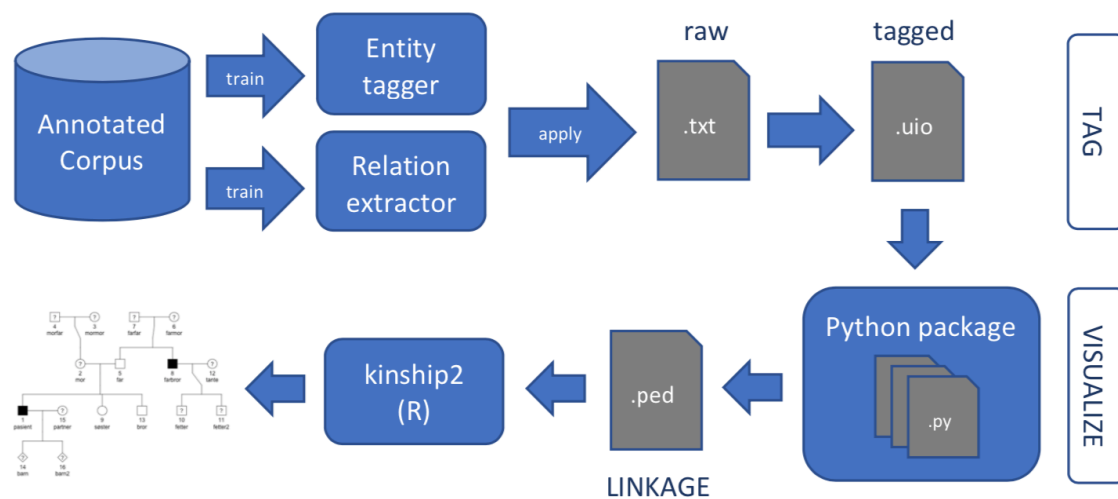


# From named entity tags and their relations to pedigree charts

## Status report

### 1. Introduction

This document describes work in progress on developing a prototype for visualizing pedigree charts from family history texts annotated with named entities (NE) and relations. Figure 1 shows a workflow consisting of two phases. First, an entity tagger and a relation extractor on an annotated corpus are trained and then applied on raw texts to produce a tagged output. The sub-sequent visualization phase, which this work focuses on, aims at restructuring the information from tagged files into LINKAGE<sup>1</sup> format – a standard format which can be used in a number of pedigree visualizing software, such as the *kinship2*<sup>2</sup> R package used here – to produce a pedigree chart. The generated pedigree chart contains information about family members and their affected status by a genetic trait.



**Figure 1.** Pedigree visualization workflow: from raw texts to family trees.

### 2. Method

The library for restructuring the tagged documents into LINKAGE format is being implemented in Python 3. The number of external libraries has been limited for better portability. Currently the only dependency is the *spacy\_udpipe*<sup>3</sup> Python package used for lemmatization and R. The *kinship2* package for visualizing the pedigree chart is launched with an R script and it is ran directly from within the Python package so that restructuring and visualizing can be done by executing a single script. The NE tags and relations used are from the NorSynthClinical corpus<sup>4</sup>. Restructuring NE tags and relations into LINKAGE format is implemented with a rule-based approach and a small ontology of Norwegian family relation

<sup>1</sup> <https://www.mv.helsinki.fi/home/tsjuntun/autogscan/pedigreefile.html>

<sup>2</sup> <https://cran.r-project.org/web/packages/kinship2/index.html>

<sup>3</sup> <https://pypi.org/project/spacy-udpipe/>

<sup>4</sup> <https://github.com/lrgoslo/NorSynthClinical>

names created for this purpose. Patient gender is inferred based on pronouns tagged as 'SELF'. For negation (NEG) tags simple heuristics based on a small set of verbs distinguishes certain cases of doubt (e.g. *vet ikke*) from actual negations (e.g. *ikke syk*).

### 3. Preliminary results

Figure 2 shows an example (1) from the NorSynthClinical corpus (example1.txt), its tagged version (2) together with its automatic LINKAGE format (3) and the generated pedigree chart<sup>5</sup> (4). Family terms are included only for convenience, but can be changed/omitted.

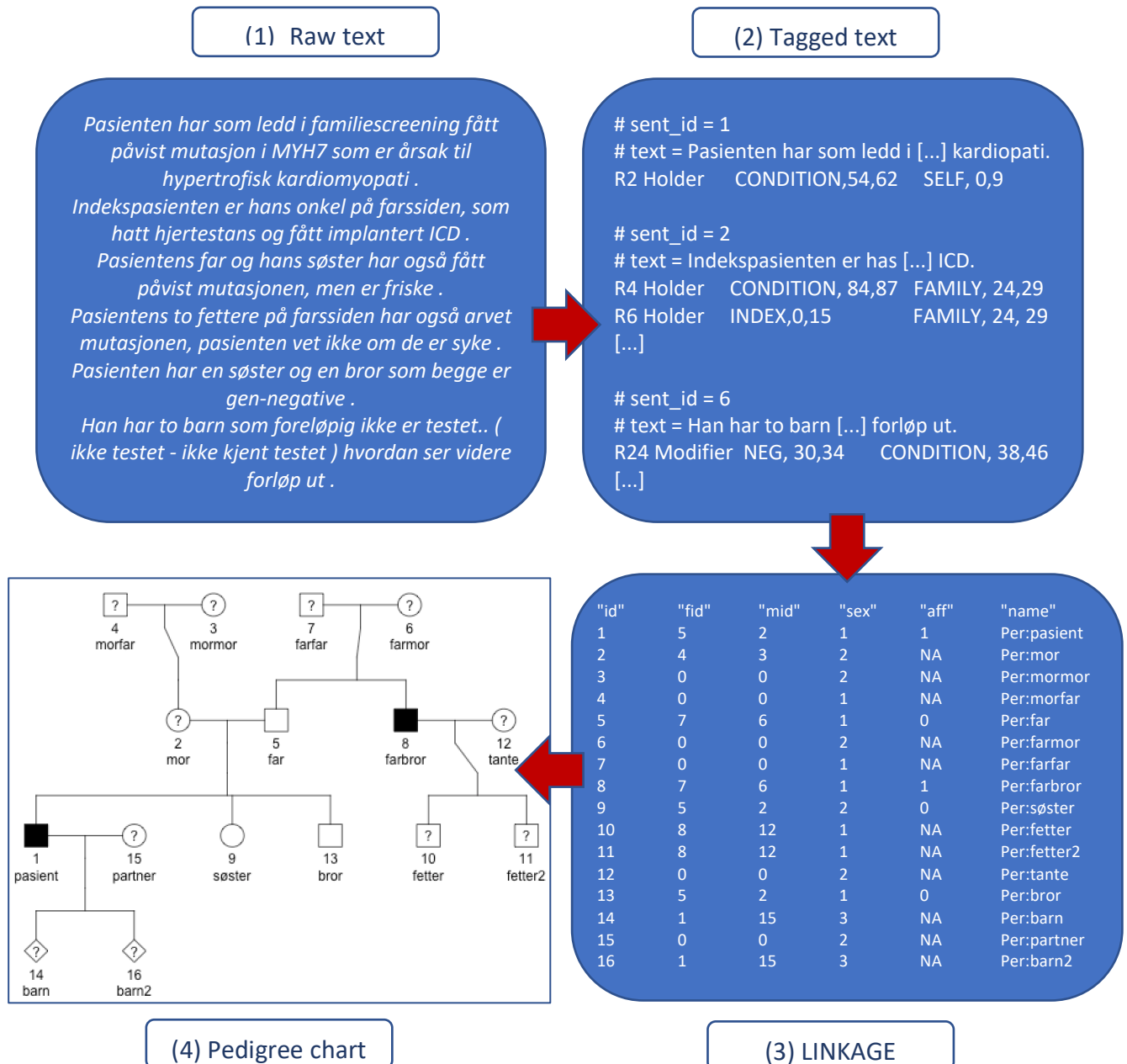



Figure 2. Example outputs.

<sup>5</sup> Affected status values may be incorrect, how to map these exactly will still need to be discussed.

The tagged format (2) presented here is the one used in the NorSynthClinical corpus, in which, for each sentence, the type of relation and the two related entity tags (including their start and end character indices) are presented on separate lines.

The LINKAGE format (3), on the other hand, is a space- (or tab)-separated file which includes information about one family member per line with the ID of the father (*fid*), the ID of the mother (*mid*), sex (with value 1 if male, 2 if female, 3 if unknown) and affected status (*aff*; value 1 if affected, 0 if not, NA if unknown) and as an optional column, a family member name to display (included in post-madeup LINKAGE versions<sup>6</sup>).

The pedigree chart (4) shows a family tree, where family members have a numeric ID and an optional name. Names are sometimes changed during the re-structuring process compared to the originally occurring term in the text to incorporate information from the SIDE tag (e.g. *onkel* to *farbror*). Male members are presented by squares, female members by circles and unknown gender is associated with a square on its point. Affected individuals have a darkened (filled) shape, unaffected ones have a white (empty) shape, while unknown status is indicated by '?'.  


#### 4. Current status

Table 1 shows the current implementation status per relation type and entity pair.

Relation	Entity 1	Entity 2	Status
Holder	CONDITION / EVENT	SELF / FAMILY	Done*
Holder	INDEX	SELF / FAMILY	Done
Modifier	AMOUNT	FAMILY	Done*
Modifier	AMOUNT	CONDITION / EVENT	To do
Modifier	AGE	SELF / FAMILY / INDEX	Needed?
Modifier	SIDE	FAMILY	Done*
Modifier	NEG	FAMILY	Done
Modifier	NEG	CONDITION / EVENT	Done*
Modifier	TEMPORAL	CONDITION / EVENT	Needed?
Related_to	FAMILY	FAMILY	Ongoing
Related_to	FAMILY	SELF	Done*
Subset	FAMILY	FAMILY	To do
Partner	FAMILY	FAMILY	To do

**Table 1.** Relation types, possible entity combinations and their implementation status.  
 (\* indicates known issues or rare phenomena not handled)

Currently the implementation of the Related\_to relation between two FAMILY entities is ongoing, which means that family member mentions are mostly assumed to be from the

<sup>6</sup>[https://www.mv.helsinki.fi/home/tsjuntun/autogscan/pedigreefile.html#:~:text=Post%2Dmakeped%20format%20comes%20after,\(FASTLINK\)%20package%20and%20others.&text=Makeped%20recoded%20pedigree%20and%20individual,end%20of%20the%20each%20line](https://www.mv.helsinki.fi/home/tsjuntun/autogscan/pedigreefile.html#:~:text=Post%2Dmakeped%20format%20comes%20after,(FASTLINK)%20package%20and%20others.&text=Makeped%20recoded%20pedigree%20and%20individual,end%20of%20the%20each%20line)

perspective of the patient (SELF), except for sibling and parent relations (e.g. *fettere* is assumed to be the patient's cousin). Since some relations and affected status values might not be completely solvable with the current annotations, it is planned to implement a mechanism for issuing warnings for relations not yet handled which would enable a semi-automatic use even of an incomplete system. Users could then post-edit a LINKAGE file to add (change) the relations not (correctly) handled.

## 5. Implementation decisions and assumptions

The current version reflects a number of implementation decisions and assumptions that might need to be discussed and modified. These are:

- The pedigree chart is generated per single document per patient.
- Family members only up to three generations back from the patient are handled.
- Some family members are included even if not mentioned to help establish other mentioned relations (e.g. *mor*, *morfar* in Figure 2.). These could be removed in the future.
- Words with CONDITION tags that are referring to not being ill (e.g. *frisk*) are assumed to always refer to the genetic trait / 'index condition'.
- Words which are labelled as (part of) a FAMILY entity, but which are not part of the family member ontology, are excluded (e.g. *familie*, *mange*, *andre*, *av*, *den*). Some of these excluded terms could be signs of not using a minimal span when annotating FAMILY tags.
- The generated LINKAGE file contains values need for plotting with kinship2 which is slightly different from some 'standard' LINKAGE formats required by other tools (e.g. affected status values, unknown gender).

## 6. Limitations of the current implementation

There are a number of known limitations of the current implementation, some of which are due to the properties of natural language and some are related to the annotations.

- No generalizable solutions are possible (without hard-coding) with the current annotations in some cases:
  - Determining affected status
    - information about a variety of CONDITION / EVENT types, not only the main genetic trait -> how to handle this?
    - NEG used both for doubt and negation, the target entity varies and it might not be a CONDITION: e.g. *ikke syk* vs. *vet ikke om de er syke* where *vet ikke* annotated as NEG of *syke* (but affected status should be NA not 0 for the latter)
- Ambiguities left
  - FAMILY entity
    - with SIDE: 'Pasientens to fettere på farssiden' (if father has more than one sibling)
    - with FAMILY: 'Pasientens far og hans ssøster' (if patient male) -> sometimes disambiguated in annotation using SELF
  - CONDITION entity

- Synonyms referring to the same condition for determining affected status
- Unclear order of appearance of relations
  - Not alphabetic for relation or for entity tokens / not length-based -> find out what determines the order to handle better inter-dependent entity tags and relations (e.g. Modifier with AMOUNT, NEG, SIDE)
  - Inter-dependent relations (Modifer, Related\_to) might not update all necessary entries currently.
- Direction of relations can go both ways sometimes:
  - Modifier with AMOUNT-FAMILY and FAMILY-AMOUNT, e.g.:
    - *to fettere* (FAMILY-AMOUNT) but *to barn* (AMOUNT-FAMILY)
    - > handled now, but it would be simpler to manage a consistent positioning for entities in general
  - Related\_to
    - *Pasientens bestefar* (SELF-FAMILY) vs. *Pasientens to fettere* (FAMILY-SELF) -> could be unclear who possesses after lemmatization without ordering, e.g. *far(s) - kusin* vs. *kusin(s) far*
- The implementation relies entirely on the tag set of the NorSynthClinical corpus.
- The implementation is language specific (e.g. family relation names) and it would need some restructuring, especially for non-Scandinavian languages.