

Magdalena Zawisławska

Uniwersytet Warszawski

zawisla@uw.edu.pl

ORCID: 0000-0003-4525-4509

## MIKROKORPUS BŁĘDÓW JĘZYKOWYCH WE WSPÓŁCZESNEJ POLSZCZYŹNIE

### 1. WSTĘP

Mikrokorpus błędów językowych powstał w ramach projektu *Opracowanie technologii redukcji błędów typu false positives wykrywanych przez algorytmy korekty tekstów na potrzeby wykonania platformy do automatycznej korekty kontekstowej dokumentów z elementami redakcji, wykorzystującej metody uczenia maszynowego i modele językowe*<sup>1</sup>. Mikrokorpus błędów był jednym z zasobów<sup>2</sup> wykorzystywanych w projekcie do trenowania modeli językowych opartych na głębokich sieciach neuronowych. Głównym celem było stworzenie modelu korekty błędów dla współczesnego języka polskiego. W tym artykule skupiam się wyłącznie na opisie korpusu, ponieważ w moim przekonaniu jest on zasobem pionierskim, który może być interesujący dla językoznawców zajmujących się poprawnością języka polskiego oraz dla badaczy, którzy chcieliby podobne korpusy budować w przyszłości.

Do tej pory nie stworzono analogicznego zasobu językowego dla języka polskiego. Dotychczas przedmiotem zainteresowania badaczy były najczęściej błędy w tekstach pisanych przez uczniów – przykładowo omawiali takie kwestie Saloni (1971) czy Dobkowska i Hącia (2012) – lub błędy popełniane przez obcokrajowców uczących się języka polskiego (Dąbrowska 2004; Dąbrowska, Pasięka 2014; Skura 2013, 2022). W ramach badań glottodydaktycznych powstały dwa korpusy uczniowskie – korpus DAMA zawierający fragmenty prac cudzoziemców (Dąbrowska, Pasięka 2014) oraz najnowszy Korpus Uczniowski Języka Polskiego PoLko<sup>3</sup> (Kaczmarska, Za-

---

<sup>1</sup> Projekt NBCiR nr POIR.01.01.01-00-1790/20-00, realizowany przez Lingventę sp. z o.o. w latach 2021–2023.

<sup>2</sup> Drugim zasobem był korpus błędnych tekstów stworzony automatycznie przez generator błędów. Oba zasoby były wykorzystane do trenowania modeli językowych w projekcie.

<sup>3</sup> <https://jakobson.korpus.cz/teitok/polko/>.

sina 2020) zawierający prace pisemne osób uczących się języka polskiego z różnych krajów, w różnym wieku i na różnym poziomie kompetencji. Własne doświadczenia korektora wykorzystał Mackiewicz (2018) w książce *497 błędów. Jak nie zbłądzić w zawitościach polszczyzny*. Z kolei autorzy najnowszego *Słownika błędów językowych* (Rudnicka i in. 2020) deklarują, że:

Do opisu słownikowego kwalifikowaliśmy błędy często popełniane: zarówno te mające ogólnie dużą liczbę wystąpień w tekstach, jak i te, które wprowadzie dużej liczby wystąpień nie mają (bo wyraz czy wyrażenie są rzadko używane), ale przeważają liczebnie nad formami poprawnymi (Rudnicka i in. 2020: 7).

Niestety nie wyjaśnia się czytelnikom, co dla autorów oznacza „częsty błąd”, ponieważ autorzy ani we wprowadzeniu, ani przy hasłach nie podają żadnych danych ilościowych. Jako źródła referencyjne dla omawianego słownika autorzy wskazują słowniki (*Wielki słownik języka polskiego*, *Nowy/Wielki słownik poprawnej polszczyzny*, *Inny słownik języka polskiego*, *Uniwersalny słownik języka polskiego*, *Słownik gramatyczny języka polskiego*, *DobrySłownik.pl*), nowe poradniki językowe oraz opinie poradni językowych (np. poradni PWN-u). Do weryfikacji frekwencji omawianych form i kolokacji wykorzystywany był m.in. Narodowy Korpus Języka Polskiego i HASK Pelcra (Rudnicka i in. 2020: 7). Należy tu podkreślić, że zarówno NKJP, jak i HASK (który wykorzystuje dane z NKJP) są już bazami bardzo nieaktualnymi<sup>4</sup> (do NKJP nowe teksty przestały być włączane w 2010 roku). Nie jest jasne, czemu autorzy jako źródła frekwencji nie wykorzystali cały czas aktualizowanego korpusu MONCO<sup>5</sup>.

## 2. STRUKTURA KORPUSU

Mikrokorpus błędów językowych został zbudowany z próbek losowo wybranych tekstów reprezentujących różne gatunki internetowe oraz tekstów literackich. Zawiera łącznie:

- 3027 tekstów (5371835 znaków),
- 942084 tokenów ze stopwordami i interpunkcją,
- 780073 tokenów ze stopwordami i bez interpunkcji,
- 491412 tokenów bez stopwordów i interpunkcji.

Ponieważ w przypadku korpusu błędów dość trudno mówić o zrównoważeniu (dla tego też używam w odniesieniu do niego określenia *mikrokorpus*, inaczej *korpus rzeźmiślniczny* – czyli korpus, który z założenia nie jest zrównoważony, ponieważ powstał w innych celach niż prezentacja stanu języka z danego okresu), materiały zostały dobrane tak, aby autorzy mogli wykorzystać korektor błędów do analizy poprawno-

<sup>4</sup> Teraz jest jeszcze dostępny najnowszy Korpus Współczesnego Języka Polskiego (kwjp.pl).

<sup>5</sup> <http://monco.frazeo.pl>.

ściowej swoich tekstów. Nie ma zatem w korpusie żadnych tekstów mówionych czy bardzo swobodnych (i zazwyczaj niegramatycznych) wypowiedzi internautów na popularnych forach. Na 3027 tekstów w korpusie składają się 964 fragmenty współczesnych tekstów literackich (w stanie przed korektą i redakcją) oraz 2637 najnowszych tekstów internetowych (z lat 2021–2022). Teksty internetowe obejmują możliwie jak najszerszy zakres różnych gatunków – pochodzą z portali internetowych (gazeta.pl, onet.pl), gazet cyfrowych (np. gazeta.policja.pl, gazetalesna.pl, se.pl), blogów prywatnych i blogów firmowych (np. toyotabank.pl, bankowymokiem.pl, ksiazkana-prezent.blogspot.com), stron internetowych instytucji państwowych (np. gov.pl, parp.gov.pl). Niewielką częścią korpusu (ok. 5%) są teksty z tematycznych forów internetowych (np. historycy.org, forum.gazeta.pl), które w założeniu miały reprezentować język bardziej swobodny, w którym istniało większe prawdopodobieństwo znalezienia nowych typów błędów. Autorami tekstów zgromadzonych w korpusie są zatem najczęściej osoby z wyższym lub średnim wykształceniem. Część tekstów (np. z gazet internetowych) przeszła również redakcję lub korektę. Można zatem założyć, że korpus prezentuje błędy popełniane przez bardziej wykształconą i prawdopodobnie bardziej świadomą językowo część społeczeństwa.

### 3. ANOTACJA KORPUSU

Wszystkie próbki pobrane z tekstów literackich oraz tekstów internetowych liczyły 1800 znaków. Przed udostępnieniem tekstów anotatorom teksty były sprawdzane i przygotowywane do dalszej obróbki (np. przycinane do wymaganej liczby znaków czy czyszczone ze zbędnych elementów, typu urwane zdania, odsyłacze do stron internetowych). Próbki całkowicie nienadające się do anotacji były usuwane.

Anotacja odbywała się w narzędziu AnnTool, specjalnie dostosowanym do potrzeb projektu. Po otwarciu przez anotatora nowego zadania próbka tekstu była widoczna w oknie narzędzia. Kiedy anotator zauważył błąd, zaznaczał wyraz, frazę lub całe zdanie (w przypadku błędów szyku) i w nowym oknie dokonywał ich korekty oraz zaznaczał odpowiednią kategorię błędów z wyświetlanej listy. Wszyscy anotatorzy byli polonistami (z tytułami magistra, doktora lub doktora habilitowanego), z wieloletnim doświadczeniem w redakcji i korekcie tekstów.

Po zakończeniu anotacji wszystkich próbek została przeprowadzona superanotacja korpusu. Dwóch superanotatorów ponownie czytało wszystkie zanotowane wcześniej teksty i wprowadziło zmiany, jeśli było to niezbędne. Wprowadzenie superanotacji wynikało z konieczności ujednoczenia korpusu w celu trenowania sieci neuronowych.

Istotnym zadaniem było sporządzenie listy błędów językowych. Kategorie błędów wyróżniane w poradnikach lub słownikach (np. Markowski 2012) były nazbyt

ogólne. Lista błędów przygotowana na potrzeby anotacji korpusu jest bardziej rozbudowana i szczegółowa, nie ogranicza się do typowych, ale zbyt obszernych kategorii, jak błędy interpunkcyjne, ortograficzne, fleksyjne, składniowe i stylistyczne.

W narzędziu AnnTool na liście błędów wyodrębnione zostały następujące grupy i podgrupy:

**1. Błąd w zapisie (literówka)**

**2. Błąd interpunkcyjny**

- 2.1. brak znaku
- 2.2. zbędny znak
- 2.3. niewłaściwy znak
- 2.4. niewłaściwa kolejność znaków

**3. Błąd ortograficzny**

3.1. pisownia łączna i rozdzielna

- błędna pisownia łączna
- błędna pisownia rozdzielna

3.2. mała/wielka litera

- zbędna wielka litera
- zbędna mała litera
- błędna wielka litera – pozycja w zdaniu<sup>6</sup>
- błędna wielka litera – nazwa pospolita<sup>7</sup>
- błędna wielka litera – nazwa własna
- błędna mała litera – pozycja w zdaniu
- błędna mała litera – nazwa pospolita
- błędna mała litera – nazwa własna

**4. Błąd leksykalny**

- 4.1. pleonazm
- 4.2. tautologia
- 4.3. zbędny wyraz
- 4.4. brak wyrazu
- 4.5. błędny wyraz

**5. Błąd frazeologiczny**

- 5.1. błędny element
- 5.2. nadmiarowy element
- 5.3. brak elementu
- 5.4. kontaminacja

---

<sup>6</sup> Dotyczy sytuacji, gdy autor błędnie używa wielkiej litery nie na początku zdania i nie jest to uzasadnione innymi regułami ortograficznymi, np. *Opakowania na torty – Torty potrafią dużo ważyć.*

<sup>7</sup> Dotyczy sytuacji, gdy autor zapisuje wyraz wielką literą, choć nie jest to nazwa własna, tylko nazwa pospolita, np. *Został powołany przez Jezusa i wybrany na Apostoła.*

## 6. Błąd fleksyjny

### 7. Błąd składniowy

- 7.1. naruszona łączliwość składniowa
- 7.2. błędny szyk
- 7.3. błędne użycie imiesłowowego równoważnika zdania
  - 7.3.1. niezgodność podmiotów
  - 7.3.2. brak równoczesności czynności
  - 7.3.3. inny
- 7.4. inny błąd składniowy

### 8. Błąd składu

- 8.1. błędne przeniesienie
- 8.2. inny

### 9. Zbędny fragment tekstu

### 10. Inny

### 11. Kursywa

Punkty 8–11 nie były formalnie błędami językowymi, ale zostały uwzględnione ze względu na potrzeby projektu. W kolejnej części artykułu skupię się jedynie na omówieniu błędów stricte językowych (z pominięciem również błędów zapisu).

Do narzędzia AnnTool zostały ponadto dodane dwa znaczniki: *konsekwencja zmiany* oraz *ignoruj*. Pierwszy z tych znaczników dotyczył konieczności wprowadzania zmian np. po poprawieniu i zmianie błędnego przyimka. Nowy przyimek wymagał innych form fleksyjnych, ale tej zmiany anotator nie mógł już oznaczyć jako błąd, zatem stosował znacznik *konsekwencja zmiany*, dzięki czemu model trenowany na korpusie uczył się, że jest to poprawka, a nie błąd. Drugi ze znaczników – *ignoruj* – dotyczył np. form archaicznych, gwarowych lub celowych błędów stosowanych przez autorów np. w celu stylizacji utworu.

Praca nad ujednoczeniem korpusu (superanotacja) pokazała kilka kwestii, które mogą być interesujące dla badaczy zajmujących się językoznawstwem normatywnym. Po pierwsze, mimo kierunkowego wykształcenia anotatorów oraz doświadczenia w pracy nad tekstem można było zauważyć dość znaczące rozbieżności w oznaczaniu błędów. Zdarzało się, że anotator uznawał za błąd coś, co błędem nie było (np. zbyt długie zdania lub szyk przestawny, którego w żaden sposób nie można było potraktować jako niepoprawny). Inne błędy z kolei nie były oznaczane (np. niepoprawne używanie przyimka *poprzez* czy czasownika *posiadać*). Również często zdarzało się, że błędy były zaliczane do niewłaściwej kategorii (np. błędy składniowe były uznawane za fleksyjne). W moim przekonaniu wynika to niekoniecznie z niestaranności anotacji, ale raczej z niejasności związanych z samym pojęciem błędu językowego i rozmyciem jego kategorii. W *Wielkim słowniku poprawnej polszczyzny PWN* (s. 1621) błąd językowy jest zdefiniowany następująco:

[B]łąd językowy, innowacja językowa w tekście albo systemie języka, którą trudno uzasadnić, gdyż nie wyraża nowych treści, nie przekazuje w nowy sposób ekspresji, nie usprawnia porozumiewania się itd. [...] Jako odstępstwo od dotychczasowych zwyczajów i przyjętych norm językowych, zarówno poziomu normy wzorcowej, jak i normy użytkowej, błąd razi osoby świadomie posługujące się językiem.

Nie jest jasne, jak należy oceniać, czy dana innowacja jest uzasadniona – nie ma żadnych kryteriów, wedle których taka ocena miałaby przebiegać. Zatem zwykle przy sprawdzaniu tekstów redaktorzy czy korektorzy wykorzystują dostępne słowniki czy poradniki językowe. Wychwytywane więc będą jedynie te negatywnie ocenione innowacje, które są już utrwalone w istniejących zasobach poprawnościowych.

Na ogromne rozbieżności w ocenie błędów przez osoby sprawdzające szkolne prace egzaminacyjne zwróciły uwagę Dobkowska i Hącia (2012). Wykazały one zaskakująco duże rozbieżności między egzaminatorami oceniającymi te same prace uczniowskie: „Rozbieżności w przypadku liczby błędów językowych sięgały niekiedy 9, 10 a nawet 152, co świadczy o tym, że egzaminatorzy nie zauważali błędów lub z jakichś powodów nie wliczali ich do wykazywanej kwoty” (Dobkowska, Hącia 2012: 94). Równie problematyczne jest zaliczenie błędu do określonej kategorii. Pisał już o tym Saloni (1971: 21):

Zebrane błędy językowe podzieliłem na kilka grup według tego, jakich dziedzin języka dotyczą. Jest to klasyfikacja robocza, celem jej jest tylko ułatwienie przeglądu. Szczególnie zbyt wiele było tu wypadków pogranicznych, takich, gdy wykołajenie wiąże się z dwoma działami gramatyki lub gdy stoi na pograniczu gramatyki i leksyki. Charakterystycznym przykładem mogą tu być wykołajenia rekcji, które można zaliczyć zarówno do gramatyki, jak i do słownictwa. Czysto umowne jest także zaliczanie błędów w użyciu przyimków i spójników do działu składni, a błędów w użyciu zaimków – do słownictwa. Trudne było rozgraniczenie błędów stylistycznych i właściwych błędów językowych.

Wydaje się, że najbardziej mglistą kategorią są błędy leksykalne i frazeologiczne. Nie zawsze da się ocenić, czy faktycznie mamy do czynienia np. z naruszeniem łączliwości leksykalnej, ponieważ dostępne źródła językowe są bardzo ograniczone. W zasadzie jedynym współczesnym słownikiem języka polskiego, który wprost podaje łączliwość wyrazów, jest *Wielki słownik języka polskiego*, natomiast *Słownik dobrego stylu, czyli wyrazy, które się lubią* pod redakcją Bańki jest niestety dość niewielki (obejmuje 5000 rzeczowników, czasowników i przymiotników). W przypadku frazeologizmów oddzielenie błędu od wariantu również nie jest zadaniem prostym. Ponadto leksyka zmienia się w tak błyskawicznym tempie, że nie ma żadnych źródeł poprawnościowych, które byłyby w stanie na bieżąco oceniać nowe zapożyczenia czy neologizmy. Nie ułatwia oceny fakt, że sama norma podlega zmianie, na co wskazuje fragment definicji błędu językowego z *Encyklopedii szkolnej WSiP. Nauka o języku* (s. 34), por.:

„Ewolucja normy językowej, pozostającej pod stałym naciskiem zwyczaju językowego, powoduje, że pewne formy językowe, które były błędami (nie mieściły się w normie), z czasem mogą zostać uznane za dopuszczalne, a wreszcie za równoprawne”.

## 4. STATYSTYKI BŁĘDÓW W KORPUSIE

Wydaje się, że najciekawsze dla językoznawcy-polonisty są statystyki błędów w korpusie. Całkowita liczba błędów językowych wynosi 33 959. Jeśli odliczy się błędy zapisu (ewidentne literówki – 4369 przypadków), pozostanie 29 590 błędów. Średnio zatem na jeden tekst w korpusie przypada 11 błędów. Najwięcej autorzy popełniali błędów interpunkcyjnych – 20 848 (70,5% wszystkich błędów językowych). Znacznie mniej było błędów ortograficznych – 3239 (11%). Błędów składniowych autorzy tekstów zgromadzonych w korpusie popełnili w sumie 2705 (9,1%). Nieco mniej było błędów leksykalnych – 2523 (8,5%). Bardzo rzadkie były błędy fleksyjne – zaledwie 185 (0,6%) oraz frazeologiczne – 90 (0,3%).

### 4.1. Błędy interpunkcyjne

Problem z interpunkcją we współczesnych tekstach został już zauważony przez wielu badaczy. Mimo że dysponujemy licznymi wydawnictwami poprawnościowymi, słownikami i poradnikami, błędy interpunkcyjne należą do najczęstszych typów błędów popełnianych przez użytkowników języka polskiego, i to niezależnie od wykształcenia (por. Łuczyński 2016: 43). Na przykład jednym z typowych błędów interpunkcyjnych (co potwierdza korpus) jest rozdzielanie przecinkiem grup składniowych. Jak piszą Sikora i Rak (2011: 192):

Do najczęstszych uchybień w przestankowaniu świadczących o rosnącej roli czynników prozodycznych należy na przykład oddzielanie przecinkiem rozbudowanego przydawkami podmiotu od orzeczenia albo też określań przyczasownikowych usytuowanych na początku zdania (zwłaszcza okoliczników).

Autorzy uważają, że to wpływ prozodii, ale być może wydzielenie okoliczników znajdujących się na początku zdania jest wynikiem wpływu języka angielskiego. Korpus pokazuje, że liczne błędy interpunkcyjne (pomijanie przecinków) występują w zdaniach złożonych, np. autorzy często zapominają, że zdania składowe wtrącone muszą być wydzielone przecinkami obustronnie, np.

- (1) Przykładem takiej firmy, która w działaniu „Programy Akceleracyjne” (POIR), otrzymała grant na rozwój swojego produktu jest Roboticon<sup>8</sup>.

<sup>8</sup> We wszystkich przykładach w artykule została zachowana oryginalna pisownia.

Niestety, stosowanie wyłącznie zasad ogólnych (przed *że* stawiamy przecinek) generuje również bardzo liczne błędy, gdzie przecinkiem rozdzielane są spójniki złożone, np.: *jako że, mimo że, tym bardziej że, a także, a że, chyba że, chyba żeby, czyli że, dlatego że, dzięki temu / na skutek tego / pomimo że, szczególnie że, tyle że, tym bardziej że, zwłaszcza że*. Korektory wbudowane w edytory tekstów (np. Word) mogą również błędnie sugerować konieczność postawienia w takich miejscach przecinków.

## 4.2. Błędy ortograficzne

W przypadku błędów ortograficznych stosunkowo zaskakujące jest, że najwięcej problemów autorzy mają z poprawnym użyciem małej i wielkiej litery (2460 błędów w korpusie). Nadużywana jest wielka litera w wyrazach oznaczających np. funkcje i zawody, np. *Minister, Naczelnik, Pielęgniarka, Pełnomocnik, Przewodniczący, Redaktor, Rolnik, Prezes*<sup>9</sup>. Z nieznanymi powodami autorzy używają wielkiej litery w zapisie takich jednostek leksykalnych, jak *Policja, Redakcja, Pomnik, Raport, Program, Rysunek, Sanskryt*. Wyraźnie pod wpływem języka angielskiego autorzy mają tendencję do zapisywania wielkimi literami wszystkich wyrazów składających się na tytuły książek.

Błędów w zakresie pisowni łącznej lub rozdzielnej jest znacznie mniej – tylko 377. Widać, że autorzy tekstów mają najczęściej problem z zapisem takich wyrazów, jak *\*nie ważne* (zamiast: *nieważne*), *\*nie raz* (zamiast: *nieraz*), *\*po za* (zamiast: *poza*), *\*po krótcie* (zamiast: *po krótko*), *\*z resztą* (zamiast: *zresztą*), *\*nie trudno* (zamiast: *nietrudno*), *\*co raz* (zamiast: *coraz*), *\*na prawdę* (zamiast: *naprawdę*), *\*na raz* (zamiast: *naraz*), *\*po niżej* (zamiast: *poniżej*), *\*pod czas* (zamiast: *podczas*). Niepoprawną pisownię można zaobserwować bardzo często w formach z częstką *-by*, np. *\*co by, \*chciała bym, \*byle bym, \*dawało by, \*mogli byśmy, \*musiała bym, \*pozostał by*. Duże problemy sprawia również zapis partykuły *nie*, np. *\*nie bagatelnie* (zamiast: *niebagatelnie*), *\*nie całe* (zamiast: *niecałe*), *\*nie dobrze* (zamiast: *niedobrze*), *\*nie jeden* (zamiast: *niejeden*), *\*nie mały* (zamiast: *niemały*), *\*nie planowany* (zamiast: *nieplanowany*), *\*nie przypadkowo* (zamiast: *nieprzypadkowo*). Z kolei łącznie zapisywane są wyrazy: *\*codzień, \*co najmniej, \*nawet, \*popołudniu, \*prosto, \*poza, \*wogóle*. Warto podkreślić, że korektory regułowe<sup>10</sup> (np. w edytorze Word) nie są w stanie takich błędów wychwycić.

<sup>9</sup> Dodajmy, że nie można w tych kontekstach uzasadnić użycia wielkiej litery względami uczuciowymi i grzecznościowymi.

<sup>10</sup> Systemy regułowe wykorzystują po prostu zestaw reguł zaczerpniętych np. ze słowników poprawnościowych czy interpunkcyjnych i wprowadzonych do programu. W zasadzie wszystkie dostępne aplikacje (LanguageTool, iKorektor, KorektorTekstu, moduły korekty w programach Microsoft Word, Open Office, Google Docs) wykorzystują systemy regułowe lub automaty skończone.

### 4.3. Błędy składniowe

Jeśli chodzi o błędy składniowe w korpusie, najwięcej z nich (1573) dotyczy naruszenia łączliwości składniowej, por.:

- (2) \***Towarem, o który** powinniśmy szczególnie zadbać i **wyeksponować** jest pieczywo.
- (3) \*Szabłok to kwintesencja codzienności, bo zawiera wszystko to, **czym obfitowały** Kujawy i Pomorze.
- (4) \***Brytfanny emaliowane** to **naczynie, które** sprawdzi się w każdej kuchni.
- (5) Wielbiciele tej niejednoznacznej i kapryśnej postaci, którą Hiddleston ożywił mocą swojego talentu, na pewno są **usatysfakcjonowani z takiego obrotu spraw**.

W przykładzie (2) mamy do czynienia z typowym skrótem składniowym: *zadbać* (o kogo, co?), ale *wyeksponować* (kogo, co?). W zdaniu (3) naruszone zostały wymagania składniowe czasownika *obfitować* (w kogo, co?) (WSJP). W kolejnym przykładzie (4) autor nie dopasował do podmiotu *brytfanny* wartości kategorii liczby wyrażen podrzędnych (*to naczynia, które sprawdzą się...*). W ostatnim zdaniu (5) autor nie uwzględnił wymagań składniowych przymiotnika *usatysfakcjonowany* (czym?) (WSJP).

Drugi typ błędów składniowych o stosunkowo wysokiej frekwencji (620) to naruszenia szyku, por.:

- (6) Chroni skórę przed rozstępami i wzmacnia ją.
- (7) Głównie kamienne groby z czytelnymi tablicami nie są rzadkością.
- (8) Na chwilę obecną, była to jedyna polska grupa, która kiedykolwiek wzięła udział w tak ważnym dla świata kulinarnego turnieju.

Nieco rzadziej (473 przykłady) autorzy tekstów zgromadzonych w korpusie popełniali błędy związane z użyciem imiesłowowych równoważników zdania. Więcej problemów sprawiała równoczesność czynności (372 przykłady) niż zgodność podmiotów (101 przykładów), por.:

- (9) Będąc z kimś, kto nas pociąga i na kim możemy bezwarunkowo polegać, życiowe perturbacje są łagodniejsze, ma się więcej odwagi do podbijania świata, a sukcesy bardziej cieszą.
- (10) Ojciec przyszłego rosyjskiego króla popu, Bedros Filippowicz, wywodził się za to z ormiańskiej rodziny, która uciekając przed tureckimi pogromami, osiadła w Warnie, zmieniając nazwisko na Kirkorow.
- (11) Zliczając głosy jurorskie wraz z wynikami głosowania SMS, zwycięzcą tegorocznej Eurowizji 2022 została Ukraina!

W przykładzie (9) mamy do czynienia z niezgodnością podmiotów – w zdaniu głównym podmiotem są *życiowe perturbacje*, natomiast podmiotem w zdaniu prze-

kształconym z imiesłowowego równoważnika zdania (*będąc z kimś*) podmiotem byłby zaimek *my*. W przykładzie (10) mamy do czynienia z brakiem równoczesności czynności. Zdarzenia ze zdania (10) nie mogły się rozegrać jednocześnie – ormiańska rodzina najpierw uciekła, potem osiadła w Warnie i zmieniła nazwisko. Z kolei w przykładzie (11) nie ma ani zgodności podmiotów (kto inny liczy głosy, kto inny – Ukraina – zostaje zwycięzcą), ani równoczesności czynności (najpierw zliczono głosy, a dopiero na podstawie tego wyniku ogłoszono, kto został wygrał).

#### 4.4. Błędy leksykalne

Nieco rzadsze w korpusie są błędy leksykalne (2523 przykłady), z czego zdecydowana większość to użycie błędnego wyrazu w danym kontekście (1191 przykładów), pomijanie przez autorów wyrazów, które powinny się znaleźć w danym zdaniu (659 przykładów), lub używanie wyrazów zbędnych (617 przykładów), zdecydowanie najrzadsze zaś były konstrukcje redundantne – pleonazmy i tautologie (tylko 56 przykładów), por.:

- (12) Według dziennikarki Julii Davis, która na bieżąco **monituje** rosyjskie media propagandowe i pokazuje na Twitterze najciekawsze fragmenty programów, to wyraźny znak, że Rosja wycofuje się z planu ogłoszenia „wygranej” w Dzień Zwycięstwa, przypadający na 9 maja.
- (13) Autor książek oraz programów kulinarnych. Studiował w Westminster Catering College, we Francji **pobierał również praktyki**.
- (14) Poniżej przedstawiamy najważniejsze informacje, które powinien **wiedzieć** absolutnie każdy przed zakupem takiego urządzenia.
- (15) Pod tym względem kasze nie mają konkurencji, zawierają one zdecydowanie więcej wartości odżywczych niż taka sama **dawka** ziemniaków, makaronu czy nawet ryżu.

W przykładzie (12) autor pomylił podobnie brzmiące słowa *monitować* i *monitorować*. W zdaniu (13) naruszona została łączliwość wyrazu *praktyka* (praktyki można odbywać, ale nie można ich pobierać)<sup>11</sup>. Autor przykładu (14) używa czasownika *wiedzieć* zamiast poprawnego w tym kontekście słowa *znać*. W zdaniu (15) zamiast wyrazu *dawka* powinien być użyty rzeczownik *porcja*.

W korpusie można zauważyć, że nadużywane są wyrazy typu *dedykowany* (zamiast *przeznaczony*), *implementacja* (zamiast *wprowadzenie*), *aktualnie* (zamiast *obecnie*), *aplikacja* (zamiast *wniosek*), *artykułować* (zamiast *wyrazić*), *design* (zamiast *wygląd*). Niewątpliwie ma na to wpływ język angielski. Regularnie są również mylone przyimki *dzięki* i *przez* oraz nadużywany jest przyimek *poprzez* traktowany jako synonim *przez*.

<sup>11</sup> Kategoryzacja tego błędu może być dyskusyjna – można go też uznać za błąd frazeologiczny.

#### 4.5. Błędy fleksyjne

Błędy fleksyjne są w korpusie bardzo nieliczne (jedynie 186 przykładów). Dotyczy to zarówno odmiany nazw własnych, jak i pospolitych, por.:

- (16) Nazywają pana perłą niepokornej inteligencji warszawskiej, tak właśnie jest – powiedział prezydent Andrzej Duda zwracając się do Antoniego **Libin-Libery** po tym, jak odznaczył go Orderem Orła Białego.
- (17) Co może publikować firma oferująca przewóz **ludzkiego tłuszczu**?
- (18) Z doniesień medialnych wynika, że w poniedziałek 23 sierpnia do **Usnarzu Górnego** dotarła duża grupa funkcjonariuszy policji z Podlasia.

Zwraca uwagę nieodmienianie nazw własnych (głównie obcych, ale też polskich), np. *Barclays* zam. *Barclaysa*, *Bellemare* zam. *Bellemare’a*, *Berthe* zam. *Berthe’owi*, *Hurricane* zam. *Hurricane’y*, *Pierre* zam. *Pierre’a*, *Radczenko*<sup>12</sup> zam. *Radczenką*, *Sachajko* zam. *Sachajką*, *Wawrzeczko* zam. *Wawrzeczki*, *Ziobro* zam. *Ziobry*.

#### 4.6. Błędy frazeologiczne

W korpusie najrzadsze są błędy frazeologiczne (tylko 90 przykładów). Najczęściej (62 przykłady) autorzy używali w związkach frazeologicznych jakiegoś błędnego elementu (np. *najstabszy element* zamiast *najstabsze ogniwo*). Znacznie rzadziej dodawali do związków jakiś element nadmiarowy, np. *jak na wyciągniętej dłoni*. Najrzadziej w przykładach autorzy pomijali jakiś element idiomu (8 przykładów) lub tworzyli kontaminacje idiomów (również 8 przykładów), np. *kolej losu* (połączenie związków frazeologicznych *koleje losu* i *kolej rzeczy*), por.:

- (19) Problem ten można rozwiązać różnie, niektórzy zwyczajnie odpuszczają (NIE POLECAM), inni kombinują coś na tej **Bogu winnej** kartce aż do skutku.
- (20) Ale mimo mojego strachu i wątpliwości, ani razu nie pomyślałam: „*nie wiem, czy to na pewno dobre mieszkanie*” i **uważam to za dobrą monetę**.
- (21) Dobór zegarka, perfum, odpowiednie buty i fryzura **stanowią kropkę nad i** naszej stylizacji.

W przykładzie (19) w idiomie brakuje elementu *ducha* (*Bogu ducha winnej*). W zdaniu (20) poprawna forma frazeologizmu powinna brzmieć *brać/wziąć coś za dobrą monetę*. W przykładzie (21) autor nie znał ani poprawnej formy, ani znaczenia frazeologizmu *stawiać/postawić kropkę nad i* ‘ktoś zakończył coś ostatecznie, w sposób niepozostawiający wątpliwości’ (WSJP).

---

<sup>12</sup> Brak odmiany nazwisk męskich zakończonych na -o (np. *Bańko*, *Radczenko*, *Ziobro*) jest dopuszczalny w normie potocznej, jeśli odmienione jest imię, jednak norma wzorcowa nie przewiduje tu wyjątków. W anotacji korpusu zastosowana była norma wzorcowa.

## 5. PODSUMOWANIE

Pierwszy korpus błędów we współczesnej polszczyźnie jest niewątpliwie niewielki i nie można uznać go za zasób zrównoważony. Wynika to przede wszystkim z założeń projektu, w ramach którego powstał. Wydaje się jednak, że na podstawie jego tworzenia oraz statystyk da się wysnuć kilka istotnych wniosków.

Po pierwsze, jeśli weźmiemy pod uwagę główny cel projektu, czyli trenowanie modeli językowych, możemy stwierdzić, że korpus jest niewątpliwie zbyt ograniczony – zawiera za mało błędów i są one niewystarczająco zróżnicowane. Dlatego poza anotowanym ręcznie korpusem w ramach projektu powstał korpus błędów utworzony sztucznie – za pomocą generatora błędów. Jednocześnie jednak korpus złożony z autentycznych, niepreparowanych tekstów, z których przecież spora część przeszła redakcję i korektę, nadal zawiera bardzo dużo błędów, zwłaszcza interpunkcyjnych, których najwyraźniej nie wychwytyją dostępne narzędzia do korekty (tzw. regułowe). Potwierdzają to testy dostępnych aplikacji do korekty przeprowadzone przez Witkowską (2021: 110–111). Zauważyła ona, że systemy regułowe radzą sobie wyłącznie z rażącymi usterkami ortograficznymi, fleksyjnymi czy leksykalnymi. Są w stanie wychwycić proste błędy interpunkcyjne. Nie radzą sobie za to z bardziej skomplikowanymi przypadkami błędów fleksyjnych, składniowych czy interpunkcyjnych, nie wychwycają również błędów wynikających z paronimii (np. *monitować* zamiast *monitorować*). Potwierdza to konieczność trenowania i doskonalenia modeli językowych opartych na sieciach neuronowych na możliwie jak najbardziej rozbudowanych korpusach błędów.

Po drugie, mikrokorpus błędów w jakimś (choć zapewne ograniczonym) zakresie pokazuje faktyczną znajomość normy językowej wśród lepiej wykształconych Polaków. Poradniki językowe i słowniki błędów lub poprawnej polszczyzny w dużej mierze czerpią z zasobów poradni językowych. Należy jednak pamiętać, że osoby zgłaszające do poradni swoje wątpliwości dotyczące poprawności danych form mają świadomość, że mogą one być niezgodne z normą. Korpus natomiast pokazuje błędy popełniane przez użytkowników nieświadomie i chyba nieco lepiej charakteryzuje realne problemy związane z kulturą języka polskiego, które powinny być wzięte pod uwagę przez osoby odpowiedzialne za kształcenie językowe na poziomie szkoły i studiów polonistycznych. Na podstawie danych z korpusu widać na przykład, że autorzy tekstów w nim zebranych nie popełniają w zasadzie błędów fleksyjnych. Na ten fakt zwrócił już uwagę Łuczyński (2015), który próbował ustalić kryteria gramatyczności wypowiedzi. Pisał on, że „[n]iekwestionowane (oczywiście) błędy gramatyczne są w wypowiedziach użytkowników języka polskiego bardzo rzadkie, wobec tego nie mogą być podstawowym miernikiem jakości tekstu” (Łuczyński 2015: 62). Bez wątplenia na wysoką poprawność fleksyjną tekstów w korpusie mają też wpływ powszechnie dostępne regułowe korektory tekstów (które

dość sprawnie proste błędy fleksyjne wychwytyją, por. Witkowska 2021: 110–111) oraz bardzo dobre zasoby dostępne bezpłatnie w Internecie, jak np. *Słownik gramatyczny języka polskiego*<sup>13</sup> lub *Wielki słownik języka polskiego*<sup>14</sup>.

Niepokojąca jest za to liczba błędów interpunkcyjnych, które anotatorzy znajdowali w każdym typie tekstów zgromadzonych w korpusie. Pokazuje to, że interpunkcji nie uczy się w szkole w efektywny sposób – choćby z powodu braku powiązania zasad przestankowania z nauką składni czy też ograniczania się do podawania uczniom tylko zasad ogólnych (np. niestawianie przecinka przed *i* lub stawianie przecinka przed każdym *że*).

Mikrokorpus błędów językowych nie jest zasobem językowym, który spełniałby wszystkie wymagania, jakie stawiane są współczesnym korpusom – nie jest zrównoważony, nie jest wystarczająco obszerny i nie jest dostępny w sieci – wytycza on jedynie nowe ścieżki badawcze. Wydaje się, że warto by kontynuować prace nad podobnymi zasobami językowymi, które byłyby dostępne nieodpłatnie w Internecie. Nowe korpusy błędów powinny bez wątpienia być znacznie większe i obejmować dużo bardziej zróżnicowane teksty.

## Bibliografia

- Bańko, M. 2013. *Słownik dobrego stylu, czyli wyrazy, które się lubią*. Warszawa: PWN.
- Dąbrowska, A. 2004. *Najczęstsze błędy popełniane przez cudzoziemców uczących się języka polskiego jako obcego*. W: *Opisywanie, rozwijanie i testowanie znajomości języka polskiego jako obcego*, red. A. Seretny, W. Martyniuk, E. Lipińska, s. 105–136. Kraków: Universitas.
- Dąbrowska, A., Pasieka, M. 2014. *Badania błędów cudzoziemców prowadzone w Szkole Języka Polskiego i Kultury dla Cudzoziemców UW*. W: *40 lat wrocławskiej glottodydaktyki polonistycznej: Teoria i praktyka*, red. A. Dąbrowska, U. Dobesz, s. 331–342. Wrocław: Oficyna Wydawnicza ATUT.
- Dobkowska, J., Hącia, A. 2012. Ocena poprawności językowej prac egzaminacyjnych uczniów III klasy gimnazjum: wewnętrznojęzykowe przyczyny trudności w ocenie, wyniki zastosowania skali egzaminacyjnej, zalecenia dla systemu egzaminacyjnego. *Edukacja 2*, s. 93–117. Online: <https://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-8b662545-2d96-4407-8184-8935839f2108> [dostęp: 7.02.2025].
- Kaczmarek, E., Zasina, A.J. 2020. Błędy walencyjne w tekstach obcokrajowców uczących się języka polskiego w świetle korpusu PoLko. *Prace Filologiczne 75* (1), s. 197–213. DOI: 10.32798/pf.657.
- Kaleta, R. 2009. Dyskusja nad błędem językowym w wybranych polskich pracach językoznawczych wydanych w latach 1978–2008 (przegląd). *Lingwistyka Stosowana 1*, s. 152–171. Online: <https://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-edf153a9-34d4-4b90-8c4e-885b69d804b3> [dostęp: 7.02.2025].

<sup>13</sup> [sgjp.pl](http://sgjp.pl).

<sup>14</sup> [wsjp.pl](http://wsjp.pl).

- Łuczynski, E. 2015. Jak ustalać poziom gramatyczności wypowiedzi? *Poradnik Językowy* 9, s. 53–67. Online: <https://www.ceeol.com/search/article-detail?id=441989> [dostęp: 7.02.2025].
- Łuczynski, E. 2016. Dlaczego nasza interpunkcja sprawia kłopoty piszącym? *Poradnik Językowy* 4, s. 43–55. Online: <https://www.ceeol.com/search/article-detail?id=419322> [dostęp: 7.02.2025].
- Mackiewicz, Ł. 2018. *497 błędów. Jak nie zbłądzić w zawitościach polszczyzny*. Elbląg: Wyd. Łukasz Mackiewicz.
- Markowski, A. red. 2004. *Wielki słownik poprawnej polszczyzny*. Warszawa: Wydawnictwo Naukowe PWN.
- Markowski, A. red. 2006. *Encyklopedia szkolna WSiP. Nauka o języku*. Warszawa: Wydawnictwo WSiP.
- Markowski, A. 2012. *Wykłady z leksykologii*. Warszawa: PWN.
- Rudnicka, E. i in. 2020. *Słownik błędów językowych. Słowa, zdania, wyrażenia (tworzenie i stosowanie)*. Poznań: Agencja Nomen.
- Saloni, Z. 1971. *Błędy językowe w pracach pisemnych uczniów liceum ogólnokształcącego. Próba analizy językoznawczej*. Warszawa: Państwowe Zakłady Wydawnictw Szkolnych.
- Sikora, K., Rak, M. 2011. Nowe tendencje w interpunkcji – przecinek (na materiale internetowym). *Język Polski* 203, s. 188–194. Online: <https://jzyk-polski.pl/index.php/jp/article/view/914> [dostęp: 7.02.2025].
- Skura, M. 2013. Błędy wynikające z interferencji kulturowej popełniane przez Niemców uczących się języka polskiego jako obcego. *Acta Universitatis Lodzianis. Kształcenie Polonistyczne Cudzoziemców* 20, s. 149–158.
- Skura, M. 2022. *Błędy popełniane przez Niemców uczących się języka polskiego*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego. Online: <http://hdl.handle.net/11089/4511> [dostęp: 7.02.2025].
- Witkowska, K. 2021. O projekcie kontekstowego rozumienia języka pisanego na potrzeby systemu automatycznej poprawy błędów dla języka polskiego. *Prace Językoznawcze* 23 (3), s. 105–113. DOI: 10.31648/pj.6839.

### ***Microcorpus of language errors in contemporary Polish***

#### Summary

The article discusses the process of compiling the first corpus of errors in contemporary Polish and its possible applications. The main goal of the corpus was to use it to train language models based on deep neural networks. However, during the annotation, several problems that may be of interest to linguists (especially those involved in prescriptive linguistics) were identified. Difficulties in annotation suggest that the very concept of error is unclear, as is categorization of language errors. Corpus statistics give an approximate picture of how well educated Poles know the linguistic norm and what types of errors most commonly appear in texts. Such information can be used for the purpose of language education at the school level and in Polish studies.

**Keywords:** Polish – language errors – typology of errors – corpus – linguistic norm.

Adj. Marta Falkowska