Meertens
Online Reports

Meertens Instituut, Amsterdam, Netherlands
www.meertens.knaw.nl

ISSN 2352-2135

The Meertens Tune Collections: The Annotated Corpus (MTC-ANN) Versions 1.1. and 2.0.1

Peter van Kranenburg, Berit Janssen, Anja Volk

2016-1

The Meertens Tune Collections: The Annotated Corpus (MTC-ANN) Versions 1.1 and 2.0.1

Peter van Kranenburg,[1] Berit Janssen,[1] and Anja Volk[2]
[1]Meertens Institute, Amsterdam, [2]Utrecht University

Contents

1 Introduction

This report describes versions 1.1 and 2.0.1 of the Annotated Corpus (MTC-ANN), which are part of the Meertens Tune Collections (Van Kranenburg et al., 2014). The Meertens Tune Collections (MTC) consist of a number of data sets containing digitized Dutch folk song data from the rich archives of the Meertens Institute (Amsterdam). The metadata are maintained in the Dutch Song Database, which is online accessible at `http://www.liederenbank.nl`. In the Meertens Tunes Collections several subsets of the contents of the Dutch Song Database are provided as downloadable files. The MTC are accessible at `http://www.liederenbank.nl/mtc`.

The Annotated Corpus is a relatively small set of 360 folk song strophes in 26 tune families. The corpus comes with a rich set of manual annotations concerning various aspects of the melodies and lyrics. The selection and annotation procedures for this set are described in detail by Volk and Van Kranenburg (2012).

The previously released version 1.0 contains the digitized melodies and texts as used in various publications, but it does not contain the annotations. These are included in version 1.1, which is described in the current report. For version 2.0, all melodies and lyrics were carefully compared to the original sources — either audio recordings or printed song books — and corrected where necessary. This resulted in a large number of minor improvements. The metadata, including the annotations, have been adapted accordingly, and new sets of annotations have been added. Moreover, an effort was made to make all variants within a tune family better comparable concerning key and metrical differences in the notations of the melodies. In version 2.0.1 the lyrics of the songs have been added in separate files, and a few field names in the metadata have been changed. For the rest, this version is identical to 2.0.

General information about file formats and conventions throughout the Meertens Tune Collections can be found in Van Kranenburg et al. (2014). Some of that information — especially Section 3 — is crucial for understanding the current document, but will not be repeated here.

2 Availability

The data can be downloaded from `http://www.liederenbank.nl/mtc`. Long-term storage and access are guaranteed by the Meertens Institute.

3 Annotated Corpus Version 1.1

The package contains the following directory structure:

```
MTC-ANN-1.1
 |__CHANGES.txt
 |__COPYRIGHT.txt
 |__README.txt
 |__VERSION.txt
 |__krn
 |__ly
 |__lyrics
 |__metadata
 |__mid
 |__pdf
 |__png
 |__wce
```

The **metadata** directory contains the tables with metadata, including the manual annotations. The other directories contain various representations of the songs.

3.1 Content Files

The songs are provided in various formats and renderings, which are described in detail in Van Kranenburg et al. (2014). Each file contains one strophe of a song. The manual input of the melodies is contained in the witchcraft editor (wce) files. The Humdrum **kern files, the lyrics, and the LilyPond source files have been generated from the wce files. The midi files have been generated from the **kern files using the humdrum-extra command `hum2mid`,[1] and the pdf and png representations of the scores have been generated from the LilyPond source files.[2]

3.2 Metadata

The following sets of metadata are included:

- Tune family membership of each strophe
- A reference song for each tune family
- Motif classes and occurrences
- Similarities between phrases (not in version 2.0)
- Similarities between songs (not in version 2.0)
- Form (not in version 2.0)

Each of this sets of annotations is provided as a comma-separated text file.

3.2.1 Metadata Table MTC-ANN-tune-family-labels

This table contains a tune family label for each of the strophes. The name of the tune family has been converted to basic ASCII encoding and the spaces have been replaced by underscores. This results in a unique identifier for each tune family that is convenient for use in e.g., shell scripts. The table has two fields:

songid the basename of the file the metadata record refers to (string).

tunefamily the name of the tune family (string).

3.2.2 Metadata Table MTC-ANN-referencemelodies

For each tune family, the collection specialists of the Meertens Institute selected one strophe, the melody of which they regarded as the best or prototypical example of all melodies in that tune family. The table contains the following two fields:

tunefamily the name of the tune family (string).

reference_song_id the file basename of the melody that is considered the prototypical example of the tune family (string).

3.2.3 Metadata Table MTC-ANN-motifs

This table contains information on motif occurrences in the melodies. Within each tune family an annotator defined several motif classes, and for each motif class all motif occurrences are recorded in the metadata table. The annotators were instructed to assign arbitrary names to the motif classes. In most cases they chose names that reflect the contents of the motif classes (e.g., `3:bagcb` indicates that the motif class has got identification number 3 and typically consists of the notes b, a, g, c, and b). Since there can be considerable variation in the occurrences of a certain motif class, there is no guarantee that the name of the motif class corresponds with each of the occurrences. Each line of the metatdata table describes one motif occurrence. The table contains the following fields:

songid the basename of the song in which the motif occurs (string).

[1] `http://extras.humdrum.org/`

[2] `http://www.lilypond.org/`

from_line the line (phrase) number in which the motif starts - the first line is 1 (integer).

from_bar the bar number in which the motif occurrence starts, relative to the phrase - the first full bar in the line is 1, upbeat is 0 (integer).

from_note the note number relative to the bar at which the motif occurrence starts (integer).

length the length of the motif - either s (short) or l (long), or an integer representing the number of notes.

comments_on_occurrence free text field that can contain a comment of the annotator (string).

motifclass the name of the motif class the motif occurrence belongs to (string).

type the type of the motif class (string).

comments_on_motifclass free text field that can contain a remark of the annotator for the motif class (string).

tunefamily the tune family the song in which the motif occurs is member of (string).

3.2.4 Metadata Table MTC-ANN-phrase-similarities

This table contains similarity assessments of pairs of melodic phrases. The degree of melodic similarity has been determined for two aspects (contour and rhythm) on a three-valued scale: 0 for not similar, 1 for somewhat similar, and 2 for very similar. The annotators systematically compared melodic phrases from a reference melody to phrases of the other melodies within a tune family. The annotators were free to choose pairs of phrases since annotating all possible pairs would have been too laborious. The exact rules to determine the degree of similarity are explained in Volk and Van Kranenburg (2012). In cases where the musical intuition of the annotator was not in accordance with the degree according to the rules, a question mark has been added to the similarity value. The table contains the following fields:

song1_id file basename of the strophe containing phrase 1 (string).

phrase_no_song1 phrase number in song 1 - the first phrase is 1 (integer).

song2_id basename of the strophe containing phrase 2 (string).

phrase_no_song2 phrase number in song 2 - the first phrase is 1 (integer).

contour_similarity similarity between the phrases concerning contour on a three-valued scale ({0, 1, 2}{?, ∅}).

rhythm_similarity similarity between the phrases concerning rhythm on a three-valued scale ({0, 1, 2}{?, ∅}).

In some cases two phrases are taken together, or half a phrase has been taken for comparison. These cases are encoded as follows. Two phrases are represented by concatenating the phrase numbers. For example, phrases 8 and 9 are encoded as 89, and phrases 1 and 2 are encoded as 12. A half phrase is represented by a minus (-) followed by the phrase number, followed by 1 for the first half and 2 for the second half. For example, the second half of the third phrase is encoded as -32.

3.2.5 Metadata Table MTC-ANN-song-similarities

This table contains similarity assessments of pairs of strophes. The degree of similarity has been determined for four aspects (contour, rhythm, motifs, and lyrics) on a three-valued scale: 0 for not similar, 1 for somewhat similar, and 2 for very similar. The annotators systematically compared the reference melody of a tune family with the other melodies within the tune family. The exact rules to determine the degree of similarity are explained in Volk and Van Kranenburg (2012). In cases where the musical intuition of the annotator was not in accordance with the degree according to the rules, a question mark has been added to the similarity value. If one of the four aspects was crucial for recognizing the song as member of the tune family, an exclamation mark has been added to the similarity value. Rhythm and lyrics were optional. Values for these aspects have not been entered for each pair of songs. If no value has been entered, the field is empty. The table contains the following fields:

song1_id file basename of the strophe of song 1 (string).

song2_id file basename of the strophe of song 2 (string).

contour_similarity similarity between the strophes concerning contour on a three-valued scale ({0, 1, 2}{!, ?, ∅}).

rhythm_similarity similarity between the strophes concerning rhythm on a three-valued scale. This field is optional, and therefore can be empty. ({0, 1, 2, ∅}{!, ?, ∅}).

motifs_similarity similarity between the strophes concerning motifs on a three-valued scale ({0, 1, 2}{!, ?, ∅}).

lyrics_similarity similarity between the strophes concerning lyrics on a three-valued scale. This field is optional, and therfore can be empty. ({0, 1, 2, ∅}{!, ?, ∅}).

3.2.6 Metadata Table MTC-ANN-form

For each melody, the form has been determined by the annotators. The form of a melody is represented by a string of letters reflecting the similarities between the phrases. E.g., AxABB′B″. Each letter corresponds to a phrase of the melody. If a phrase is exactly repeated it is represented by the same letter. A varied repetition (e.g., an ouvert-clos structure) is represented by an accented letter. A next variant is represented by a double accented letter. Some very short phrases, which were not considered as independent phrases by the annotators, are represented by an x. In the given example, AxABB′B″, the strophe consists of 6 phrases. The first and third are exact repetitions. The second is too short to be considered an independent phrase. The fifth is a varied repetition of the fourth, and the sixth is also a varied repetition of the fourth, but in a different way than the fifth.

The table contains the following fields:

songid the basename of the file the metadata record refers to (string).

form the form of the melody (string).

4 Annotated Corpus Version 2.0.1

The package contains the following directory structure:

```
MTC-ANN-2.0.1
 ├─ CHANGES.txt
 ├─ COPYRIGHT.txt
 ├─ README.txt
 ├─ VERSION.txt
 ├─ krn
 ├─ ly
 ├─ lyrics
 ├─ metadata
 ├─ mid
 ├─ pdf
 ├─ pdf_tunefamilies
 ├─ png
 ├─ png_motifs
 └─ png_motifs_withoffsets
```

```
└── wce
```

The **`metadata`** directory contains the tables with metadata, including the manual annotations. The other directories contain various representations of the songs.

4.1 Changes with Respect to Version 2.0

- The lyrics of the songs have been added in separate text files.
- In metadata table MTC-ANN-songs, the field “filename” has been renamed to “songid” and the field “songid” has been renamed to “NLB_record_number”. As a result, the field names are consistent accross the tables within MTC-ANN-2.0.1.
- in metadata table MTC-ANN-pitch-duration-normalization, the field “filename” has been renamed to “songid”.
- The extension `.csv` has been added to the file `MTC-ANN-referencemelodies-fieldnames`.

4.2 Content Files

The songs are provided in various formats and renderings, which are described in detail in Van Kranenburg et al. (2014). Each file contains one strophe of a song. The manual input of the melodies is contained in the witchcraft editor (wce) files. The Humdrum **kern files, the lyrics, and the LilyPond source files have been generated from the wce files. The midi files have been generated from the **kern files using humdrum extra command `hum2mid`,[3] and the pdf and png representations of the scores have been generated from the LilyPond source files.[4]. The directory **`pdf_tunefamilies`** contains one pdf for each tune family which contains all strophes belonging to that tune family. The directory **`png_motifs`** contains visualizations of the motif occurrences as annotated by the collection specialists of the Meertens Institute. And the directory **`png_motifs_withoffsets`** contains the same visualizations, but with the note index and score time of the onset added to each note.

4.3 Metadata

The following sets of metadata are included:

- Various metadata for the strophes
- Tune family membership of each strophe
- A reference song for each tune family
- Motif classes and motif occurrences
- Similarity labels for the individual phrases
- A list of sources
- A list of singers
- Information for normalization of the key and metric notation.

Each of this sets of annotations is provided as a comma-separated text file.

[3] `http://extras.humdrum.org/`

[4] `http://www.lilypond.org/`

4.3.1 Metadata Table MTC-ANN-songs

This table contains basic metadata for each strophe. It has the following fields:

songid the basename of the file the metadata record refers to (string).

NLB_record_number the record number of the song in the Database of Dutch Songs (integer).

source_id identifier of the source of the song.

serial_number the serial number of the song in the source (string).

page the page number of the song in the source (string).

singer_id_s identifiers of one or more singers (list of integers).

date_of_recording date of recording (DD-MM-YYYY).

place_of_recording place of recording, in most cases the name of the municipality (string).

latitude latitude of place of recording (float).

longitude longitude of place of recording (float).

title title of the song (string).

firstline first line of the lyrics (string).

strophe_number serial number of the strophe that is in the file (integer).

4.3.2 Metadata Table MTC-ANN-tune-family-labels

This table contains a tune family label for each of the strophes. The name of the tune family has been converted to basic ASCII encoding and the spaces have been replaced by underscores. This results in a unique identifier for each tune family that is convenient for use in e.g., shell scripts. The table has two fields:

songid the basename of the file the metadata record refers to (string).

tunefamily the name of the tune family (string).

4.3.3 Metadata Table MTC-ANN-referencemelodies

For each tune family, the collection specialists of the Meertens Institute selected one strophe whose melody they regarded as the best or prototypical example of all melodies in that tune family. The table contains the following to fields:

tunefamily the name of the tune family (string).

songid the file basename of the melody that is considered the prototypical example of the tune family (string).

4.3.4 Metadata Table MTC-ANN-motifs

For MTC-ANN 2.0 we considerably improved the annotations of motif classes and motif occurrences with respect to the initial annotations as are released in MTC-ANN 1.1. The metadata table MTC-ANN-motifs contains all motif occurrences. The field “change” registers for each motif occurrence in what way(s) the data of the occurrence were changed with respect to MTC-ANN 1.1.

In the initial annotation procedure, the domain experts were not forced to register the length of a motif in an absolute number. They could also choose to indicate whether the motif occurrence was short (`s`) or long (`l`). The annotators gave explicit lengths to the motifs in only 681 out of the original 1406 annotations. To compare the output of pattern matching or pattern discovery algorithms, the absolute length of the motifs is indispensable. The descriptions of the motifs (e.g. "jump (fourth)") did make it possible to infer the intended length of the motif. For MTC-ANN 2.0, a secondary annotator used this information to determine the absolute lengths of all motifs. For all motif occurrences for which the length was converted from `s` or `l` to the exact number of notes, the field "change" contains the value `length`.

In some cases inconsistencies were found in the start positions of the occurrences of a motif class, possibly based on different versions of songs, or simply caused by errors of the annotators. These cases were corrected, also based on the descriptions of the motifs. This happened in twelve cases. In these cases the field "change" contains the value `start`.

Some tune families had phrase repetitions, in which the original annotations only contained a motif occurrence for the first phrase, probably under the assumption that it was self-explaining that this motif was also found correspondingly in repetitions of the phrase. To make the metadata useable for computational corpus studies, all instances of the motif classes were explicitly registered, which resulted in the addition of 233 motif occurrences. For these cases, the field "change" contains the value `pr`.

In one tune family, two motif classes consisting of the same melodic material were annotated under different names. The two motif classes were consolidated as one class, after the assurance by the original annotator that this indeed was an inconsistency in the annotation procedure. This new, consolidated motif class has 26 occurrences, and is represented by `cons` (for consolidated) in the "change" field.

Finally, the annotations were double-checked one more time with a new version of the melodies, to make sure that the positions of the motifs still corresponded to the original positions. The annotator who undertook this procedure also noted a number of cases (52 in total) where motif classes ought to have been annotated, but were not. These motif occurrences were added to the table, and in all these cases the field "change" contains the value `add`.

The metadata table contains the following fields:

motifid unique identifier for the motif occurrence (string).

tunefamily the tune family the song in which the motif occurs is member of (string).

songid the basename of the song in which the motif occurs (string).

begintime score time of the starting time of the motif occurrence in quarter note units (rational number).

duration duration of the motif occurrence in quarter note units (rational number).

endtime end time of the motif occurrence in quarter note units (rational number).

startindex index of the first note of the motif occurrence - first note of the melody is 0 (integer).

endindex index of the last note of the motif occurrence - first note of the melody is 0 (integer).

numberofnotes number of notes in the motif occurrence (integer).

motifclass the name of the motif class the motif occurrence belongs to (string).

description description of the motif class (string).

annotator the annotator who annotated the tune family ({`ann1, ann2, ann3`}).

changes changes - if any - with respect to MTC-ANN 1.1. `add` if the motif occurrence has been added; `cons` if the motif class has been merged with another motif class; `length` if the length of the motif has been adapted; `pr` if the motif occurs in a repeated phrase and has not been annotated initially; `start` if the starting time of the motif occurrence has been adjusted. (list of {`add, cons, length, pr, start, ∅`}).

4.3.5 Metadata Table MTC-ANN-phrase-similarity

The goal of the phrase similarity annotations was to create a ground truth of phrase occurrences in the annotated corpus. Per tune family, three annotators were instructed to assign labels to all phrases, in which they would give the same label to two phrases if they thought they were identical or just slightly varied, or different labels if they were different.

The annotators stated that it would help their overview if they were to make three rather than two distinctions, encoded by a letter combined with a number: if two phrases were "almost identical", they would receive the same letter and number, if they were "related but varied", they would receive the same letter, but different numbers, and if they were "different", they would receive different letters as well. This became therefore the agreed procedure.

Per tune family, the annotators compared the phrases of the variants, assigning the labels of letter and number combinations. The annotators worked independently of each other, so as to be able to compare the three expert judgements afterwards.

songid the file basename of the strophe in which the phrase occurs (string).

phrase_id the index of the phrase - first phrase is 0 (integer).

ann1 the label provided by annotator 1 (string).

ann2 the label provided by annotator 2 (string).

ann3 the label provided by annotator 3 (string).

4.3.6 Metadata Table MTC-ANN-sources

source_id identifier of the source (integer).

title title of the source (string).

author author(s) of the source (list of string).

place_publisher place(s) and publisher(s) of the source (list of strings). Place and publisher are separated by a colon (":").

dating (approximate) dating of the source (string).

sorting_year year that can be used for sorting (integer).

type type of source ({`manuscript, print, audio`}).

copy_used library siglum of the copy that was used for digitization (string).

scan_url url of scanned images of the source, in case the source is publicly available online (string).

4.3.7 Metadata Table MTC-ANN-singers

singer_id identifier of the singer (integer).

year_of_birth year of birth of the singer (integer).

place_of_birth place of birth of the singer, in most cases the name of the municipality (string).

latitude latitude of place of birth.

longitude longitude of place of birth.

4.3.8 Metadata Table MTC-ANN-pitch-duration-normalization

In the Annotated Corpus, as in the remaining Folk Song collections in the MTC, variants are not always represented in the same key, or the same meter. In the metadata table MTC-ANN-pitch-duration-normalization, we provide normalization factors for pitches and durations of the notes. After applying the adjustments as indicated in the table, for each tune family, all member melodies are comparable concerning key and score time.

To remove time dilation, the durations of all notes and rests have to be divided by the value of the field "time_stretch".

For key adjustment, the values in field "pitch_shift" can be used to add the indicated number of semitones to the original pitch. This will ensure that all variants from the same tune family are in the same key.

The metadata table contains the following fields:

songid the basename of the file the metadata record refers to (string).

time_stretch the value (power of two) by which to divide the durations in the melody (float).

pitch_shift the number of semitones to add to the pitches (integer).

5 Abbreviations for Tune Familiy Names

Because the names of the tune families in the annotated corpus are quite long, we standardized short names for the tune families. We recommend using these short names in running text in articles. The following table provides an overview of short and long names.

Tune Family (short)	**Tune Family (long)**	**Size**
Heer	Daar_ging_een_heer_1	16
Jonkheer	Daar_reed_een_jonkheer_1	12
Ruiter2	Daar_was_laatstmaal_een_ruiter_2	17
Maagdje	Daar_zou_er_een_maagdje_vroeg_opstaan_2	10
Dochtertje	Een_Soudaan_had_een_dochtertje_1	13
Lindeboom	Een_lindeboom_stond_in_het_dal_1	9
Zoeteliefjes	En_er_waren_eens_twee_zoeteliefjes	16
Ruiter1	Er_reed_er_eens_een_ruiter_1	27
Herderinnetje	Er_was_een_herderinnetje_1	11
Koopman	Er_was_een_koopman_rijk_en_machtig	17
Meisje	Er_was_een_meisje_van_zestien_jaren_1	15
Vrouwtje	Er_woonde_een_vrouwtje_al_over_het_bos	12
Femmes	Femmes_voulez_vous_eprouver	13
Halewijn2	Heer_Halewijn_2	11
Halewijn4	Heer_Halewijn_4	11
Stavoren	Het_vrouwtje_van_Stavoren_1	8
Zomerdag	Het_was_laatst_op_een_zomerdag	17
Driekoningenavond	Het_was_op_een_driekoningenavond_1	12
Stad	Ik_kwam_laatst_eens_in_de_stad	18
Stil	Kom_laat_ons_nu_zo_stil_niet_zijn_1	11
Schipper	Lieve_schipper_vaar_me_over_1	15
Nood	O_God_ik_leef_in_nood	8
Soldaat	Soldaat_kwam_uit_de_oorlog	17
Bruidje	Vaarwel_bruidje_schoon	11
Verre	Wat_zag_ik_daar_van_verre_1	15
Boom	Zolang_de_boom_zal_bloeien_1	18

6 Relations between MTC-FS 1.0, MTC-ANN 1.0, MTC-ANN 1.1, and MTC-ANN-2.0.1

The songs that are included in MTC-ANN all are included in MTC-FS 1.0 as well.[5] There are however some differences. Because the digitization of the songs and the curation of the data and metadata is a long-lasting, and still ongoing process at the Meertens Institute, each release of a data set in the Meertens Tune Collections contains a snap-shot that includes the songs and metadata as they are present in the Dutch Song Database at a certain point of time. MTC-ANN 1.0 was compiled almost six years before MTC-FS 1.0, and MTC-ANN 2.0 was compiled a year after MTC-FS 1.0. In MTC-ANN 2.0.1 only the lyrics have been added in separate files. For the rest this version is identical to MTC-FS-ANN 2.0. In between the releases, the songs have been subject to many kinds of changes: spelling errors could have been corrected, phrase endings could have been adapted, individual notes could have been changed, tune family membership could have been changed, etc. Because of these differences, great care should be taken when combining songs and metadata from the different data sets.

6.1 Tune Family Membership

The songs have been grouped into tune families by the collection specialists of the Meertens Institute. The melodies of the songs that have been assigned to the same tune family are believed to be related to each other through the process of cultural transmission. The collection specialists had to start this process with many thousands of songs and no pre-existing information on tune families in Dutch oral tradition at all. Their approach was to group songs with similar melodies (Volk and Van Kranenburg, 2012). Given the amount of songs, it is not surprising that the process of grouping songs, and defining tune families, included many revisions, and still is a 'work in progress'.

MTC-ANN 1.0 contains 360 songs in 26 tune families. Each song is member of exactly one tune family. The tune family memberships in MTC-ANN 1.0 reflect the situation in 2008. Based on new insights, some changes have been made afterwards. However, we decided to retain the tune family memberships of MTC-ANN 1.0 in versions 1.1 and 2.0.

In MTC-FS, instead, the updated tune family memberships are included. This comprises only a few differences:

- `NLB072638_01` lost its tune family membership. This song was considered to belong to tune family `Zoetelief jes` in 2008. But later on, the collection specialists concluded that the melody is too different from the other melodies in `Zoetelief jes` to retain this tune family membership.
- Tune family Halewijn 2 has been renamed to Halewijn 1. Therefore, the songs that belong to Halewijn 2 in MTC-ANN belong to Halewijn 1 in MTC-FS.
- Tune family Halewijn 4 has been renamed to Halewijn 3. Therefore, the songs that belong to Halewijn 4 in MTC-ANN belong to Halewijn 3 in MTC-FS.
- `NLB074603_01` had been assigned to Halewijn 4 in MTC-ANN, and has been reassigned to Halewijn 1 in MTC-FS.

Furthermore, MTC-FS includes songs of the annotated tune families that are not part of MTC-ANN because these had not been digitized at the time MTC-ANN was compiled.

6.2 Identifiers

In MTC-ANN-2.0 and 2.0.1 two identifiers have changed with respect to MTC-ANN-1.0, 1,1, and MTC-FS 1.0:

[5] For a description of MTC-FS see Van Kranenburg et al. (2014)

	MTC-ANN 1.0, 1.1, MTC-FS 1.0	MTC-ANN 2.0, 2.0.1
identifier	`NLB074378_01`	`NLB074378_02`
identifier	`NLB073754_01`	`NLB073754_02`

The reason for this is that during a revision of the transcription an initial strophe was transcribed that had not been transcribed before, causing all existing strophe numbers to increase with 1.

6.3 Differences in Contents and Annotations

As preparation for the release of MTC-ANN-2.0 all melodies and lyrics were carefully and systematically compared to the original sources — either audio recordings or printed song books — and corrected where necessary. This resulted in a large number of minor improvements, including spelling and punctuation corrections, changes of individual notes and rests, and removal of unnecessary ties.

Furthermore, the segmentations of the strophes into phrases has been adapted such that the segmentations of all strophes within a tune family correspond with each other. As a consequence, the annotations of the form, the phrase similarities, and the song similarities as included in MTC-ANN 1.1 do not apply anymore. Therefore, these have been left out in MTC-ANN 2.0 and 2.0.1. The phrases similarity annotations in MTC-ANN 2.0 were newly made and are unrelated to those in MTC-ANN 1.1.

6.4 Using MTC-FS as a background corpus for MTC-ANN

It could be desirable to use the large set of melodies that is provided by MTC-FS as background corpus for the melodies in MTC-ANN. For that purpose, it is necessary to remove from MTC-FS all melodies that are somehow related to the melodies in MTC-ANN. This could be accomplished by the following procedure:

1. Identify all melodies in MTC-FS that are also in MTC-ANN, possibly using the information from Section 6.2.

2. For each of these melodies, obtain the tunefamily_id from the metadata of MTC-FS (table MTC-FS in file MTC-FS.csv).

3. From the tunefamily_ids remove the part after the underscore (the *sub-identifier*) to retain the *main-identifier*. E.g. `9744_1` becomes `9744`.

4. Remove all 611 melodies from MTC-FS that are in tune families that have the main-identifiers that are in the list that results from step 3. See below for a full list.

5. Remove all 834 melodies from MTC-FS that do not have a tune family identifier at all.

The reason to disregard the sub-identifier of tunefamily_id is that there might be relations between tune families with the same main-identifier, but different sub-identifiers. E.g., tune family `9744_1` (Er reed er eens een ruiter 1) might be related to `9744_2` (Er reed er eens een ruiter 2). This is not always the case, but to be absolutely sure to remove all melodies that are related to melodies in MTC-ANN it is better to remove all tune families with the same main-identifier. For the same reason, it is necessary to remove all melodies without tune family label. Among these might be members of one of the MTC-ANN tune families that have not been identified yet.

This is the full list of tune families that need to be removed in step 4 from MTC-FS-1.0 to serve as a background corpus for MTC-ANN: ∅, 1419_1, 1419_2, 1419_3, 1419_4, 1419_5, 1507_1, 1507_2, 1507_3, 1507_4, 1507_5, 1507_6, 2694_1, 2694_2, 2694_3, 2774_0, 3676_1, 3676_2, 5301_1, 5301_2, 5301_3, 5301_4, 5301_5, 5301_6, 6382_0, 8592_0, 9664_1, 9664_2, 9665_1, 9665_2, 9665_3, 9665_4, 9665_5, 9668_1, 9668_2, 9668_3, 9668_4, 9668_5, 9673_1, 9673_2, 9673_3, 9673_4, 9673_5, 9673_6, 9675_0, 9676_0, 9677_1, 9677_2, 9686_1, 9686_2, 9686_3, 9693_0, 9700_1, 9715_1, 9715_2, 9722_1, 9722_2, 9722_3, 9722_4, 9743_1, 9743_2, 9744_1, 9744_2, 9744_3, 9744_4, 9744_5, 9749_0, 9779_0, 10043_0.

The resulting background corpus consists of 3,509 melodies that are unrelated to the melodies in MTC-ANN.

7 License and Attribution

If you use the data for a research paper, please cite this report:

Peter van Kranenburg, Berit Janssen, and Anja Volk. 2016. “Meertens Tune Collections: The Annotated Corpus (MTC-ANN) Versions 1.1 and 2.0.1”, *Meertens Online Reports*. No. 2016-1. Meertens Institute, Amsterdam.

Acknowledgements

We are grateful to collection specialist Ellen van der Grijn, who has thorough knowledge of the collection and contributed greatly to all data in the MTC. We thank Ellen van der Grijn, Marieke Klein, and Mariet Kaptein, for doing the first round of annotations that are released in MTC-ANN 1.1. We thank Sanneke van der Ouw and Jorn Janssen, who carefully revised the transcriptions and metadata and created the new annotations as released in MTC-ANN 2.0. We thank Martine de Bruin for maintaining the Dutch Song Database and for her extensive support for the releases of the annotated corpus.

Berit Janssen and Peter van Kranenburg are supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the Tunes & Tales project. For further information, see `http://ehumanities.nl`. Anja Volk is supported by the Netherlands Organisation for Scientific Research through an NWO-VIDI grant (276-35-001).

References

Van Kranenburg, P., M. De Bruin, L. P. Grijp, and F. Wiering (2014). The meertens tune collections. Meertens Online Reports 2014-1, Meertens Institute, Amsterdam.

Volk, A. and P. Van Kranenburg (2012). Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae 16*(3), 317–339.