

Korpus przykładowy: Poznański Czerwiec 1956

Materiał ćwiczeniowy do warsztatu „**Sprzątanie danych narzędziami AI**” (doktoranci, WPiA UAM, 28.06.2026). Korpus jest **celowo brudny i wielopostaciowy** — to surowiec, na którym ćwiczymy czyszczenie, strukturyzację i ekstrakcję danych modelami językowymi (Bielik / LM Studio i usługi zewnętrzne).

[!WARNING] ## Charakter materiału — przeczytaj zanim użyjesz

- To korpus **dydaktyczny i częściowo syntetyczny**. Część plików (relacje, wycinki, CSV, lista jednostek) to **rekonstrukcje** napisane na potrzeby ćwiczenia i **celowo zniekształcone** (błędy „OCR”, niespójne formaty, duplikaty, literówki).
- **Imiona i nazwiska świadków są fikcyjne**. Postacie publiczne (Cyrankiewicz, gen. Popławski, Romek Strzałkowski), nazwy jednostek, miejsca i daty pochodzą z udokumentowanego przebiegu wydarzeń.
- **Nie cytować tych plików jako źródła historycznego**. Zweryfikowane, prawdziwe materiały (dosłowne cytaty, liczby, fotografie, bibliografia, linki) → ZRODLA.md; gdzie szukać samemu → GDZIE-SZUKAC.md.
- Każdy plik-rekonstrukcja ma w nagłówku znacznik **ŹRÓDŁO: rekonstrukcja dydaktyczna**.

Co tu jest (mapa „brudu” → use-case z use-cases.md)

Plik / katalog	Na czym polega „brud”	Czego uczy
relacje-swiadkow/relacja-0*.txt	brak interpunkcji, wtrącenia (yyy, no więc), nieoznaczeni mówcy, fikcyjne dane osobowe	transkrypcja #7 · anonimizacja #9 · streszczenie #2
nagrania/przemowienie-cyrankiewicza-1956-06-29.mp3	PRAWDZIWE nagranie radiowe (Cyrankiewicz, 29.06.1956, Radio Poznań) — pełne + klip 90 s	transkrypcja Whisperem #7
nagrania/wspomnienia/relacja-5.mp3	PRAWDZIWE relacji świadków (audio, 2,5–6 min): Banasiak, Kozłowska-Siejak, Biegański, Majchrzak, Lamęcki	transkrypcja #7 · anonimizacja #9 · streszczenie #2
relacje-swiadkow/relacja-audio-1956-06-29.mp3	syntetyczne nagranie (gdzieby brakło audio) — już zastąpiony realnym nagraniem w nagrania/	transkrypcja #7
opracowania/opracowanie-pl-fragmenty-1956-06-29.md	niepójne przypisy, literówki w nazwiskach, brak DOI	streszczenie #2 · porównanie #5 · bibliografia #12 · RAG #1
opracowania/study-en-excerpt.md	fragment EN do tłumaczenia / porównania	tłumaczenie #3 · porównanie #5
wycinki-prasowe/wycinek-0*-ocr.txt	tekst „po OCR”: rn m, l l, O 0, ucięte dzielenie wyrazów, gubione diakrytyki	OCR #13 · ekstrakcja encji #10 · tematy #11
wycinki-prasowe/wycinek-skan.pdf	obraz „skanu” wycinka (do OCR od zera)	OCR #13

Plik / katalog	Na czym polega „brud”	Czego uczy
statystyki/ofiary-i-aresztowania	nieczyste separatory, przecinki dziesiętne, nagłówki PL+EN, duplikaty, różne formaty dat	strukturyzacja · ekstrakcja #10
oddzialy-wojskowe/sily-wojskowe	ta sama jednostka pisana na 3 sposoby, wolny tekst + pół-tabela	ekstrakcja/normalizacja encji #10
fotografie/*.jpg	prawdziwe zdjęcia PD z Wikimedia (np. „ŻĄDAMY CHLEBA”) — nie rekonstrukcje	OCR z obrazu #13 · kontekst · ekstrakcja encji #10
fotografie/aipn-poznanpl/*.jpg	48 prawdziwych zdjęć (AIPN / poznan.pl) z opisami: demonstracje, walki, ofiary (Romek Strzałkowski), procesy, zatrzymania	kontekst · ekstrakcja encji #10 · OCR wycinków #13
ZRODLA.md · GDZIE-SZUKAC.md	meta: zweryfikowane źródła / gdzie szukać	—

Sugerowana kolejność sprzątnia (ścieżka warsztatu)

1. **OCR** — wycinek-01-ocr.txt: poproś model o rekonstrukcję poprawnego tekstu.
2. **Strukturyzacja** — ofiary-i-aresztowania.csv: do czystej, jednolitej tabeli (jeden separator, jeden format daty).
3. **Normalizacja encji** — sily-wojskowe.md: ujednolicić nazwy jednostek, wyciągnij listę: jednostka, dowódca, sprzęt.
4. **Transkrypcja + streszczenie** — relacja-01-surowa.txt: interpunkcja, podział na mówców, 5-zdaniowe streszczenie.
5. **Anonimizacja** — ta sama relacja: zamień dane osobowe na pseudonimy/placeholdery (RODO).
6. **RAG / porównanie** — opracowania/: zadaj pytanie do obu tekstów i porównaj ujęcia (np. liczba ofiar 57 vs 58).

Uwaga o liczbach

Źródła podają **57 lub 58 ofiar śmiertelnych** (śledztwo IPN z 2009 r.: 58 — w tym 50 cywilów, 4 żołnierzy, 1 milicjant, 3 funkcjonariuszy UB; co najmniej 573 rannych, min. 746 zatrzymanych i aresztowanych — dosłowne cytaty i URL w ZRODLA.md sek. C). Ta rozbieżność jest w korpusie **celowa** — to dobry materiał do ćwiczenia „porównaj, co mówią różne źródła” (#5).