily weighting the corpus toward tracks with multiple releases and remixes. The texts were collected from azlyrics.com, initially with BootCat, but were manually cleaned by the compiler. Minor corrections to certain lyrics were made and noted within a corrections tag. For the purpose of making the corpus more diachronically comparable, release years were binned into three groups: 2012-2013, 2015-2017, and 2019-2024.

Metadata

The following metadata were collected for each track and stored in xml-style tags:

- Primary Artist – <artist></artist> – (currently only Loreen, but was included to make the corpus expandable in the future)

- Track Title – <title></title>
- Album – <album></album>
- Release Year – <release-year></release-year>
- Bin – <bin></bin>
- Language – <language></language>
- Writers – <writers><names><name id=”1”></name></names></writers>
- Word Count – <words></words>
- Lyrics Source – <source></source>
- Corrections – <corrections></corrections>

The corpus and all metadata, excluding writers, are compatible with AntConc Version 4.3.1 (Anthony, 2024) . Metadata can be loaded into AntConc from the metadata.csv file.

Corpus Profile

Word Counts

Total Word Count	9,621 words
English Word Count	8,425 words
Swedish Word Count	1,196 words

Word Count by Release Year

2012	3,146 words
2013	228 words
2015	648 words
2017	2,728 words
2019	264 words
2020	917 words
2021	236 words
2022	319 words
2023	536 words
2024	599 words

Word Count by Release Year (Binned)

2012-2013	3,374 words
2015-2017	3,376 words
2019-2024	2,871 words

Word Count by Album

“Heal”	3,146 words
“Nude”	786 words
“Ride”	1,739 words
“Så mycket bättre 2020”	791 words
singles	3,159 words

Additional Information

The tagged and untagged corpora each contain the same number of words and the same metadata. A full list of urls from which texts were sourced may be found in the appendix. All texts are song lyrics, and as such, not written in prose. As such, the tagged corpus was kept in horizontal format to retain line breaks. It is a spoken, or more accurately, sung, corpus. All text falls into the broad genre of European Pop music.

References

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc

Anthony, L. (2024). TagAnt (Version 2.1.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/TagAnt