

ANONYMIZATION CORPUS FOR AUTOMATED ANONYMIZATION/ENCODING OF GDPR-RELEVANT DATA



VYTAUTAS
MAGNUS
UNIVERSITY
Faculty of
Informatics

AUTHORS: Danguolė Kalinauskaitė danguole.kalinauskaite@vdu.lt Milita Songailaitė milita.songailaite@vdu.lt Justina Mandravickaitė justina.mandravickaite@vdu.lt Justinas Juozas Dainauskas justinas.dainauskas@vdu.lt Agnė Paulauskaitė-Tarasevičienė agne.paulauskaite-taraseviciene@ktu.lt Tomas Krilavičius tomas.krilavicius@vdu.lt

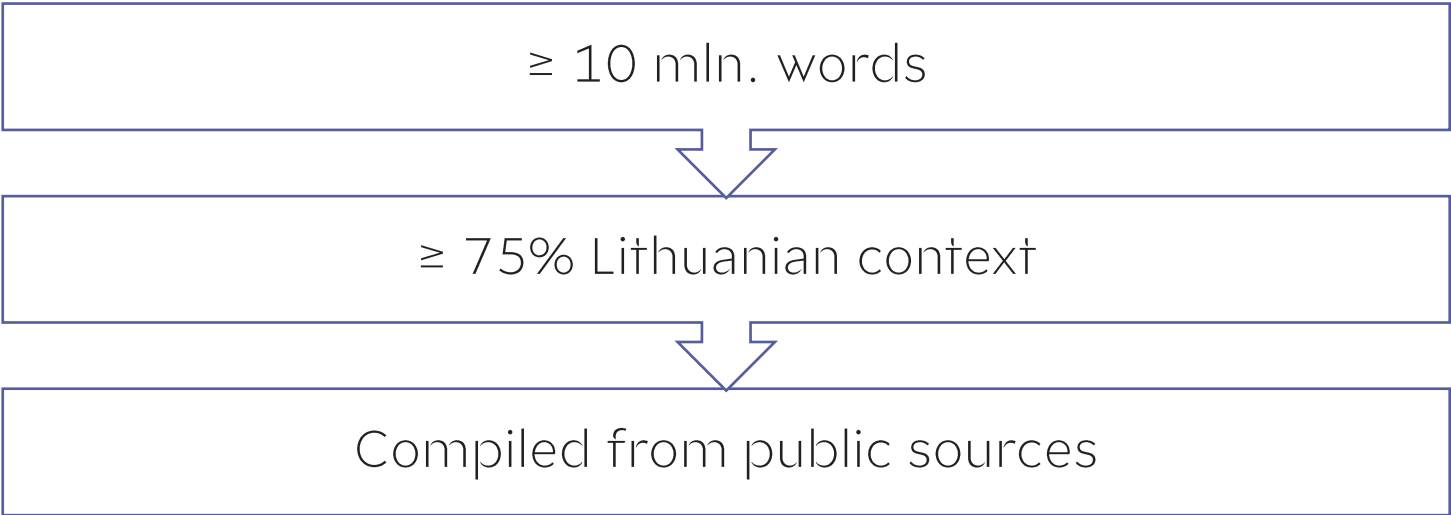
1 INTRODUCTION

Personal data in text must be protected following GDPR requirements, but manual anonymization is slow, error-prone, and hard to scale.

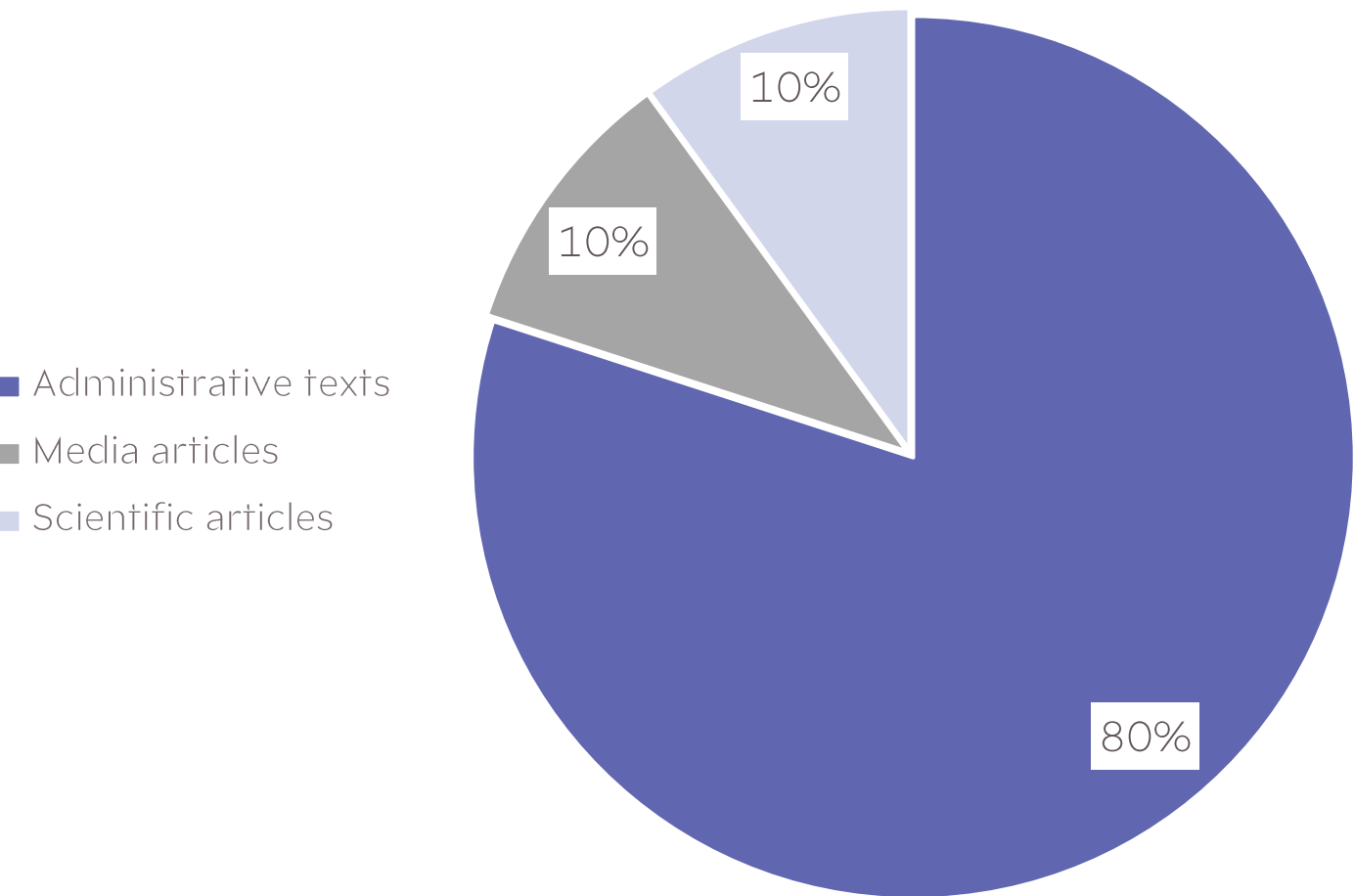
High-quality, domain-specific training data is needed to train a model for detecting and transforming sensitive entities across diverse text types.

We present an **anonymization corpus IN PROGRESS** designed to train and evaluate automated anonymization systems, enabling more reliable protection of GDPR-relevant data.

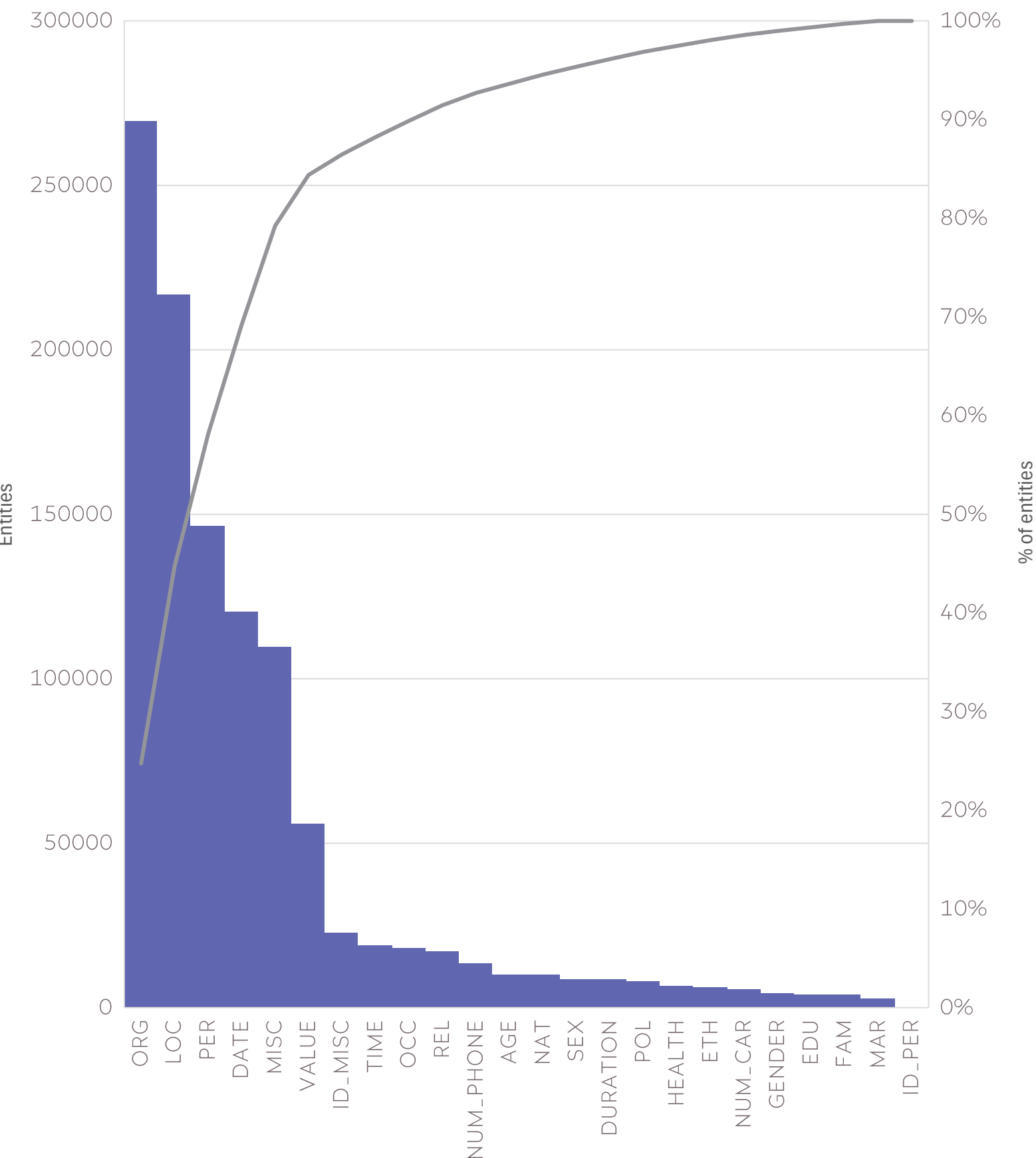
3 REQUIREMENTS



4 COMPOSITION



6 NAMED ENTITY COVERAGE



2 SOURCES & SELECTION

- 1. Administrative sources:**
- a) Lithuanian municipalities websites (e.g., <https://www.akmene.lt/>)
 - b) Lithuanian governmental pages (e.g., <https://ird.lrv.lt/lt/>)
 - c) Stocks, bonds and obligations documents (e.g., <https://nasdaqbaltic.com/lt/>)
 - d) Parliament transcripts (e.g., https://www.lrs.lt/sip/portal.show?p_r=35694&p_k=1)
 - e) Public university senate decisions (e.g., <https://www.vdu.lt/lt/apie-vdu-kaune/struktura-ir-valdymas/senatas/senato-nutarimai/>)
 - f) Generated documents based on real document templates
- 2. Scientific sources:**
- a) Various scientific journals (e.g., <https://www.journals.vu.lt/actahas/en/>)
 - b) Students' bachelor's and master's theses (e.g., <https://www.vdu.lt/cris/home>)
- 3. Media sources:**
- a) Various news articles from <https://www.lrt.lt/>
- Named entities should reflect general personal information about **real-world persons**.

5 STATISTICS

Type of source	Documents	Total words	Words per doc. (avg.)
Media	19 481	6 944 725	377.71
Scientific	1 116	3 182 471	4 545.17
Administrative	42 080	13 326 932	316.70

7 ANNOTATION

Metadata

Source: 4-zin-503-833715 Category: ziniasklaida Word count: 194

General Entity Types

PER 1 ID_PER 2 ID_MISC 3 NUM_CAR 4 NUM_PHONE 5 LOC 6

ORG 7 NAT 8 OCC 9 EDU 0 AGE q TIME w DATE e

DURATION t VALUE a MISC s

Sensitive Entity Types

REL d POL f SEX g GENDER z MAR x FAM c ETH v HEALTH b

Text to Annotate

„Oro navigacijos“ atstovas: GPS ryšio trikdžių paskutinį ketvirtį fiksuojama nemažai Orlaivių įgulių pranešimų apie GPS ryšio trikdžius per metus padaugėjo daugiau nei 6 kartus, sako įmonės „Oro navigacija“ Operacinės veiklos departamento vadovas Tomas Montvila.

„Paskutinį ketvirtį jų nemažai fiksuojama“, – penktadienį LRT RADIJUI teigė T. Montvila. Jis tvirtino, kad ketvirtadienio vakarą iš Londono skridęs ir Vilniuje turėjęs leistis keleivinis „Ryanair“ lėktuvas nenusileido dėl ryšio problemų.

„Mūsų duomenimis, orlaiviui buvo GPS trukdžiai ir tūpimo metu suveikė vadinamoji avionikos perspėjimo sistema, kad orlaivis gali susidurti su žeme ir jie priėmė sprendimą skristi į atsarginį oro uostą Varšuvoje. Ten pasitikrino savo avionikos sistemas ir grįžo į Vilnių“, – teigė T. Montvila.

Pasak jo, GPS trikdžiai gali kilti dėl įvairių priežasčių.

„Daug priežasčių gali būti. Tai ir nedraugiškos šalys, ir visa kita“, – tvirtino „Oro navigacijos“ atstovas.

Susiekimo ministro patarėjo Luko Paškevičiaus teigimu, visi kiti lėktuvai tiek ketvirtadienio vakare, tiek penktadienio ryte visuose šalies oro uostuose kilo ir leidosi saugiai.

Ketvirtadienio vakarą iš Londono skridęs ir Vilniuje turėjęs leistis keleivinis „Ryanair“ lėktuvas jau buvo atvykęs į sostinės oro uostą, tačiau jame nenusileido – apsigrėžė ir nuskrido į Varšuvą. Po priverstinės stotelės Lenkijos sostinėje 23.47 val. atskrido į Vilniaus oro uostą ir šį kartą nusileido.

8 FUTURE PLANS

- The corpus will be used:
- For training models for automated texts anonymization
 - For analysis of various aspects of named entities in the Lithuanian language
 - As a resource for comparing anonymization and named entity recognition tools for Lithuanian

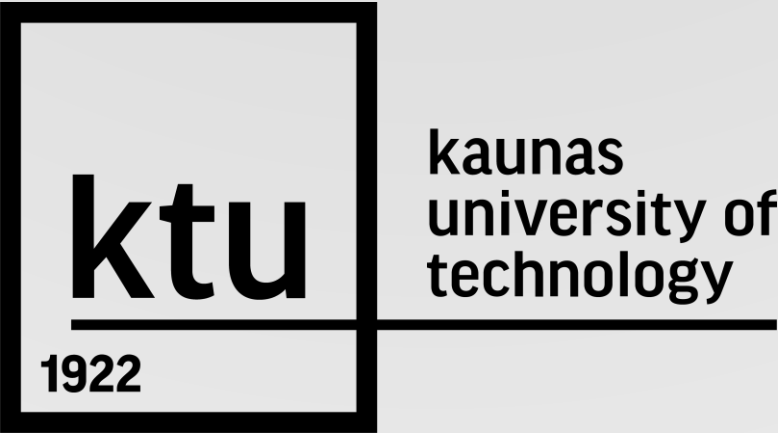
9 ASK US

- How do we deal with tricky cases?
- How was the annotation scheme defined?
- How were the texts selected and obtained?

ACKNOWLEDGEMENTS

These results are part of the project “Developing an Anonymization Corpus” implemented from the funds of the Economic Recovery and Resilience Plan “New Generation Lithuania”, project No. 02-100-K-0001. We thank Eglė Rimkienė for helping to prepare the presentation of the results.

SustAIN
Liv
Work



CARD
CENTRE
FOR APPLIED
RESEARCH
AND
DEVELOPMENT