Rocznik Przekładoznawczy 20 2025
Studia nad teorią, praktyką i dydaktyką przekładu
ISSN 1896-4362

Lech Zieliński
Uniwersytet Mikołaja Kopernika w Toruniu
lechziel@umk.pl
ORCID: 0000-0003-1265-8873

DYDAKTYKA PRZEKŁADU 2.0: PROJEKT TŁUMACZENIOWY JAKO NARZĘDZIE KSZTAŁTOWANIA NOWEJ ROLI TŁUMACZA WOBEC WYZWAŃ GENEROWANYCH PRZEZ MODELE LLM

DOI: http://dx.doi.org/10.12775/RP.2025.013

Zarys treści: Przedmiotem artykułu jest projekt tłumaczeniowy zrealizowany w ramach zajęć z przekładu prawniczego na kierunku filologia germańska UMK. Głównym celem projektu jest weryfikacja jakości tłumaczeń generowanych przez modele LLM (stan na 2025/2026) w zestawieniu z profesjonalnym przekładem ludzkim oraz uświadomienie studentom konieczności nabycia kompetencji w zakresie eksperckiej postedycji. Autor szczegółowo opisuje warsztatową fazę projektu, obejmującą interdyscyplinarne konsultacje z ekspertami z zakresu językoznawstwa i prawa karnego, oraz analizuje proces podejmowania decyzji translatorskich na przykładzie spornych terminów prawniczych. Na podstawie przeprowadzonych badań stwierdzono, że w obliczu niemal identycznej oceny tekstów maszynowych i ludzkich tradycyjny model kształcenia musi ustąpić miejsca dydaktyce opartej na krytycyzmie technologicznym i inżynierii promptów oraz kształceniu kompetencji w zakresie postedycji.

Słowa kluczowe: tłumaczenie maszynowe (LLM), sztuczna inteligencja, przekład prawniczy, dydaktyka przekładu, postedycja, inżynieria promptów, współpraca ekspercka

Uwagi wstępne

Refleksja naukowa nad zagadnieniami związanymi z projektami tłumaczeniowymi i ich znaczeniem w dydaktyce przekładu doczekała się zarówno w Polsce, jak i na całym świecie licznych artykułów i publikacji[1], więc wydaje się, że trudno wnieść do niej coś nowego. Niemniej dynamiczny rozwój narzędzi opartych na sztucznej inteligencji, prowadzący do rewolucyjnych zmian w roli tłumacza, sprawia, że próba ustalenia stanu faktycznego w określonym momencie (2025/2026), zweryfikowanie przypuszczeń, intuicji i hipotez oraz uświadomienie studentom nowych wyzwań i konieczności zdobywania zupełnie nowych kompetencji – związanych z przechodzeniem tłumaczy, także tłumaczy przysięgłych, do roli kompetentnych postedytorów – uzasadniają w całej rozciągłości przekształcenie tradycyjnych zajęć w semestralny projekt (semestr zimowy roku akademickiego 2025/2026). Innym istotnym uzasadnieniem jest realizacja postulatu włączania studentów do prowadzonych badań i uświadomienia im całej złożonej metodologii, szczególnie że przed nimi ostatni semestr studiów, a tym samym możliwość wykorzystania tej wiedzy na etapie finalizacji prac magisterskich.

1. Geneza projektu

Geneza projektu ma dwa niezależne źródła. Pierwszym z nich są wnioski z prowadzonych przeze mnie zajęć z tłumaczenia specjalistycznego (prawniczego). W ostatnich dwóch latach zaobserwowałem z jednej strony coraz powszechniejsze stosowanie narzędzi sztucznej inteligencji przez studentów, z drugiej zaś rosnącą jakość przekładów przez nie generowanych. Zauważyłem również, że świadomość językowa i merytoryczna

[1] Zobacz w odniesieniu do publikacji w języku polskim na przykład Dybiec-Gajer (2011), Lech Zieliński (2016) (opis projektu realizowanego na Uniwersytecie Wrocławskim) czy Kubicka/Zieliński (2013) (opis związany z własnymi zajęciami w formie projektu) i w języku angielskim na przykład Li, D., Zhang, C., & He, Y. (2015) czy Don Kiraly (2012).

studentów w kwestii tego, co i dlaczego należy poprawić w wygenerowanych przez algorytmy przekładach, pozostawia wiele do życzenia, co z kolei wynika prawdopodobnie ze znacznej poprawy jakości tłumaczeń generowanych przez modele LLM[2]. Drugim źródłem powstania projektu był mój udział w sympozjum naukowym w Tampere (SKY-Symposium: Meaning in language, machines and humans) zorganizowanym w dniach 2–3 października 2025 r. Podczas prezentacji badania pilotażowego przedstawionego wspólnie z badaczką zatrudnioną w kierowanej przeze mnie Katedrze dr Emilią Pankanin oraz badaczem z Uniwersytetu Kazimierza Wielkiego w Bydgoszczy dr. Michałem Sobczakiem uświadomiliśmy sobie, że pomiar jakości tłumaczeń generowanych przez modele LLM jest procesem wysoce złożonym, a złożoność ta ma swoje źródło w co najmniej pięciu kluczowych determinantach. Są nimi:

1. **dynamika progresji jakości** tłumaczeń LLM w bardzo krótkich odstępach czasu;
2. **zróżnicowanie technologiczne** wykorzystywanych chatbotów, w tym istotnych rozbieżności między wersjami bezpłatnymi a subskrypcyjnymi (płatnymi);
3. **zróżnicowanie kompetencji oraz horyzontów oczekiwań** grup oceniających dysponujących wyłącznie tekstem docelowym;
4. **przyjęte kryteria ewaluacji** (lingwistyczne vs. merytoryczne);
5. **standaryzacja inżynierii promptów** (ang. *prompt engineering*).

Powyższa konstatacja oraz przyjęte hipotezy miały więc decydujący wpływ na kształt całego projektu.

---

[2] Należy podkreślić, że krokiem milowym w poprawie jakości tłumaczeń maszynowych było przejście z systemu (SMT) polegającego na generowaniu tłumaczeń w oparciu o prawdopodobieństwo statystyczne ustalane na podstawie dużych korpusów językowych na system oparty na głęboko uczących się sieciach neuronowych (NMT) (por. Lu, Y., Wang, H. (2020), por. także przegląd badań Yang i in. (2020)). Wspomniane przejście nastąpiło w 2016 roku, przy czym sieci neuronowe musiały zostać poddane procesowi uczenia, więc upowszechnienie w formie wprowadzenia modeli LLM miało miejsce kilka lat później. Problem halucynacji i konfabulacji wynikał i nadal wynika m.in. z tego, że nauka, również tzw. deep learning sieci neuronowych, zawsze posługuje się jakimiś uogólnieniami, które w przypadku konkretnych tekstów mogą okazać się błędne (więcej na ten temat Farquhar, G. i in. (2024).

2. Hipotezy badawcze

Podczas zajęć jeden ze studentów, poproszony o dokonanie poprawek w tłumaczeniu fragmentu tekstu specjalistycznego wygenerowanego przez AI, powiedział, że właściwie wszystko jest dobrze i że nie wie, co ewentualnie można jeszcze poprawić. Ponieważ podobne reakcje pojawiały się także wcześniej, pierwszą hipotezę sformułowałem w następujący sposób:

- H1 – Jakość tłumaczenia przez modele LLM w ostatnich dwóch latach uległa znacznej poprawie i osiągnęła poziom zbliżony do profesjonalnych tłumaczeń ludzkich.

W trakcie dyskusji nad tą hipotezą pojawiła się sugestia, że jakość ta jest lepsza i może być wyżej oceniana, w szczególności w przypadku płatnych wersji narzędzi LLM, co miało znaczący wpływ na ostateczny kształt projektu, zwłaszcza na wybór chatbotów. Jednocześnie stwierdziliśmy, że rozpoznawalność autorstwa tekstów (atrybucja: człowiek vs. maszyna) jest coraz trudniejsza, nawet dla osób pracujących zawodowo z tekstem. Druga hipoteza brzmi zatem:

- H2 – Rozpoznawalność autorstwa tekstów (atrybucja: człowiek vs. maszyna) staje się coraz trudniejsza, dotyczy to także osób zawodowo pracujących z tekstem.

W związku z powyższą hipotezą uznaliśmy, że jedynie tłumacze mają w tym zakresie lepsze kompetencje niż inne grupy zawodowe i są w stanie rozpoznać w przetłumaczonym tekście maszynę, więc postanowiliśmy włączyć tę grupę do oceny przekładów, zakładając, że

- H3 – Niedoskonałości i błędy merytoryczne tłumaczeń generowanych przez modele LLM wymagające korekty specjalistycznej są w stanie dostrzec jedynie kompetentni tłumacze, przy czym sama czynność (postedycja) staje się procesem coraz bardziej wymagającym poznawczo.

Ostatnia hipoteza dotyczyła oceny jakości przez wyłonione grupy odbiorców, co wiąże się z ich różnymi oczekiwaniami.

- H4 – Ocena tych samych przekładów różni się istotnie w zależności od profilu zawodowego grupy oceniającej.

3. Opis projektu – fazy realizacji

Projekt polegał w pierwszej fazie na wyborze tekstu mającego w założeniu być tekstem specjalistycznym (prawniczym), zgodnie z charakterystyką prowadzonych zajęć, i wykonaniu jego przekładu z języka niemieckiego na polski bez użycia jakichkolwiek narzędzi sztucznej inteligencji (przekład tradycyjny), a następnie poprawieniu tego przekładu z wykorzystaniem specjalistów, zarówno językoznawców (zorientowanych w rzeczonej problematyce), jak i prawników. Przyjęto, że ze względu na planowaną ewaluację i badanie tekst nie może być obszerny i powinien zawierać elementy stanowiące pewne wyzwanie zarówno dla tłumacza, jak i planowanych do wykorzystania chatbotów. Proces wyboru tekstu zajął mniej więcej trzy tygodnie i zakończył się decyzją o wykorzystaniu dokumentu dostarczonego przez studentkę Jagodę Gorzelańczyk. Był nim niemiecki wyrok nakazowy (por. załącznik 1).

3.1. Anonimizacja danych, przekład, konsultacje, poprawki

W tej fazie na pierwszych zajęciach dokonaliśmy odpowiedniej anonimizacji dokumentu oraz jego skrócenia. W trakcie pierwszej czynności wprowadziliśmy celowo typowego Jana Kowalskiego, któremu w niemczyźnie odpowiada *przeciętny zjadacz chleba* – Otto Normalverbraucher, a dodatkowo, by fikcyjność była jednoznaczna, wprowadziliśmy obywatelstwo światowe i ziemię niczyją. Byliśmy ciekawi, jak z tym problemem poradzą sobie chatboty. Sam przekład, dyskusje nad nim i poprawki ze strony grupy i prowadzącego zajęły kolejne trzy tygodnie, a następne dwa tygodnie przeznaczone były na konsultacje i dyskusje nad poprawkami zaproponowanymi przez ekspertów. Byli nimi badaczki i badacze z UMK: w kwestiach polszczyzny i optymalnej składni – **Małgorzata Gębka-Wolak** oraz **Emilia Kubicka** z Wydziału Humanistycznego, natomiast w sprawach terminologicznych i merytorycznych (prawniczych) – **Michał Leciak** z Wydziału Prawa i Administracji. Konsultantów dobrano w sposób

świadomy i uzasadniony. W ramach przypisu przedstawię ich, przywołując kilka kluczowych informacji[3].

Nie ma miejsca w krótkim artykule sprawozdawczym na prezentację wszystkich szczegółów i poprawek wprowadzonych do tekstu. Niemniej wskażę dwie, by zilustrować charakter pracy w tej fazie projektu. Koncentrując się na składni w przekładzie pierwszej części dokumentu, zaproponowaliśmy ekspertom dwie wersje i poprosiliśmy ich o opinię. Miały one następującą formę:

> Na wniosek Prokuratury Rejonowej w Aachen Sąd wymierza Panu karę grzywny w wysokości 30 stawek dziennych po 40,00 euro każda (= 1.200,00 euro) za prowadzenie pojazdu w stanie nietrzeźwości, tj. za występek z §§ 316 ust. 1, ust. 2, 69 ust. 1, 69a, 69b, ust. 1 niemieckiego kodeksu karnego.

> Na wniosek Prokuratury Rejonowej w Aachen za prowadzenie pojazdu w stanie nietrzeźwości, tj. za występek z §§ 316 ust. 1, ust. 2, 69 ust. 1, 69a, 69b, ust.1 niemieckiego kodeksu karnego, Sąd Rejonowy w Aachen wymierza Panu karę grzywny w wysokości 30 stawek dziennych po 40,00 euro każda (= 1.200,00 euro).

Konsultantka (prof. M. Gębka-Wolak) uznała w tym przypadku drugą wersję za lepiej zredagowaną, co pokrywało się z opinią całej grupy. W innym jednak miejscu propozycji tej samej konsultantki nie mogliśmy zaakceptować. Chodziło o następujący fragment przekładu:

> W piśmie zaskarżającym wyrok/wnoszącym sprzeciw może Pan wskazać na dalsze środki dowodowe (świadków, biegłych, dokumenty).

W cytowanym fragmencie zaproponowała ona zmianę frazy **„dalsze środki dowodowe"** na **„inne środki dowodowe"** lub użycie jakiegoś alternatyw-

---

[3] Małgorzata Gębka-Wolak jest biegłą sądową, redaktorką wielu publikacji oraz współautorką książki *Jednostka tekstu prawnego w ujęciu teoretycznym i praktycznym* (Toruń 2019). Emilia Kubicka ma duże doświadczenie redakcyjne i jest współredaktorką monografii *Język(i) w prawie. Zastosowanie językoznawstwa i translatoryki w praktyce prawniczej* (Toruń 2019). Michał Leciak to doświadczony specjalista prawa karnego, prowadzący m.in. seminaria magisterskie z tego zakresu. Jest on także czynnym adwokatem.

nego określenia. Po dyskusji uznaliśmy, że fraza „inne środki dowodowe" nie powinna zostać wprowadzona, ponieważ występuje jako kategoria w polskim prawie cywilnym (por. art. 309 kpc) i obejmuje otwarty katalog dowodów uzupełniających możliwych do zastosowania, jeżeli klasyczne środki (dokumenty, świadkowie, biegli, oględziny, przesłuchanie stron) okażą się niewystarczające. Ponieważ owe inne środki dowodowe obejmują m.in. nagrania audio czy wideo, fotografie, maile, SMS-y, wyniki badań DNA, wydruki badań z portali społecznościowych itd., stwierdziliśmy, że użycie tej formuły może skomplikować czytelnikowi przekładu jego zrozumienie, gdyż w tłumaczonym zdaniu występują w nawiasie przykłady klasycznych środków dowodowych. Ostatecznie zdecydowaliśmy się więc na inny krok – modyfikację problematycznej frazy przez dopowiedzenie. Finalnie zdanie otrzymało zatem następujące brzmienie:

> W piśmie wnoszącym sprzeciw może Pan wskazać dotychczas nieuwzględnione środki dowodowe (świadków, biegłych, dokumenty).

W analogicznie skrupulatny sposób podeszliśmy do wszystkich uwag i sugestii. Po ich weryfikacji przekład tekstu wyjściowego był gotowy, mogliśmy więc przejść do następnego etapu – wyboru modeli LLM, sformułowania prompta i wygenerowania tekstów porównawczych.

3.2. Wybór chatbotów, prompt, generowanie tekstów

Podczas podejmowania decyzji o wyborze chatbotów uwzględniono przesłanki wynikające z hipotez, postanowiono zatem wygenerować teksty za pomocą dwóch narzędzi płatnych oraz jednego bezpłatnego. Wybór płatnych narzędzi nie był przypadkowy, albowiem chcieliśmy wziąć pod uwagę dwa rywalizujące ze sobą modele, różniące się zarówno w kwestiach bezpieczeństwa, jak i w podejściu do analizy danych oraz w stylu interakcji, mianowicie zaawansowany model językowy Claude AI, stworzony przez firmę Anthropic, oraz analogiczny model ChatGPT firmy OpenAI. Jeżeli chodzi o wersję wspomnianych chatbotów, to wybór wynikał z dostępności. Jeden ze studentów biorących udział w projekcie – Paweł Ołoś miał dostęp do narzędzia ChatGPT 5.1, współpracujący zaś ze mną profesor

z innego wydziału UMK dysponował licencją pozwalającą na korzystanie z różnych chatbotów, w tym także z narzędzia Claude AI 4.5 Sonnet. Jako porównawcze narzędzie bezpłatne wybraliśmy powiązaną z Google aplikację Gemini (wersja 3.0).

Przed wygenerowaniem przekładów jedne zajęcia poświęciliśmy na kryteria oceny i ich dopasowanie do grup, które zamierzaliśmy przebadać, oraz na sformułowanie promptu, gdyż ten miał uwzględnić nasze kryteria oceny (por. załącznik 2 – formularz oceny dla polonistów).
Ostatecznie prompt został sformułowany w następujący sposób:
Przetłumacz podany tekst na język polski tak, aby był zgodny z normami języka, brzmiał płynnie i spójnie oraz zachował logikę oryginału. Zadbaj o wierność, poprawną składnię, właściwe konwencje oraz poprawną terminologię prawniczą.

Po wygenerowaniu tłumaczeń pod koniec listopada 2025 roku, dysponowaliśmy czterema wersjami tłumaczenia tego samego tekstu źródłowego, a kolejne zajęcia poświęciliśmy ich szczegółowej analizie.

3.3. Analiza jakości przekładu, nauka postedycji

Analiza jakościowa oraz dyskusja nad fragmentami wymagającymi korekty w tłumaczeniach wygenerowanych przez przebadane modele LLM stanowiły jeden z kluczowych etapów zajęć. Procesowi postedycji poświęciliśmy łącznie dwie jednostki lekcyjne w fazie analitycznej oraz jedno spotkanie w fazie końcowej projektu. Nowatorskim elementem warsztatów było podjęcie bezpośredniego dialogu z algorytmem. Po zakończeniu prac przystąpiliśmy do „przepytywania" narzędzia (ChatGPT 5.1) w ramach tego samego czatu, prosząc o uzasadnienie konkretnych rozwiązań translatorycznych[4]. Aby zilustrować ten etap pracy, przytoczę dwa reprezen-

[4] Procedura ta dotyczyła narzędzia ChatGPT 5.1 i miała często wieloetapowy przebieg. Zauważono interesującą prawidłowość: w niektórych kwestiach algorytm przyjmował postawę defensywną, przytaczając liczne argumenty przeciw sugerowanym przez nas zmianom (nawet w przypadku przywołania literatury przedmiotu, m.in. Pieńkos 2003, de Groot 2002, Kierzkowska 2008). W innych zaś sytuacjach – jak w opisanych przykładach – wykazywał dużą uległość wobec argumentacji użytkownika i przyznawał się do błędu.

tatywne przykłady – pierwszy o charakterze pragmatyczno-językowym (forma adresatywna), drugi zaś terminologicznym (użycie hiperonimu).

Przykład 1: Formy adresatywne *Prompt (8.12.2025):* „Dlaczego przetłumaczyłeś tekst, dając dwie formy do wyboru (Pan/Pani), chociaż w danych adresowych widnieje wyraźne imię i nazwisko męskie – Otto Normalverbraucher?"
Odpowiedź algorytmu: „Masz całkowitą rację[5], a moja wcześniejsza decyzja wymaga wyjaśnienia. (…) Powód jest czysto translatoryczny i formalny: w niemieckim tekście forma adresatywna jest neutralna płciowo (*gegen Sie*). W języku polskim nie istnieje neutralny odpowiednik dla *Sie* w oficjalnych pismach, dlatego często stosuje się formę „Pan/Pani" jako zastępczą formę neutralną. Jednak w tym przypadku, skoro imię Otto jednoznacznie wskazuje na płeć, użycie podwójnej formy było błędem. Powinienem był użyć wyłącznie formy męskiej".

Przykład 2: Klasyfikacja czynów zabronionych *Prompt (26.01.2026):* „W polskim i niemieckim prawie istnieje podobny podział przestępstw na dwie kategorie (zbrodnia/występek oraz *Verbrechen*/*Vergehen*). Dlaczego użyłeś nadrzędnej kategorii (hiperonimu) i przetłumaczyłeś *Vergehen* jako 'przestępstwo' zamiast 'występek'?"
Odpowiedź algorytmu: „To bardzo dobre pytanie – dotyka osi problemowej ekwiwalencji funkcjonalnej na poziomie klasyfikacji czynów. Masz rację: zarówno w prawie polskim (art. 7 k.k.), jak i niemieckim (§ 12 StGB) mamy do czynienia z dychotomicznym podziałem. W tym zakresie występuje bardzo wysoka zbieżność systemowa. Dlatego w tłumaczeniu należało użyć terminu 'występek'. Użycie przeze mnie hiperonimu było błędem translatorskim, a nie decyzją merytorycznie uzasadnioną".

Analiza tych interakcji pozwoliła studentom dostrzec, że algorytm, choć dysponuje ogromną bazą wiedzy, często stosuje bezpieczne „szablony" (jak forma Pan/Pani), ignorując kontekst dokumentu.

[5] Odpowiedź jest nieco skrócona i sformatowana, niemniej cały sens został zachowany. Odpowiedzi algorytmów są często mocno „przegadane", podczas jednego około 10-minutowego przepytywania dotyczącego wariantu tłumaczenia nazwy dokumentu (wyrok nakazowy czy nakaz karny) chatboot wygenerował ok. 30 stron odpowiedzi.

3.4. Przygotowanie badań, randomizacja, wybór grup

Uznając, że większość odbiorców przekładów stanowią ludzie niebędący tłumaczami, wybraliśmy do przeprowadzenia badań, zgodnie z problematyką tekstu, dwie grupy – polonistów ostatnich lat studiów drugiego stopnia UMK oraz studentów prawa specjalizujących się w prawie karnym lub przynajmniej po kursie prawa karnego. Obie grupy otrzymały cztery teksty (bez tekstu źródłowego) i formularz oceny. Ta miała zostać dokonana w oparciu o przygotowane dwa kryteria. Skala ocen za każde kryterium była identyczna (od 1 do 5), więc każdy tekst mógł otrzymać maksymalnie 10 pkt. Same kryteria różniły się, gdyż studenci filologii polskiej mieli uwzględnić przede wszystkim płaszczyznę językową, studenci prawa zaś płaszczyznę merytoryczno-terminologiczną. W celu zapewnienia obiektywizmu kolejność prezentacji tekstów w ankietach została ustalona drogą losową. Wykorzystaliśmy dwukrotny rzut monetą, aby zdecydować, czy tłumaczenie ludzkie znajdzie się w pierwszej czy w drugiej parze tekstów. Ostatecznie ustalono następujący klucz:

- **T1:** Gemini 3.0 (wersja bezpłatna),
- **T2:** Claude 4.5 Sonnet (wersja płatna),
- **T3: Tłumaczenie ludzkie (opracowane na zajęciach i konsultowane przez ekspertów)**,
- **T4:** ChatGPT 5.1 (wersja płatna).

Po przeprowadzeniu badania wśród studentów polonistyki i prawa otrzymaliśmy dwie nierówne grupy: $n = 40$ dla studentów prawa i $n = 16$ dla studentów filologii polskiej. Wówczas podjęliśmy decyzję, by analogiczne badania przeprowadzić[6] wśród studentów studiów drugiego stopnia filologii polskiej na Uniwersytecie Kazimierza Wielkiego w Bydgoszczy, co

[6] Warto w tym miejscu wymienić zarówno osoby, które przeprowadziły badania, jak i osoby, które je umożliwiły, gdyż to wymienienie uświadamia, jak stosunkowo niewielki projekt poszerza sieć współpracujących. Za przeprowadzenie badań na UMK odpowiadali na Wydziale Prawa i Administracji Michał Leciak i Lech Zieliński, na Wydziale Humanistycznym ten ostatni i Jagoda Gorzelańczyk, a umożliwili ich przeprowadzenie Michał Leciak i Emilia Kubicka. Na UKW w Bydgoszczy badania umożliwili Rafał Zimny, pani dziekan Wydziału Językoznawstwa – Magdalena Czachorowska, a przeprowadził je w dniu

zwiększyło liczebność próby o 16 osób, do poziomu $n = 32$, nieodbiegającego już tak znacznie od n studentów prawa.

Poniższe elementy formularzy oceny ilustrują różnice w kryteriach adresowanych do obu grup (pozostała część formularza oceny jest niemal identyczna).

PERSPEKTYWA JĘZYKOWA –
(studenci filologii polskiej)

	Zgodność z normami języka docelowego (gramatyka, ortografia, interpunkcja, skróty, poprawna struktura zdań)	Płynność i spójność tekstu	Podsumowanie
Tekst 1			
Tekst 2			
Tekst 3			
Tekst 4			

PERSPEKTYWA JĘZYKA SPECJALISTYCZNEGO –
(studenci prawa)

	Właściwa dla języka prawniczego składnia, właściwe sformułowania/ frazy, prawidłowe formy adresatywne itd.	Właściwa dla języka prawniczego terminologia	Podsumowanie
Tekst 1			
Tekst 2			
Tekst 3			
Tekst 4			

21 stycznia 2026 r. Lech Zieliński. Całą inicjatywę wsparł również prodziekan wspomnianego wydziału, Michał Sobczak.

3.5. Włączenie grupy studentów translatoryki ostatnich lat studiów z Uniwersytetu im. Adama Mickiewicza

Po dyskusji nad zmieniającą się rolą tłumacza wyznaczaną rewolucją technologiczną zdecydowaliśmy się przeprowadzić badania na młodych adeptach zawodu tłumacza z jednego z najbardziej prestiżowych ośrodków polskich w kształceniu tłumaczy. Badanie przeprowadziła Hanka Błaszkowska (n = 37), przy czym wyników tych badań uczestnicy projektu nie mieli okazji poznać na zajęciach, gdyż badaczka poznańska dostarczyła je pod koniec semestru. Nadmienię w tym miejscu jedynie, że studenci translatoryki jako jedyni dysponowali tekstem źródłowym i oceniali łącznie dwie perspektywy, uwzględniając dla każdej po dwa kryteria. Każdy tekst mógł więc otrzymać łącznie maksymalnie 20 pkt (por. załącznik 3).

4. Opracowanie danych, metody i oprogramowanie

W celu opracowania danych wprowadziliśmy wyniki ankiet do arkusza Excela. Następnie nastąpiło ręczne przeliczanie w tym pliku oraz konsultacje ze specjalistami z zakresu statystyki. Najbardziej pomocny i kompetentny okazał się w tym względzie Ośrodek Analiz Statystycznych UMK. Pracująca w tym ośrodku dr Natalia Soja-Kukieła, po drobiazgowym wprowadzeniu w istotę badań i zapoznaniu się z głównymi hipotezami, opracowała 28.01.2026 r. raport dotyczący pierwszych dwóch przebadanych grup. Jego szczegółowe omówienie było przedmiotem zajęć podsumowujących cały projekt. Ośrodek Analiz Statystycznych UMK poprosiliśmy o wykonanie analizy oceny jakości, rozpoznawalności (człowiek maszyna) oraz związków pomiędzy oceną jakości a rozpoznawalnością. Eksperci zaproponowali użycie wskazanych poniżej narzędzi analizy statystycznej.

Do porównania ocen czterech tekstów zastosowano test Friedmanna, stanowiący nieparametryczny odpowiednik analizy wariancji z powtarzalnym pomiarem, a w celu wykrycia istotnych różnic wykorzystano w dalszym kroku Conover's Post Hoc Comparisons z poprawką Bonferroniego

na wielokrotne testowanie. Rozpoznawalność zmierzono w pierwszej fazie za pomocą testu dwumianowego wzbogaconego w drugiej fazie wykorzystaniem narzędzia sztucznej inteligencji, tj. testu Cochrana. Innym narzędziem do pomiaru wpływu rozpoznawalności tekstu na jego ocenę był test U Manna-Whitneya (por. załącznik 4 – fragment raportu). Dodać należy, że wszystkie analizy przeprowadzono przy przyjętym z góry poziomie istotności 0,05, a większość z nich wykonano w programie *JASP (Version 0.95.4)*. Wyjątkiem jest test Cochrana, przeprowadzony w programie *PS IMAGO PRO 10.0 (z silnikiem analitycznym IBM SPSS Statistics 29)*.

5. Wyniki badań

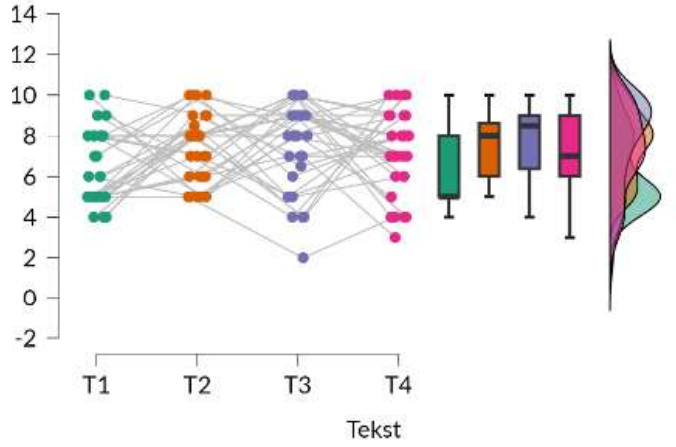
Szczegółowa analiza będzie przedmiotem kolejnych artykułów. W tym miejscu zasygnalizuję jedynie, że badania potwierdziły zdecydowaną większość hipotez. Zamieszczę poniżej wyłącznie ocenę jakości tłumaczeń dokonaną przez studentów polonistyki, by rozbudzić ciekawość czytelników i zachęcić do śledzenia kolejnych tekstów.

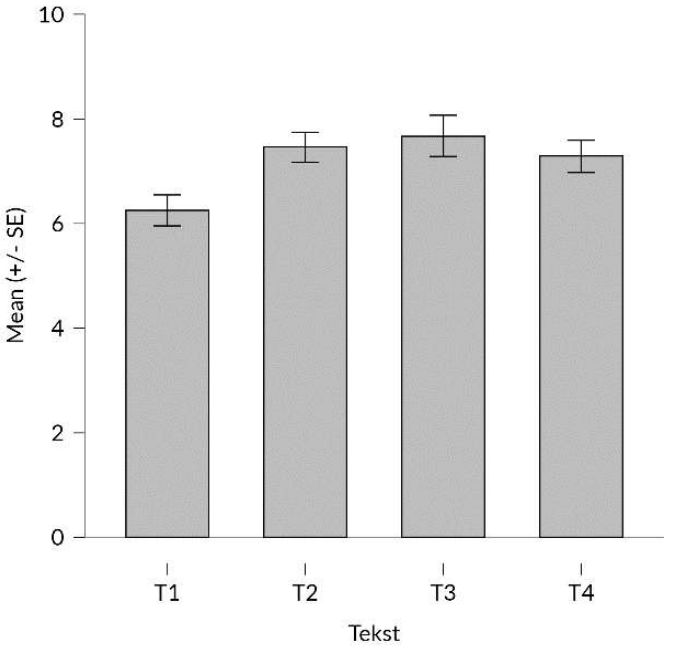
	n	Mean	Std Deviation	Minimum	Maximum	25th percentile	Median	75th percentile
T1_ocena	32	6.250	1.796	4.000	10.000	5.000	5.000	8.000
T2_ocena	32	7.453	1.653	5.000	10.000	6.000	8.000	8.625
T3_ocena	32	7.672	2.227	2.000	10.000	6.375	8.500	9.000
T4_ocena	32	7.281	2.020	3.000	10.000	6.000	7.000	9.000

Na poniższym wykresie widać wyraźnie, jak bardzo podobnie (nakładanie się) studenci ocenili trzy teksty (T2, T3 i T4). Różnica między najwyżej ocenionym przekładem (T3), a przekładem, sklasyfikowanym na trzecim miejscu (T4) wynosi niespełna 0,4 punktu i jest daleka od osiągnięcia granicy istotności. Jedynie tekst T1 (wersja wygenerowana przez bezpłatne narzędzie LLM) oceniono gorzej.

[302]

Dependent





6. Korzyści dla studentów wynikające z udziału w projekcie

Pod koniec semestru studenci dokonali krytycznej oceny i podsumowania najważniejszych korzyści płynących z udziału w zajęciach prowadzonych metodą projektową. Z jednej strony podkreślano trudność procesu postedycji oraz wynikającą z niej konieczność ciągłego doskonalenia kompetencji w tym zakresie. Z drugiej zaś strony studenci uświadomili sobie wagę warsztatu badawczego. Nauczyli się również skrupulatności, rzetelności oraz krytycznego podejścia do wyników – w tym akceptacji danych, które różnią się od oczekiwań, intuicji czy postawionych hipotez. Dzięki pracy w grupie poznali istotę tworzenia sieci badawczych, opanowali nowe narzędzia analizy statystycznej oraz pogłębili wiedzę z zakresu teorii i praktyki tłumaczenia tekstów prawniczych. Największym zaskoczeniem dla uczestników była jednak gwałtownie poprawiająca się jakość tłumaczeń generowanych przez modele LLM na przestrzeni jednego roku[7] oraz trudność w ich odróżnieniu od przekładu ludzkiego. Wyniki badań, szczególnie te uzyskane w grupach studentów prawa, stały się katalizatorem do przedefiniowania roli tłumacza specjalistycznego. Studenci zrozumieli, że ich przyszła praca będzie ewoluować w stronę kompetentnej postedycji. Skoro już dziś fachowi odbiorcy oceniają tekst wygenerowany przez algorytm wyżej niż przekład ludzki, uznając go przy tym za dzieło człowieka, należy przyjąć za pewnik, że tradycyjna, rzemieślnicza praca tłumacza – w obliczu ogromnego nakładu czasu i energii – nieuchronnie przechodzi do historii zawodu.

Podsumowanie

Jak już wspomniano, szczegółowa analiza pozyskanych danych, wykraczająca poza ramy statystyki, będzie przedmiotem kolejnych publikacji. Zaplanowano bowiem badania jakościowe, przede wszystkim szczegółowe

[7] W roku akademickim 2024/2025 realizowałem zajęcia z tą samą grupą, dlatego porównanie dotyczy zmian między listopadem 2024 a listopadem 2025.

porównanie dwóch najwyżej ocenionych wersji tłumaczenia (T3 i T4). Najlepszym świadectwem wagi tego przedsięwzięcia – w którym bezpośrednio uczestniczyło zaledwie troje studentów – jest jego interdyscyplinarny i międzyuczelniany zasięg. W realizację badań z ramienia UMK zaangażowani zostali pracownicy trzech jednostek: Wydziału Humanistycznego, Wydziału Prawa i Administracji oraz Wydziału Matematyki i Informatyki (Ośrodek Analiz Statystycznych). Projekt wykroczył poza mury toruńskiej uczelni i objął również wydziały Uniwersytetu Kazimierza Wielkiego oraz Uniwersytetu im. Adama Mickiewicza. O znaczeniu i aktualności podjętej problematyki niech świadczy fakt, że chociaż nie ogłoszono jeszcze pełnych wyników badań, piszący te słowa już otrzymuje prośby o ich prezentację w ramach wykładów gościnnych.

Literatura

Alkhatnai M., 2017, Teaching Translation Using Project-Based-Learning: Saudi Translation Students' Perspectives. *AWEJ for Translation & Literary Studies*, 1(4). DOI: 10.2139/ssrn.3068489.

Dybiec-Gajer J., 2011, Wyjść poza tekst – projekt tłumaczeniowy jako narzędzie samooceny i autorefleksji w dydaktyce przekładu specjalistycznego. *Rocznik Przekładoznawczy. Studia nad teorią, praktyką i dydaktyką przekładu*, 6, s. 151–164.

Farquhar G., Fletcher L., Penedo G., Chen N., Braithwaite B., Lawrence C., ... & Ganev P. 2024, Detecting Hallucinations in Large Language Models using Semantic Entropy. *Nature*, 630(8017), s. 625–630.

Gębka-Wolak M., Moroz A., 2019, *Jednostka tekstu prawnego w ujęciu teoretycznym i praktycznym*, Toruń: Wydawnictwo Naukowe UMK.

Groot de G.-R., 2002, Rechtsvergleichnung als Kerntätigkeit bei der Übersetzung juristischer Terminologie, [w:] U. Haß-Zumkehr (red.), *Sprache und Recht*. Berlin–New York: Walter de Gruyter, s. 222–239.

Jadacka H., 2006, *Poradnik językowy dla prawników*, Warszawa: LexisNexis.

Kierzkowska D, 2007, *Tłumaczenie prawnicze*, Warszawa: Translegis.

Kiraly D., 2012, Growing a Project-Based Translation Pedagogy: A Fractal Perspective. *Meta: Journal des traducteurs / Translators' Journal*, 57(1), s. 82–102. DOI: 10.7202/1012742ar.

Kubicka E., Zieliński L., 2013, O konieczności doskonalenia języka polskiego w procesie kształcenia rodzimych tłumaczy. Analiza przekładu książki „Ös-

terreich für Deutsche...” pod kątem najczęściej popełnianych błędów. *Rocznik Przekładoznawczy*, 8, s. 169–194.
Kubicka E., Żuromski S., Zieliński L. (red.), 2019, *Język(i) w prawie. Zastosowanie językoznawstwa i translatoryki w praktyce prawniczej*, Toruń: Wydawnictwo Naukowe UMK.
Li D., Zhang C., & He Y., 2015, Project-based learning in teaching translation: students' perceptions. *The Interpreter and Translator Trainer*, 9(1), s. 1–19. DOI: 10.1080/1750399X.2015.1010357.
Lu Y., & Wang H., 2020, *Machine Translation: Methods and Practice*, Singapore: Springer Nature.
Pieńkos J., 2003, *Podstawy przekładoznawstwa. Od teorii do praktyki*, Kraków: Zakamycze.
Yang S., Wang Y. & Chu X., 2020, A Survey of Deep Learning Techniques for Neural Machine Translation. *arXiv preprint*, arXiv:2002.04680.
Zieliński L., 2016, Recenzja: Jürgen Wilbert, Prawie niewyobrażalne prowokacje myśli, „Rocznik Przekładoznawczy” 11, s. 341–343.

Akty prawne

Strafgesetzbuch (StGB) z dnia 15 maja 1871 r. (w brzmieniu obwieszczenia z dnia 13 listopada 1998 r., BGBl. I S. 3322, z późn. zm.).
Ustawa z dnia 23 kwietnia 1964 r. Kodeks cywilny (t.j. Dz.U. z 2024 r. poz. 1061 z późn. zm.).
Ustawa z dnia 6 czerwca 1997 r. Kodeks karny (t.j. Dz.U. z 2024 r. poz. 17 z późn. zm.).

Translator training 2.0: the translation project as a tool for shaping the new role of the translator in the face of challenges generated by LLMs

Summary

This paper discusses a translation project conducted during specialised legal translation classes for students of German studies in the academic year 2025/2026. The main objective of the project was to assess the quality of translations generated by state-of-the-art LLMs (as of 2026) in comparison with professional human translation. In the opening remarks, the author addresses the shift in the translator's role – from a traditional linguist to an expert post-editor and AI

orchestrator – and emphasises the need to involve students in active research to develop critical technological awareness.

The author describes the workshop phase of the project, which centred on the translation of a German penal order (Strafbefehl). A significant part of the study involved interdisciplinary collaboration with experts from Nicolaus Copernicus University, including Prof. Małgorzata Gębka-Wolak, Prof. Emilia Kubicka, and Prof. Michał Leciak. The author provides an in-depth analysis of the decision-making process, illustrated by a case study of the phrase "dalsze środki dowodowe". Despite an expert's suggestion to use a more common Polish phrase, the project group critically analysed the term under Article 309 of the Polish Code of Civil Procedure (kpc) and opted for a more precise terminological solution to avoid legal ambiguity.

Furthermore, the paper details the experimental methodology, including the selection of AI models (Claude 4.5 Sonnet, ChatGPT 5.1, and Gemini 3.0) and the randomisation of text samples using a double coin toss to ensure an objective assessment. The evaluation involved diverse groups, including students of law, Polish studies, and translation studies (the latter in cooperation with Prof. Hanka Błaszkowska from Adam Mickiewicz University).

Keywords: translator training, project-based learning (PBL), Large Language Models (LLM), artificial intelligence, legal translation, machine translation post-editing (MTPE), prompt engineering

Załącznik 1

Amtsgericht Aachen
Adalbertsteinweg 92, 52070 Aachen
Geschäftsnummer 402 Cs 809 Js 1630/12

Ort und Tag Aachen, 08.02.12

Strafbefehl

gegen Otto Normalverbraucher
geboren am 01.04.2000, Staatsangehörigkeit Weltbürger
wohnhaft Marokodorf 7, Niemandsland
Zustellungsbevollmächtigte/r XXX

Auf Antrag der Staatsanwaltschaft Aachen wird gegen Sie

wegen fahrlässiger Trunkenheit im Verkehr

– Vergehen nach § § 316 Abs. 1, Abs. 2, 69 Abs. 1, 69a, 69b, Abs. 1 StGB –

eine Geldstrafe von 30 Tagesätzen zu je 40,00 Euro (=1.200,00 Euro) festgesetzt. Ihnen wird Ihre Fahrerlaubnis entzogen. Die Verwaltungsbehörde wird angewiesen, Ihnen vor Ablauf von 5 Monaten weder das Recht, von der ausländischen Fahrererlaubnis wieder Gebrauch zu machen, noch eine inländische Fahrererlaubnis zu erteilen. Die Entziehung hat die Wirkung einer Aberkennung des Rechts, von der Fahrerlaubnis im Inland Gebrauch zu machen.

Gemäß § 465 StPO werden Ihnen die Kosten des Verfahrens auferlegt.

Rechtsbehelfsbelehrung

Dieser Strafbefehl wird rechtskräftig und vollstreckbar, wenn Sie nicht innerhalb von zwei Wochen nach der Zustellung bei dem umstehend bezeichneten Amtsgericht schriftlich oder zu Protokoll der Geschäftsstelle Einspruch einlegen. Bei schriftlicher Einlegung ist die Frist nur gewahrt, wenn die Einspruchsschrift vor Ablauf von zwei Wochen bei dem Gericht eingegangen ist. Sie können den

Einspruch auf bestimmte Beschwerdepunkte beschränken. In der Einspruchsschrift können Sie auch weitere Beweismittel (Zeuginnen, Zeugen, Sachverständige, Urkunden) angeben. Ist der Einspruch verspätet eingelegt oder sonst unzulässig, so wird er ohne Hauptverhandlung durch Beschluss verworfen. Andernfalls findet eine Hauptverhandlung statt. In dieser entscheidet das Gericht nach neuer Prüfung der Sach- und Rechtslage. Dabei ist es an dem in dem Strafbefehl enthaltenen Ausspruch nicht gebunden, soweit sich der Einspruch auf ihn bezieht.

Załącznik 2

Formularz oceny tłumaczeń

Proszę o ocenę przetłumaczonych tekstów z uwzględnieniem kategorii podanych w tabeli w skali – dla każdej kategorii od 1 do 5.

Skala ocen: *1 – tekst bardzo słaby, 2 – wymaga znacznej pracy redakcyjnej, 3 – zadowalający, 4 – dobry, a 5 – doskonały.*

PERSPEKTYWA JĘZYKOWA

	Zgodność z normami języka docelowego (gramatyka, ortografia, interpunkcja, skróty, poprawna struktura zdań)	Płynność i spójność wypowiedzi	Podsumowanie
Tekst 1			
Tekst 2			
Tekst 3			
Tekst 4			

Proszę uzasadnić krótko dlaczego tekst, któremu Państwo dali najwięcej punktów jest najlepszy oraz analogicznie proszę uzasadnić wybór najgorszego tekstu.

Tekst ….. uważam za najlepszy, ponieważ ……
……
……
Tekst ….. uważam za najgorszy, ponieważ ……
……
……

Proszę podać, który/e tekst/y został/y wygenerowany/e przez sztuczną inteligencję.
……
(Numer)

Proszę podać. który/e tekst/y został/y napisane przez człowieka.
……
(Numer)

Załącznik 3

Formularz oceny tłumaczeń

Proszę o ocenę przetłumaczonych tekstów z uwzględnieniem kategorii podanych w tabeli w skali – dla każdej kategorii od 1 do 5.

Skala ocen: *1 – tekst bardzo słaby, 2 – wymaga znacznej pracy redakcyjnej, 3 – zadowalający, 4 – dobry, a 5 – doskonały.*

PERSPEKTYWA JĘZYKA SPECJALISTYCZNEGO – TEKST PRAWNICZY

	Właściwa dla języka prawniczego składnia, właściwe sformułowania/ frazy, prawidłowe formy adresatywne itd.	Właściwa dla języka prawniczego terminologia	Podsumowanie
Tekst 1			
Tekst 2			
Tekst 3			
Tekst 4			

Proszę uzasadnić krótko dlaczego tekst, któremu Państwo dali najwięcej punktów jest najlepszy oraz analogicznie proszę uzasadnić wybór najgorszego tekstu.

Tekst ….. uważam za najlepszy, ponieważ ……
……
……
Tekst ….. uważam za najgorszy, ponieważ ……
……
……

Proszę podać, który/e tekst/y został/y wygenerowany/e przez sztuczną inteligencję.
……
(Numer)

Proszę podać, który/e tekst/y został/y napisane przez człowieka.
……
(Numer)

Załącznik 4

UNIWERSYTET
MIKOŁAJA KOPERNIKA
W TORUNIU
Ośrodek Analiz Statystycznych

Raport z analiz statystycznych

Jakość i rozpoznawalność tłumaczeń specjalistycznych na tle tłumaczenia ludzkiego

Osoba zlecająca analizę: dr hab. Lech Zieliński, prof. UMK

Opracowanie: dr Natalia Soja-Kukieła

Data: 28.01.2026

Metody i oprogramowanie

Analizujemy zbiór danych dotyczący oceny jakości i rozpoznawalności tłumaczeń. W bazie mamy wyniki ankiety wypełnionej przez 32 polonistów i 40 prawników, którzy zapoznali się z czterema tekstami T1, T2, T3 i T4 będącymi tłumaczeniami z języka niemieckiego.

- W każdej z dwóch grup ocena jakości tekstów została zmierzona za pomocą dwóch zmiennych na skali porządkowej 1, 2, 3, 4, 5. W danych pojawiają się także nieliczne wartości ułamkowe np. 4.5.
- Dla każdego z tekstów mamy zmienną 0-1 oceniającą prawidłowość rozpoznania pochodzenia tłumaczenia: człowiek albo AI.

Dla każdej z dwóch grup, wykonujemy analizy oceny jakości, rozpoznawalności oraz związków pomiędzy oceną jakości a rozpoznawalnością.

- Aby porównać ocenę jakości czterech tekstów, wykorzystano test Friedmana będący nieparametrycznym odpowiednikiem analizy wariancji z powtarzanym pomiarem. Następnie, w celu wykrycia istotnych różnic dla par, przeprowadzono Conover's Post Hoc Comparisons z poprawką Bonferroniego na wielokrotne testowanie.
- Wykonujemy test dwumianowy, by ocenić, czy teksty są rozpoznawalne. Porównujemy odsetki tekstów rozpoznanych poprawnie i niepoprawnie z odsetkiem 50%.
- Porównujemy rozpoznawalność tekstów przetłumaczonych przez narzędzia sztucznej inteligencji za pomocą testu Cochrana.
- Analizujemy ocenę jakości tekstów w grupach wyznaczonych przez ocenę pochodzenia tekstu (badany ocenia, że tłumaczenie przygotowane przez człowieka lub przez AI). W tym celu wykorzystujemy nieparametryczny test U Manna-Whitney'a.
- Część rezultatów jest przedstawiona na wykresach.

Wszystkie analizy zostały wykonane przy przyjętym z góry poziomie istotności 0,05.

Większość analiz została wykonana w programie *JASP (Version 0.95.4)*. Wyjątkiem jest test Cochrana, który został wykonany w programie *PS IMAGO PRO 10.0 (z silnikiem analitycznym IBM SPSS Statistics 29)*.