

Corpus design · Alignment · Language resources

A Ukrainian-English Parallel Corpus of Public Safety Communication

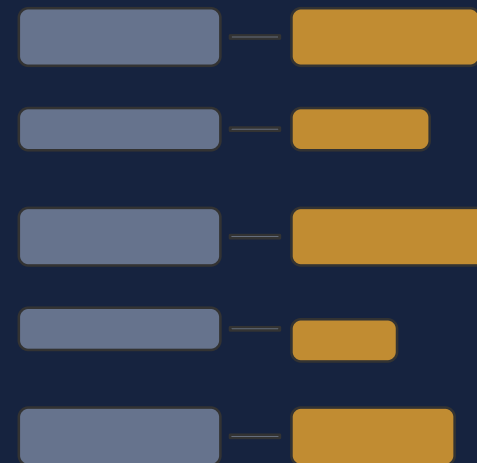
Design, Alignment, and Research Potential

Svitlana Matvieieva

Kaunas University of Technology

K. Donelaičio St. 73, 44249 Kaunas, Lithuania

ORCID 0000-0002-8357-9366 · sam141175@gmail.com



MOTIVATION

A resource gap in a high-stakes domain

Parallel corpora underpin contrastive analysis, terminology work, translation description and alignment-based applications. Yet Ukrainian still lacks small, curated, domain-specific bilingual corpora whose sources, metadata and reuse conditions are transparent.

Why public safety communication?

- short and operational
- socially consequential
- structurally unstable
- published bilingually by accountable institutions

The material

- App notices
- First-aid announcements
- Route safety notes
- Electricity-saving instructions
- Transport updates
- Mixed weekly digests

Brief, but harder to model than long conventional prose.

The argument in one line

**Bilinguality is real,
but rarely literal.**

The aim is not merely to announce a new corpus. It is to show that a Ukrainian-English corpus of public safety communication requires a specifically curated, authorial approach to document pairing, segmentation, alignment and metadata design.

The English version is often

shortened

reorganized

explicated

audience-adapted

rather than a sentence-by-sentence replica

Two institutionally accountable sources

Portal of Kyiv

kyivcity.gov.ua

Official web portal of the Kyiv City State Administration

Kyiv Digital

kyiv.digital

Municipal digital service ecosystem of the Kyiv territorial community

Why this source base is methodologically attractive

1

Paired publication

Ukrainian and English versions appear within the same item or within clearly linked bilingual page structures.

2

Transparent reuse

The portal states that its content is available under CC BY 4.0 unless otherwise noted — far clearer than most institutional environments.

3

Right genre range

Short notices, procedural guidance, app-related safety messaging and multi-item public digests.

Why generic sentence alignment fails

01

Audience-oriented restructuring

The Ukrainian page states explicitly that the English version is not a copy, but a selection of services relevant for foreign users — foregrounding air alert notifications, curfew reminders and a shelter map.

Kyiv City State Administration, 2022a

02

Non-literal title adaptation

Route names *Kuïв хрещатий* and *Kuïв застрономічний* are rendered as *Central Kyiv* and *Foodie Kyiv* rather than through literal lexical matching.

Kyiv City State Administration, 2022b

03

Terminological instability

A weekly digest advises pedestrians to wear “reflective elements or flickers”. Authentic institutional usage — preserved in the primary text, but flagged in terminological annotation.

Kyiv City State Administration, 2022c

A generic web-harvested corpus would flatten these relations or misclassify them as noise.
Our corpus treats them as central evidence.

An authorial workflow in five stages



Corpus design is adapted to the actual communicative behaviour of public safety discourse — not to the expectations of standard sentence-alignment pipelines.

Five relation tags

exact

Stable near-1:1 correspondence between the two versions.

compressed

The English segment condenses the Ukrainian source.

expanded

Local references are explicated for a non-local readership.

adapted

Information is re-authored and reorganised around a different audience.

unmatched

A segment with no counterpart in the other version.

What the tags refuse to do

- normalize away the source text
- discard non-literal correspondences as accidental deviation
- reduce alignment to sentence length

*Instead:
model the real behaviour
of bilingual institutional discourse*

Alignment in practice

exact

A stable near-1:1 correspondence — the kind of pair standard pipelines handle well.

KCSA, 2022b

UK

Для безпеки учасників на екскурсійних маршрутах передбачені укриття.



EN

For the safety of the participants, there are shelters along the excursion routes.

adapted

A broader adaptive relation: the English text does not reproduce every local detail in the same order.

KCSA, 2022a

UK

Ukrainian notice: full local description of the Kyiv Digital app and its services



EN

English version: reorganised around foreign users — air alerts, curfew, shelter map, metro and rail tickets

Metadata that survives the build

Within the interoperability logic of CLARIN, a corpus becomes reusable only when its metadata is rich enough to make it findable, interpretable and interoperable across workflows.

Branco et al., 2023; Wilkinson et al., 2016

Translation status is recorded explicitly

- a close rendering
- a reduced version
- an audience-adapted re-authoring

Document-level metadata fields

☐ Institution

☐ Date

☐ Channel

☐ Genre

☐ Topic

☐ Intended audience

☐ Territorial scope

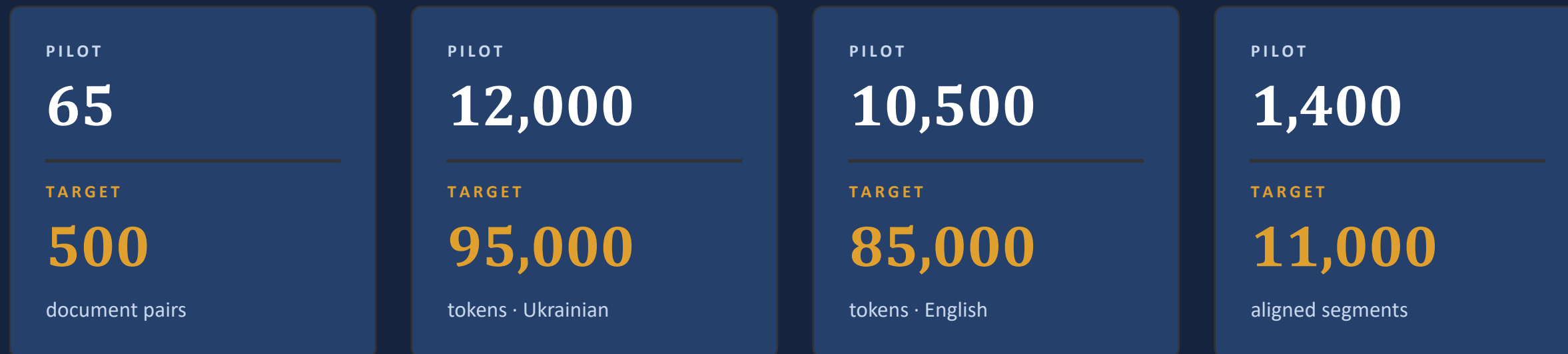
☐ Rights status

☐ Translation status

☐ Alignment confidence

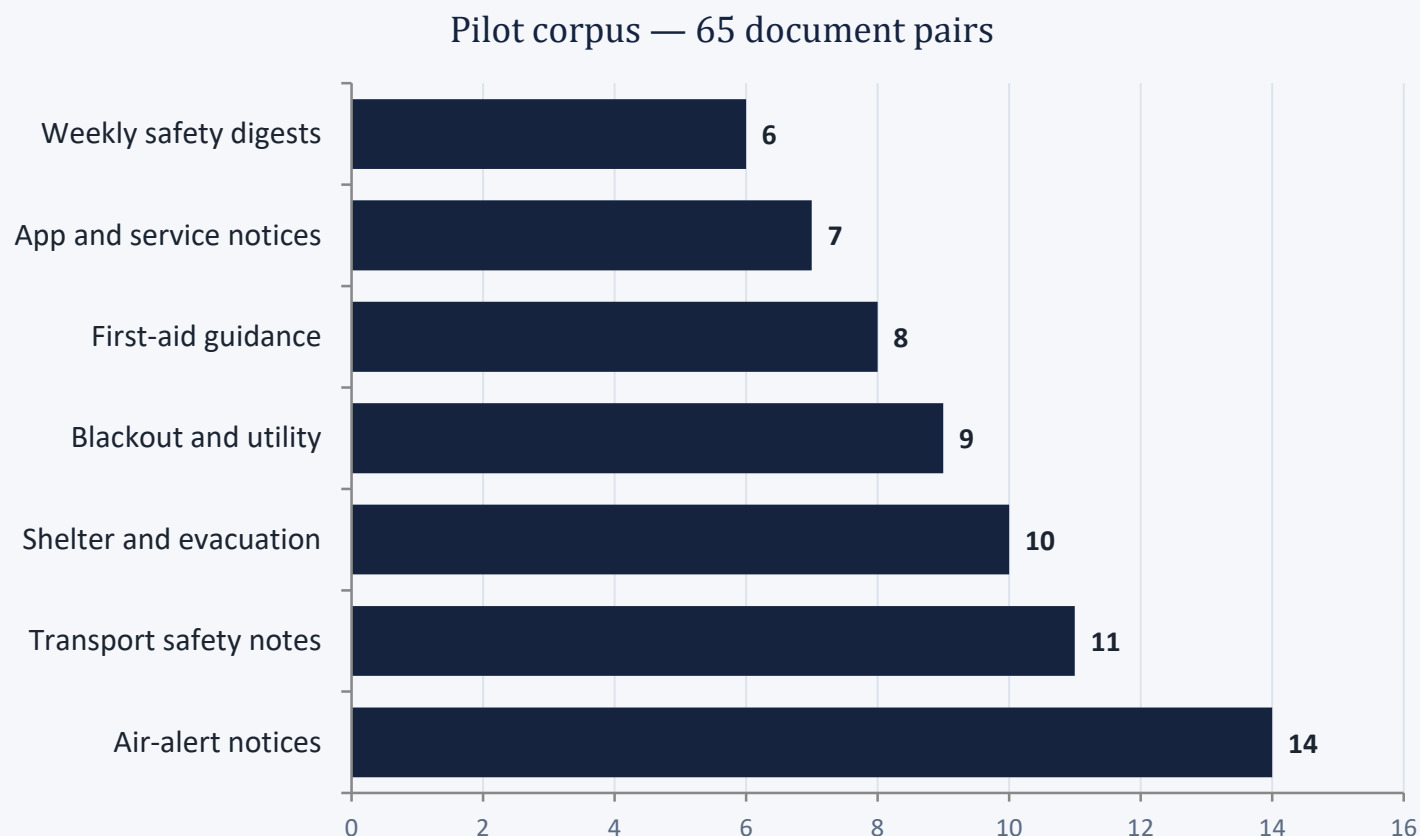
COMPOSITION

Small by design, dense by intention



These figures are deliberately modest. The intended contribution is annotation density and traceability rather than volume — a curated reference corpus, not a large-scale harvest. Sampling frame: municipal public safety communication published with an official English version, February 2022 onwards.

Genre distribution and quota sampling



Quota, not proportion

Quota sampling is applied across the seven subdomains, so that each communicative type is represented in numbers sufficient for contrastive analysis. Proportional sampling would let weekly digests dominate.

Time-spread sampling

Documents are sampled across the full time span to capture seasonal and situational variation — energy guidance in winter, mobility guidance in summer.

Target: 500 pairs, same proportions

Two limitations, recorded in the documentation

1

City-specific

The pilot draws on Kyiv municipal channels only.
Generalisation to national-level safety discourse requires additional institutional sources — planned for the second stage.

Stage two: broaden the institutional base

2

Bilingual-only eligibility

Only material published in both languages is eligible.
Ukrainian-only texts, which constitute the majority of municipal safety output, fall outside the corpus by definition.

A boundary of the design, not an oversight

Both limitations follow directly from the design and are published with the data, so that users know what the corpus does and does not represent.

Availability, licensing, sustainability

Formats

TEI-XML

master edition with segment-level alignment pointers and relation tags

TMX

export for translation-memory and CAT environments

TSV

aligned segment pairs with relation tags, for scripting and machine learning

CMDI

document-level metadata, harvestable by the CLARIN Virtual Language Observatory

A corpus that cannot be located, cited or reused under clear conditions does not fulfil its purpose, however carefully it has been annotated.

Deposit and access

CLARIN-LT repository, in which Kaunas University of Technology participates as a consortium member: persistent handle identifiers and long-term preservation. Access is public and unrestricted, and every release separately citable.

Licensing

Source texts are CC BY 4.0, so the corpus is released under the same licence. Item-level attribution to the Kyiv City State Administration, with original URL and retrieval date. Annotation layer CC BY 4.0; CMDI records CC0. Items with unclear rights are excluded.

Sustainability

Versioned releases; guidelines and pairing protocol published with the data, allowing periodic extension.

Research potential in three fields

Corpus linguistics

- formulaic warning patterns and directives
- address forms and temporal framing
- nominal compression in short institutional genres
- how safety meaning is distributed across titles, instructions and isolated warnings

Translation studies

- an asymmetry common in practice, under-described in parallel resources
- route-name adaptation and reframing of app functionalities
- locally embedded terms such as shelter map or flickers
- compression, omission and explicitation — qualitative and small-scale quantitative

NLP

- compact evaluation set for document pairing and segment alignment
- bilingual terminology extraction and controlled vocabulary building
- relation-aware segmentation and metadata-sensitive alignment
- complements OPUS-scale repositories by traceability, not volume

Classic sentence-alignment models remain foundational (Gale & Church, 1993), but this material shows that a public-safety corpus for Ukrainian and English also requires relation-aware segmentation and metadata-sensitive alignment.

CONCLUSION

What the project contributes

**Small by design,
but dense in metadata,
traceability and analytical value.**

The main contribution is methodological: a corpus model based on manual document pairing, communicative-segment alignment, relation tagging and reuse-aware source documentation — combined with an explicit specification of composition, sampling design and access conditions.

The pilot data already show that public safety communication contains exactly the kinds of translational and structural features that generic parallel resources miss. A corpus of this type is useful not only as a dataset, but as a model for building more precise bilingual resources for Ukrainian.

Bilinguality is real

but rarely literal

Curation over harvesting

manual pairing, typed relations

Open by default

CC BY 4.0, CLARIN-LT, persistent IDs

References

Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., de Jong, F., et al. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation*, 59, 4359–4390.

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1), 75–102.

Kyiv City State Administration (2022a, October 25). An English version of the Kyiv Digital city app is now available. Portal of Kyiv.

Kyiv City State Administration (2022b, July 11). Kyiv: Important info as of July 11. Portal of Kyiv.

Kyiv City State Administration (2022c, October 28). Kyiv: Highlights of the week. Portal of Kyiv.

Matvieieva, S., Lemish, N., Zernetska, A., Babych, V., & Torgovets, M. (2022). English-Ukrainian parallel corpus: Prerequisites for building and practical use. *Studies about Languages*, 40, 61–74.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC'12* (pp. 2214–2218). ELRA.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

Zanettin, F. (2017). Parallel corpora in translation studies: Issues in corpus design and analysis. In M. Olohan (Ed.), *Intercultural Faultlines*. Routledge.

Svitlana Matvieieva · Kaunas University of Technology · [ORCID 0000-0002-8357-9366](https://orcid.org/0000-0002-8357-9366) · sam141175@gmail.com