# De-identification of Dutch Medical Text

**Erik Tjong Kim Sang[1], Ben de Vries[1], Wouter Smink[2], Bernard Veldkamp[2], Gerben Westerhof[2], Anneke Sools[2]**

[1]Netherlands eScience Center
[2]Department of Psychology, Health and Technology, University of Twente
e.tjongkimsang@esciencecenter.nl

## Introduction

Medical texts often contain information revealing the identity of patients. This information must be removed before the texts can be used for scientific research. In this paper we describe our efforts to anonymize or de-identify mental health texts written in Dutch. Similar work has earlier been done by Menger et al. (2017) and Scheurwegs et al. (2013).

## Data

Our medical texts consist of emails from online therapies in which patients discuss their personal situation. These data cannot be shared. In order to facilitate an open evaluation process, we use Wikipedia biographies, which after an egofication process have been converted to autobiographies which are similar to our mental health text.

## Method

We remove all person names, locations, organizations, numbers, dates, mail addresses, urls and miscellaneous named entities from text. For identifying these entities, we rely on the tool Frog (Van den Bosch et al. 2007) and a collection of entity identifying rules.

## Translated example of processed text

PER
I (LOC, DATE - LOC, DATE) was a MISC painter, printmaker and draughtsman. I is generally considered to be one of the greatest painters and printmakers of MISC art, and as the most important MISC master of the DATE. In total, I created about NUM paintings, NUM prints and NUM drawings. My work is part of the baroque art and was influenced by caravaggism, although I has never visited LOC.

## Evaluation results

| | UNL | AVG | PER | NUM |
|---|---|---|---|---|
| TKS2019 | 84% | 55% | 70% | 70% |
| Menger2017 | 40% | 31% | 55% | 1% |

Our system outperforms previous work (Menger at al. 2017) with respect to F1 scores on all evaluated classes and class combinations

## References

Menger V, Scheepers F, Van Wijk L, Spruit M. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. Telematics and Informatics. 2017;35:727736.

Scheurwegs E, Luyckx K, Van der Schueren F, Van den Bulcke T. De-Identification of Clinical Free Text in Dutch with Limited Training Data: A Case Study. In: Proceedings of the Workshop on NLP for Medicine and Biology. RANLP, Hissar, Bulgaria; 2013. p. 1823.