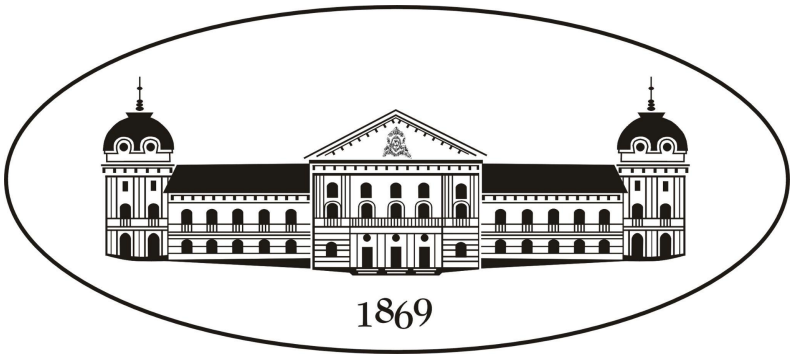
Секция по компютърна лингвистика

Институт за български език „Професор Любомир Андрейчин“

Българска академия на науките



Виктория Красимирова Петрова-Любенова

ПОЛУАВТОМАТИЧНО РАЗРАБОТВАНЕ НА МНОГОЕЗИКОВИ ТЕРМИНОЛОГИЧНИ РЕСУРСИ

Автореферат

на дисертационен труд за присъждане на образователна и научна степен „доктор“

Област на висше образование: 2. Хуманитарни науки;

Професионално направление: 2.1. Филология;

Специалност: Общо и сравнително езикознание

Научен ръководител:

Проф. д-р Светла Коева

София

2022

Съдържание

Обща характеристика на дисертацията

В периода от 60-те до 80-те години на миналия век надеждите, че машинният превод ще се развие достатъчно, че да замести човешкия превод, се оказват неоправдани[1]. Затова усилията се пренасочват към разработването на програми, които да помогнат на преводачите. Нарастващото използване на компютърнобазирани инструменти и ресурси за превод революционизира и променя начина на работа на преводачите и теорията на превода. Новите технологии са толкова неразривно свързани с преводаческата професията, че към днешна дата преводът изглежда немислим, без да се използва някакъв компютърен инструмент (било то и най-обикновен текстов редактор).

Най-голяма е необходимостта от специализирани и технически преводи, тъй като тяхното количество расте с голяма бързина. Такива преводи не могат да бъдат осъществени без познаването на правилната терминология за конкретната област. Ситуацията се усложнява допълнително от липсата на езикови ресурси за малки езици като българския. Затова настоящото изследване цели да анализира каква е ситуацията в момента, и да предложи начин за преодоляване (поне доколкото е възможно) на липсата на нужните терминологични ресурси.

Обектът на изследване включва проучване, анализ и подбор на достъпни и ефективни методи за създаване на терминологични ресурси, предназначени за системите за компютърноподпомогнат превод. Това се отнася както до съществуващите методи за създаване на терминологични бази, така и до проучване на възможностите за интегриране на други технологии, които да спомогнат за улесняването и ускоряването (където е възможно) на процеса по създаване на нови терминологични ресурси.

Предмет на изследване са терминологичните бази и тяхната употреба в системите за компютърноподпомогнат превод, както и начините, по които съвременните технологии позволяват и улесняват създаването и употребата им. Изследвани са нагласите на преводачите и средствата, които те използват за създаването на терминологични бази, анализирани са най-популярните инструменти в тази област (както платени, така и безплатни).

Езиците, които са обект на внимание са български и английски, а посоката на превод, която се разглежда в дисертацията, е от английски на български. Създаването

[1] Най-вече след доклада „Automatic Language Processing Advisory Committee“ или ALPAC от 1966 г.

на терминологични ресурси се илюстрира с паралелни корпуси, терминологични речници и терминологична база в областта на компютърната терминология. Постоянното развитие на технологиите налага и съответното обновяване на компютърната терминология.

Основната цел на изследването е да предложи методика за създаване на терминологични ресурси за дадена специализирана област. Поради това са анализирани възможностите на програмите за компютърноподпомогнат превод за създаването на терминологични бази, програмите за извличане на термини, както и наличните към момента терминологични ресурси за български език.

За постигането на **основната цел** се подхожда комплексно посредством следните **по-конкретни цели:**

- Да се опишат в детайли съвременните системи за компютърноподпомогнат превод, като се акцентира върху техните основни компоненти (преводна памет, машинен превод, терминологични бази, включително многоезикови).
- Да се опишат новите познания и умения, които се изискват от преводачите във връзка с използването на технологиите за компютърноподпомогнат превод.
- Да се проведе проучване сред преводачите в България относно използването на системите за комютърноподпомогнат превод и да се обобщят и анализират резултатите от него.
- Да се представят видовете терминологични речници и тяхната структурата в съпоставка със структурата на терминологичните бази.
- Да се опише структурата и особеностите на терминологичните бази.
- Да се представи създаването и управлението на терминологичните бази в системите за компютърноподпомогнат превод. В допълнение ще се разгледат притурки, които могат да се използват заедно със системите за компютърноподпомогнат превод, както и самостоятелно функциониращи системи за работа с терминология.
- Да се покаже използване на терминологичните бази в зависимост от различните типове текст за превод и приложението на контролирани езици.
- Да се представят накратко особеностите на компютърната терминология, като стремежът е към обобщение от гледна точка на включването на компютърните термини в терминологични ресурси.
- Да се опишат начините за създаване на терминологични ресурси, като се започне от традиционните терминологични речници и се достигне до

автоматичното извличане на термини и представянето на съществуващите инструменти за тази цел.

- Да се създаде методика за полуавтоматично създаване на терминологична база в областта на компютърната терминология и да се опишат конкретните етапи от работата по създаването на двуезикова терминологична база (за английски и български език).
- Да се представи измерването на точността на превода със системите за компютърноподпомогнат превод, съответно без или с използването на терминологични бази.
- Да се представи измерването на точността на превода с функционалности, вградени във или външни за инструментите за компютърноподпомогнат превод.
- Да се предложи класификация на грешките при превод със системи за компютърноподпомогнат превод с оглед на тяхната значимост на съдържанието на текста.
- Да се направи сравнение на ефективността на програмите за оценка на превода по предварително зададена методика.

При разработването на дисертацията са използвани различни **съвременни методи (статистически, за компютърна обработка, корпусен**), както и някои **традиционни** (**описателен, аналитичен, експериментален**). Основни подходи за постигане на изследователските цели са анализът и описанието. Към описателния метод на работа могат да се причислят прегледът на съществуващата научна литература в тази област, прегледът на известни терминология бази и прегледът на техническите инструменти, използвани в ежедневната работа на преводачите и под. Анализ се прилага при подбора на подходящи езикови ресурси и средства за тяхната обработка.

Формулирани са **хипотези**, например за нагласите на преводачите в България към новите средства за работа при превод или за количеството термини, които могат да бъдат извлечени в зависимост от изходния паралелен корпус. Хипотезите се проверяват експериментално по различен начин: посредством анкета, проведена онлайн; посредством работа с програми за автоматична обработка на текстове: за автоматично извличане на термини, за автоматично подравняване на паралелни документи по изречения, за автоматично създаване на двуезикови терминологични бази.

Работата по дисертацията изискваше професионална работа с поредица от системи за компютърноподпомогнат превод и техните компоненти (преводна памет и

терминологична база), автономни системи за автоматично извличане на информация, системи за автоматично извличане на термини и подравняване на паралелни текстове по изречения и не на последно място – автономни системи за проверка на качеството на преода.

Научната разработка се състои от **четири самостоятелни глави**.

В Първа глава – *Системи за компютърноподпомогнат превод и техните компоненти,* са разгледани нови функции на системите за компютърноподпомогнат превод, тяхното използване в зависимост от лингвистичните характеристики на различните типове текстове и изискванията на системите за компютърноподпомогнат превод към уменията на преводачите. Направено е проучване сред преводачите в България, за да се обобщят и анализират степента на използване на новите технологии сред преводачите и техните нагласи по отношение на машинния превод и системите за компютърноподпомогнат превод.

Втора глава – *Терминологични бази в системите за компютърноподпомогнат превод* представя термините, терминологичните речници и терминологичните бази в контекста на превода. Направено е сравнение на терминологичните бази с терминологичните речници и на преводната памет и терминологичните бази. Анализирани са методите за създаване и управление на терминологичните бази в системите за компютърноподпомогнат превод, притурките към основната програма, системите за управление на терминология и уменията за работа с терминология и терминологични бази.

Трета глава – *Полуавтоматично създаване на многоезикови терминологични бази* представя създаването на терминологични ресурси в областта на компютърната терминология и особеностите ѝ. Анализирани са средствата за автоматично извличане на термини и е направено сравнение между тях. Разработена е методика за полуавтоматично създаване на терминологична база в областта на компютърната терминология чрез двуезиков паралелен корпус. След извличането на термините е направена ръчна проверка и техническо оформление на терминологичната база. Описани са характеристиките на създадената терминологична база в областта на компютърната терминология.

Четвърта глава – *Измерване на качеството в системите за компютърноподпомогнат превод и техните компоненти* представя измерването на точността на превода със системите за компютърноподпомогнат превод без терминологични бази, с терминологични бази и с програми, външни на инструментите

за компютърноподпомогнат превод. Предложена е класификация на грешките и е направено сравнение на резултатите от автоматичните средства за измерване на качеството. Разгледани са международни стандарти, метрики и модели, свързани с качеството на превода.

В дисертацията има три приложения, показващи в пълнота резултатите от направената анкета сред преводачите в България, автоматично извлечените термини на английски в областта на компютърната терминология и английско българската терминологична база в областта на компютърната терминология.

ПЪРВА ГЛАВА

Системи за компютърноподпомогнат превод и техните компоненти

След неуспеха на първите прототипи на машинен превод, се предлага нов подход: система, която не превежда автоматично, а улеснява работата на преводача. Става въпрос за системите за компютърноподпомогнат превод (Computer Assisted / Aided Translation Tools) – софтуер, чрез който преводачът може да превежда максимално бързо и лесно възложените му преводи. В тези системи преводачите имат възможност да редактират, създават, съхраняват и управляват съдържанието, докато в същото време имат на разположение терминологични бази, предишни преводи, машинен превод, проверка на правописа и други полезни функции (например визуализация на оригиналния документ, функция за използване и комбиниране на макроси, опция за диктуване и др.). Основна характеристика на този вид програми е възможността им да разделят документите на сегменти (отделни изречения или абзаци, в зависимост от настройките), които се съхраняват в база от данни и които могат да бъдат използвани отново.

Създателите на системите за компътърноподпомогнат превод дават следните дефиниции за този вид програми: „Инструментите за компютърноподпомогнат превод са софтуерни приложения, които помагат за превода на текстове от един език на друг. По-конкретно, инструментите за компютърноподпомогнат превод се използват за възлагане, редактиране, управление и съхранение на преводи“ според Мемсорс[2]; „Инструментите за компютърноподпомогнат превод разделят големи многоезикови документи на сегменти (изречения и абзаци), които се съхраняват в база от данни, наречена преводна памет: вече преведените документи могат да бъдат използвани повторно по всяко време“ според мемоКю[3]; „Това е софтуер, използван от преводачи и лингвисти. Има различно предназначение, но преди всичко подпомага процеса на превода. Позволява редакция, създаване, управление на превода“ според Традос[4]. През 2015 г. са изброени шестдесет и седем системи за компютърноподпомогнат превод на

[2] https://www.memsource.com/what-are-cat-tools/
[3] https://www.memoq.com/tools/what-is-a-cat-tool
[4] https://www.trados.com/solutions/cat-tools/translation-101-what-is-a-cat-tool.html

пазара (Гарсия 2015: 22), а през 2020 г. бройката се покачва на 164[5], като вече се говори за „системи за управление на превод“ (Translation Management Systems). Софтуерът за управление на превод е предназначен да организира и управлява целия процес на превод и обикновено обхваща и функции, които не са свързани конкретно с него.

Системите за компютърноподпомогнат превод имат два основни компонента: преводна памет и терминологична база. Машинният превод също е част от някои от системите за компютърноподпомогнат превод.

Преводната памет (Translation Memory или TM) е база от данни, която съхранява изрази от вече преведени от човек документи. При въвеждането в програмата за компютърноподпомогнат превод текстът се сегментира (по подразбиране на всеки пунктуационен знак за край на изречение) и подравнява в съответстващи си двойки изходни и целеви езикови единици, или сегменти. Всички сегменти от изходния и целевия език се съхраняват в паметта. Когато даден израз бъде засечен отново в същия или нов документ, преводът му се предлага от паметта. Причината преводната памет да е изключително ценна, е, че спестява време и средства на притежателите си и спомага за увеличаване на ефективността и производителността на преводачите, защото прекарват по-малко време в търсене на правилния превод и трябва само да сравнят повторенията, идентифицирани от паметта (Мичъл-Шуитевьордер 2020: 5). Въпреки че качеството на машинния превод значително се подобрява през последните няколко години (най-вече с напредъка в областта на невронния машинен превод), все още се предпочита използването на преводна памет. По този начин не само е възможно да се превеждат „перфектно“ вече преведени изречения, но също така се предлага „почти перфектно“ качество на превода, когато се извличат от паметта подобни изречения (Булте и Тецкан 2019).

Основната единица в паметта, сегментът, обикновено е разграничена с пунктуационен знак и следователно обикновено е изречение, но може да бъде и заглавие, мото, формула и др. Сегмент на изходния език (source segment), свързан със своя превод (target segment), плюс съответните метаданни (например час, дата и име на преводача или редактора, извършил корекцията, име на клиента, тематична област и т.н.) образуват т. нар. „единица за превод“ (Translation Unit или TU). Паметта съдържа също и алгоритъм за разпознаване на съвпадения (matches): когато идентичен или подобен сегмент се срещне в нов текст. Ако в паметта се открие изходен сегмент в

[5] Според годишния доклад на Нимзи за 2020 г. https://www.nimdzi.com/nimdzi-language-technology-atlas-2020/?hilite=trados

базата от данни, който точно съвпада със сегмента от новия текст, се предлага съответният превод като точно съвпадение или 100% съвпадение. При такава ситуация това, което преводачът трябва да направи, е да провери дали сегментът може да бъде използван повторно такъв, какъвто е, или са необходими корекции заради разлики в контекста. Ако в паметта се открие изходен сегмент в базата от данни, който е подобен на превеждания сегмент, той се предлага като частично съвпадение (Fuzzy Match) заедно със степента му на подобие, посочена като процент. Подобието се изчислява по Разстоянието на Левенщайн[6], т.е. минималния брой вмъквания, изтривания или заместявания на символи; след това преводачът преценява дали предложеният превод може да бъде адаптиран, или по-малко усилие ще представлява да се направи изцяло нов превод.

Според настройките по подразбиране обикновено се предлагат и визуализират само сегменти със съвпадение от над 70% (за Традос е 75%), тъй като се приема, че съвпаденията с по-нисък процент са в по-голяма степен пречка, отколкото полезни за преводача, ако в паметта не може да бъде намерен сегмент, надвишаващ предварително зададения праг на съвпадение[7]. Случаят се нарича праг на частично съвпадение (Fuzzy Match Threshold). Когато от паметта няма предложения, е налице „липса на съвпадение“ (No Match Found) и преводачът трябва да преведе този сегмент по традиционния начин (Гарсия 2015: 71-72). По-новите версии на системите за компютърноподпомогнат превод имат функция за контекстни съвпадения (Context Matches). Контекстното съвпадение е известно още като „101% съвпадение“ и се наблюдава, когато има точно съвпадение със сегмент в преводната памет, както и съвпадение на неговия контекст: т.е. предходният и следходният или следходните сегменти също са точно съвпадение, т.е. 100%.

Преводната памет се нуждае от три различни вида софтуер, за да функционира правилно:

- система за подравняване на сегменти (aligner), посредством която се създадат файловете за преводна памет. Може да бъде както самостоятелна програма, така и вградена функция в системите за компютърноподпомогнат превод. В нея се

[6] Разстоянието на Левенщайн между два низа е броят на изтривания, вмъквания или заместявания на единичен символ, необходими за трансформиране на един низ в друг. То е известно още и като разстояние за редактиране.

[7] Трябва да се отбележи, че са възможни разлики в изчисляването на частичните съвпадения. Влияние в по-малка или голяма степен могат да окажат: словоредът, пунктуацията (понякога влиянието не е важно, а друг път може напълно да промени значението на сегмента), форматирането и таговете (елементи, които съдържат кодирана информация за формата и структурата на дадена дума, фраза или сегмент).

съпоставят изходните и целевите сегменти от вече преведения текст, за да се създаде преводната памет;

- система за управление на формàта, която е вградена опция в програмите за компютърноподпомогнат превод. Тя следи форматирането на изходния текст да съвпада с форматирането на преводния текст;
- система за управление на преводната памет, която дава възможност за преместване на съдържание между различни преводни памети, изтриване, експортирате на части от съдържанието, както и историята на всеки сегмент, намиращ се в паметта (Гаудек 2007: 271 – 272).

Най-общо казано, **машинният превод** е специален софтуер за превод на текст от един естествен език на друг, без да има човешка намеса. Основните тенденции при компютърноподпомогнатия превод в момента поставят все по-голям акцент върху машинния превод.

Двата най-популярни и разпространени модела са статистическият и невронният машинен превод. При статистическия машинен превод (Statistical Machine Translation или SMT) декодерът, който „превежда“, е по същество алгоритъм за търсене. За всяка дума и група думи в изречението за превод алгоритъмът проверява наличните ресурси, съдържащи подравнени думи и групи от думи, и извлича най-добрия еквивалент. Този метод може да се определи и като парадигматичен: всяка изходна дума създава слот, който може да бъде попълнен от дадена дума от наличните преводни думи и фрази (Боукър и Сиро 2019: 37). Невронният машинен превод (Neural Machine Translation или NMT) работи по различен начин. Декодерът не търси конкретен елемент от наличните ресурси. Използва първо невронните мрежи, за да се научи и след това да идентифицира най-добрите последователности за превод на цели изречения. Декодерът се обучава от големи по обем паралелни данни. Невронният машинен превод се опитва да изгради последователност от думи при превода линейно. Всяка последователност от предишни думи определя следващата дума.

В продължение на много години технологиите за машинен превод и тези за компютърноподпомогнат превод са противопоставяни една на друга. В последните години обаче машинният превод (най-вече след навлизането на невронния машинен превод) започва да заема все по-централна роля в работните процеси при компютърноподпомогнатия превод. Например в най-новата версия на Традос използването на машинен превод става неразделна част от процеса на всеки нов

проект: след като документите за превод, паметта и терминологичната база бъдат заредени, програмата създава машинен превод специално за този проект[8].

Първите опити за интеграция между преводната памет и машинния превод се отнасят до възможността те да се комбинират по два начина. Единият е потребителят да получава за всеки сегмент предложения от машинния превод, преводната памет и терминологичната база и те да бъдат визуализирани едновременно. Вторият начин за комбиниране е използването на двете технологии заедно, за да се подобрят резултатите в целевия език и по този начин да се увеличи производителността и да се намалят усилията след редактиране.

В същото време, въпреки че използването на технологиите за машинен превод е неизбежно, практически погледнато резултатът от всеки машинен превод може да изисква някаква форма на човешка намеса. Най-важната причина машинният превод да не може да се използва самостоятелно[9], си остава качеството на изходния текст. Един начин за подобряване на качеството на машинния превод е предварителната обработка на текстовете, използвани за обучение: те се филтрират, за да се премахнат неизвестни термини, неясноти и многозначност от различен характер; текстовете се преобразуват в „опростен“ език с кратки изречения. Друг начин за подобряване на качеството на машинния превод е чрез „обучение“ с помощта на преводна памет. От гледна точка на качеството този метод се приема за по-добър, тъй като обикновено потвърдените сегменти в паметта са одобрени от професионални преводачи.

Нарастващото интегриране на машинния превод в работния процес на проектите за превод води до така наречения „разширен превод“ (augmented translation). Разширеният превод комбинира уменията на преводачите и машинния превод. А. Аренас определя тази дейност като „преразглеждане на предварително преведен текст, генериран чрез машинен превод на изходния текст, и коригиране на възможните грешки с цел да бъдат постигнати предварително заложените критерии за качество“[10] (Аренас 2020: 333). Може да се направи разграничение на два вида редактиране на машинния превод: частично и пълно редактиране. В първия случай текстът трябва да бъде разбираем, като се допускат някои граматически и правописни грешки. Във втория случай текстът, стилът, граматиката, правописът и терминологията трябва да

[8] Свързването става по различен начин за по-новите и по-старите версии на програмата. https://www.trados.com/products/machine-translation/

[9] Не се взема под внимание употребата на машинния превод за лични цели.

[10] Важно е да се подчертае, че критериите за качество се определят заедно с възложителя на проекта преди започването на работата.

бъдат сравними с тези на текст, превеждан от човек. За работата с машинен превод са необходими и специализирани технически познания, които може да не принадлежат към компетенциите на преводачите.

Терминологичните бази са централизирани бази от данни, които съдържат специфични за дадена тематична област или даден проект за превод термини. Обикновено тези термини са предварително одобрени (в много случаи самите клиенти ги съставят). Основното предимство на терминологичната база е, че прави процеса на превод по-бърз (особено ако той също е автоматизиран), защото спестява на преводачите нуждата да търсят необходимия термин. Както при преводната памет, програмата за компютърноподпомогнат превод сканира всеки нов сегмент, за да провери дали има съвпадение с терминологичната база. Този процес на работа гарантира също и високо ниво на последователност, защото един и същ термин се превежда по идентичен начин.

На пръв поглед терминологичните бази са като речници – съдържат в себе си списък от термини. За разлика от тях обаче могат да бъдат конфигурирани от потребителя на базата, който има пълната свобода във всеки един момент да добавя, изтрива, модифицира и класифицира термините в различни категории (например забранен термин или термин, който да бъде използван само в определен контекст). **Многоезиковите терминологични бази** следват същата структура и функционират по същия начин като двуезиковите.

Механизмът, по който работят системите за компютърноподпомогнат превод, ги прави изключително подходящи за еднотипни текстове, защото повтарящите се езикови единици се превеждат значително по-бързо. Подобряването на скоростта на превода посредством технологията за съпоставяне, описана по-горе, спестява нуждата един и същ превод да бъде набиран отново. Стандартизацията на текстовете помага и за точността на превода. Чрез преводната памет вече не се налага преводачите да помнят какво са превеждали, за да превеждат по еквивалентен начин дадени фрагменти в големи по обем текстове.

П. Нюмарк в своята книга *Учебник по превод* (A Textbook in Translation) от 1988 г. заявява: „Винаги е възможен задоволителен превод, но добрият преводач никога не е доволен от него. Обикновено може да го подобри. Няма перфектен, идеален или „правилен“ превод. Преводачът винаги се опитва да разшири знанията си и да подобри изразните си средства; той винаги преследва факти и думи. Преводачът работи на четири нива: преводът е първо наука, която включва познаването и проверката на

фактите и на начина, по който ги описва – на това равнище може да се идентифицират фактически грешки; второ, това е умение, което изисква подходящ език и правилна употреба; трето, това е изкуство, което отличава добрия, творческия, интуитивния, понякога вдъхновен превод от безличното предаване на съдържание на друг език; и накрая, това е въпрос на вкус и разнообразието от качествени преводи е отражение на индивидуалните различия на преводачите“ (Нюмарк 1988: 6).

II. Отношението на преводачите към технологиите за компютърноподпомогнат превод

Профилът на преводача придобива нови измерения с развитието на технологиите. Необходимо е да се разграничат преводачите на художествени текстове от преводачите на специализирани текстове. Изискванията към последните се променят сериозно през последните години и продължават да се променят, като налагат работа с най-новите инструменти за компютърноподпомогнат превод като Традос или други инструменти и платформи. Преводачите изпълняват широк спектър от задачи, които включват, но не се ограничават до: превод, редактиране, преглед, употребата на езика, управление на терминология и много други.

Използването на системите за компютърноподпомогнат превод налага промяна в изискванията към образованието и уменията на преводачите. Профилът на преводача вече не може да бъде ограничен само до езиковите умения. Преводачите се превръщат във висококвалифицирани технически експерти поради съдържанието, което превеждат, и инструментите и софтуера, които трябва да използват в ежедневната си работа (свързана с познаването на различните системи за компютърноподпомогнат превод, управлението на преводна памет, създаването и редактирането на терминологични бази, последващото редактиране на машинен превод, необходимостта от програмиране, работата със субтитри, локализацията, обработката на изображения, обработката на файлове в различни формати и др.).

Един от начините, използвани през годините, за да се проверят нагласите и мненията на преводачите, са допитванията и анкетите. Според проучване на американската компания Сиесей[11] от 2020 г. за работата на 7 363 преводачи малко повече от една трета (37%) от тези, които използват машинен превод, смятат, че общото качество е добро. 81% забелязват сериозни вариации в качеството при текстове,

[11] Сиесей Рисърч (CSA Research) – независима американска компания за маркетингови проучвания. https://csa-research.com/

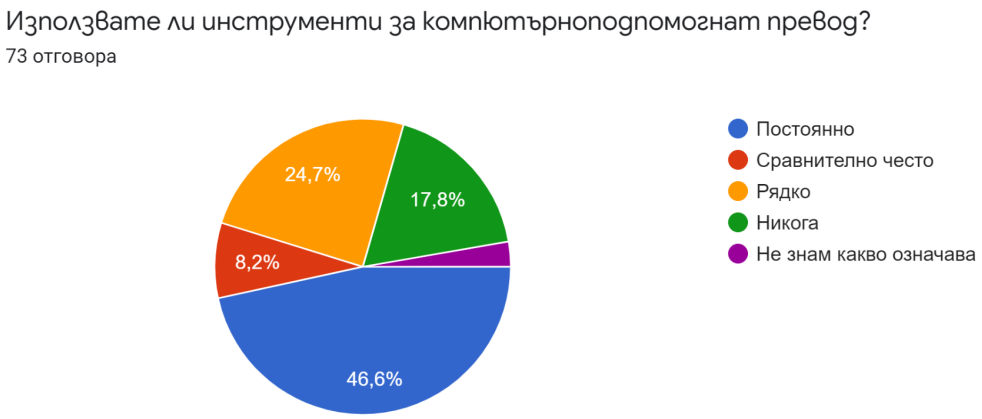
необработени предварително от човек (например елиминиране на многозначността, заместване на фразеологизми и др.). Преводачите, които използват машинен превод, обикновено предпочитат да работят с адаптивни системи, каквато е Лилт[12] (заради цялостното по-добро качество, гарантирано от персонализирането на съдържанието), вместо такива с необработен машинен превод (71%), каквито са свободно достъпните платформи като Гугъл Транслейт. Преводачите оценяват приноса на речниците към качествените резултати (91% от анкетираните предоставят по-добро качество, когато използват речници, а 76 %, когато използват преводна памет). Интересно е, че почти една четвърт от преводачите (23%) твърдят, че предоставят по-добро качество, използвайки машинен превод – като опровержение на общоприетото мнение, че крайният продукт на машинния превод е по-лош в сравнение с човешкия превод. Преводачите все още оценяват използването на машинния превод като забавяне на процеса на превода в сравнение с работата с другите инструменти, което най-вероятно е обвързано с факта, че резултатът на машинния превод не е достатъчно добър и трябва да се редактира.

В годишния доклад за *Проучване на европейската езикова индустрия* (European Language Industry Survey или ELIS[13]) за 2022 г. според констатациите: „значително по-висок процент от независимите професионалисти съобщават, че технологичното обучение, което получават от създателите на инструменти, подпомагащи превода по различни начини, е достатъчно, за да се справят успешно (48%, в сравнение с 40% през 2021 г.), но все още има 21% сред анкетираните, които не са съгласни с това заключение, което е повече в сравнение с резултата през 2018 г. – 18%“. Машинният превод показва най-висок темп на растеж по отношение на използването си, но все още е далеч зад преводната памет. По-малко от 10% от анкетираните планират да инвестират в закупуването на технологии, които подпомагат превода, с изключение на машинния превод (11%).

Направено е **проучване сред преводачите в България.** Анкетата се състои от 26 въпроса относно използването на инструментите за компютърноподпомогнат превод, преводна памет, терминологичните бази и машинния превод. Анкетата е проведена в периода 01.02 – 31.03.2022 г. и в нея участваха общо 73 професионални преводачи.

[12] Lilt https://lilt.com/
[13] https://elis-survey.org/

Най-голям процент от анкетираните са отговорили, че превеждат специализирани текстове (90,4 %), следват тези, които превеждат административни документи (69,9 %), художествена литература (23,3 %) и друг тип документи (31,5 %). С други думи, по-голяма част от групата на респондентите се занимава с документи, които са подходящи за компютърноподпомогнат превод. 46,5 % от преводачите използват системи за компютърноподпомогнат превод постоянно, а 8,2 % – сравнително често. Почти половината от анкетираните използват системите за компютърноподпомогнат превод рядко или никога (Фигура 1).



Фигура 1: Честота на използване на инструментите за компютърноподпомогнат превод

Най-популярните системи за компютърноподпомогнат превод са Традос, Мемсоурс, мемоКю, Уърдфаст, СмартКАТ. Почти половината от анкетираните ги оценяват като много полезни (49,3 %), а над една четвърт – като донякъде полезни (26 %). Около една четвърт от преводачите нямат мнение или смятат, че системите за компютърноподпомогнат превод не са полезни. От тези числа може да се направи изводът, че системите за компютърноподпомогнат превод се използват от българските преводачи.

На въпроса „Как според Вас трябва да се развият и подобрят инструментите за компютърноподпомогнат превод?“ са отговорили общо 55 от 73 участници. Тук са поместени част от получените отговори: „Да имат адекватни инструкции и на български език. Не всички преводачи ползват английски“; „Обогатяване на речника и включване на идиоматични изрази, да се работи за по-добра граматика и лексикална прецизност“; „Необходима е по-добра работа с преводачески памети и терминологични бази; възможност за директно търсене на термини“ и др. Отговорите показват, че

(интуитивно или не) преводачите имат реалистични очаквания за развитието на системите за компютърноподпомогнат превод, а също така – доста добре очертани изисквания към функционалностите на тези системи, което (без съмнение) се дължи на опита им за работа с тях.

На въпроса дали използват преводна памет, 47,9% от анкетираните са отговорили, че използват постоянно, а 12,3% – сравнително често, това дава над 60% общо в полза на използването на преводна памет.

Отговорите на въпроса дали преводачите използват терминологични бази, са съпоставими с отговорите на въпроса дали използват преводна памет: 31,5% използват терминологични бази постоянно, а 32,9% – сравнително често, отново над 60%. От това може да се заключи, че терминологичните бази се използват, което препраща към свободните отговори за употребата на системите за компютърноподпомогнат превод, според които някои преводачи искат да имат връзка за търсене на термини в терминологични хранилища. Липсата на подходящи терминологични ресурси е още един повод за намирането на лесен, универсален и достъпен начин за създаването на терминологични бази.

67,8% от преводачите, използващи терминологични бази, ги създават сами, 39,4% използват бази на клиентите, 26,8% – на агенциите, а 49,3% използват публичнодостъпни терминологични бази.

41% от участниците в анкетата използват публичнодостъпни бази. Тук трябва да се отчете фактът, че в много случаи преводачите се нуждаят от специализирани терминологични бази, които не са отразени в посочените публични източници. Последното е в подкрепа на възможността сами да ги създават.

На въпроса „Според Вас какво е значението на терминологичните бази за качеството на превода?“ 43% са отговорили сравнително голямо и 37% – голямо. От тези отговори може да се заключи, че разработването на унифицирана методика за създаването на терминологични бази може да бъде от полза за ежедневната работа на преводачите.

Интересни са резултатите при отговора на въпроса дали преводачите използват машинен превод: 42,5% го използват рядко, а 34,2% – никога. Само 5,5% от преводачите посочват, че използват машинен превод постоянно. Отговорите може да се дължат на все още незадоволителното качество на машинния превод, но и възможното нежелание да се признае, че се използват технологии, които се очаква (ако не да заместят изцяло) до голяма степен да облекчат работата на преводачите.

Голямо разнообразие от отговори има на въпроса „Как проверявате качеството на превода?“. Само 17,8% отговарят, че използват функционалностите за проверка на качеството на инструментите за компютърноподпомогнат превод; 9,6% си служат с външни програми, 61,6% (най-голямата група) повторно проверяват текста сами.

Отговорите на българските преводачи не се различават в значителна степен от другите анкети, които бяха представени. Преводачите се стремят да използват новите технологии като преводна памет и терминологични бази, тъй като те могат да допринесат за подобряване на бързината и донякъде на консистентността на превода. От друга страна, има все още преводачи, които не само не използват нови технологии, но и изразяват притеснението си от тях. Тук трябва да се посочи и липсата на публичнодостъпни терминологични бази в много области от човешкото познание, както и недостатъчно качественият машинен превод особено от езици, различни от английски. Анкетата сред български преводачи, чиято ежедневна работа е свързана с технологиите за компютърноподпомогнат превод, потвърждава тези изводи.

ВТОРА ГЛАВА

Терминологични бази в системите за компютърноподпомогнат превод

I. Термини, терминологични речници и терминологични бази

Според М. Попова „**термините** са най-информативният пласт в лексиката на научния език, реч и текст (Попова 2019: 149). Ще бъдат разгледани дефинициите за термин с оглед на употребата на термините в терминологичните бази.

М. Бейкър предлага следната дефиниция за термин: „Термините се различават от думите по това, че притежават специална форма на референция (form of reference), а именно, отнасят се до отделни концептуални единици, свойства, дейности или отношения, които се отнасят до знанието от определена предметна област. За да се направи разлика между специална и обща референция, се установява разграничение между термините, които имат специална референция в рамките на определена научна дисциплина, и думите, които функционират с обща референция в различни предметни области“ (Бейкър 2001: 216). Х. Сайер предлага следната дефиниция за термин: „Термините се различават от думите по това, че притежават специална форма на референция (form of reference), а именно, отнасят се до отделни концептуални единици, свойства, дейности или отношения, които се отнасят до знанието от определена предметна област. За да се направи разлика между специална и обща референция, се установява разграничение между термините, които имат специална референция в рамките на определена научна дисциплина, и думите, които функционират с обща референция в различни предметни области“ (Сайер 2001: 261).

А. Кис дава значително по-съкратена дефиниция: „терминът (авторът използва названието terminus technicus) е израз, принадлежащ към техническия език“ (Кис 2005: 106). А. Имре констатира, че различните дефиниции подчертават различни аспекти: някои от тях вземат предвид „формата“ на думата или израза, а други се фокусират върху значението. Терминът трябва да има само едно значение, като синонимните термини се изключват, но най-важното е, че значението на термините е добре установено (без припокриване със значението на други термини), така че да не може да бъде разширено или стеснено (Имре 2013: 127).

Съществува разграничение между широко и тясно разбиране за терминологията: при широкото разбиране за терминология се смята, че термин е всяка дума или фраза, изразяваща понятие, която се използва в определен контекст (Тръмбъл и Стивънсън 2002); тясното разбиране приема, че термин е дума или фраза, използвана с точно определено значение в дадена тематична област или дисциплина.

Терминологичните речници се характеризират с експертни дефиниции, думи и значения, които принадлежат към конкретна техническа или научна предметна област (като медицина или математика) (Де Калуве и Ван Сантен 2003: 82). Изборът на термини за речника определя неговия тематичен обхват, а когнитивната му роля се реализира или чрез представянето на семантичните отношения между термини и понятия, или чрез имплицитна семантична информация, предадена с помощта на дефинициите на включените термини. От тази гледна точка терминологичният речник се превръща в средство за трансфер на професионално знание (Лукашик 2012: 100).

Към терминологичните речници могат да бъдат включени един или повече от следните компоненти (Попова 2016: 71): научната област на термина; речникова единица – заглавка; лексикално-граматична информация – като част на речта; формалноезикова информация – граматични особености; функция; етимологични данни; информация за съчетаемостта; прагматичен параметър – особености при употребата и степента на разпространение; илюстративен параметър; регистрационен параметър (датата на регистрация на дадения термин); интерпретационен параметър; параметър на понятийно-семантичната системност; параметър на лексикалносемантичната системност; времеви параметър; метадиалектен параметър (термините на дадена научна школа); идиолектен параметър (термините на даден учен); ареален параметър; категориален параметър; сведения за източниците.

Според ISO 30042:2008 **терминологична база** е „база от данни, която съдържа информация за специализирани езикови понятия и термините, които обозначават тези понятия, заедно с допълнителна информация“. В търговски условия обаче трябва да се отбележи, че терминологичната база не е ограничена до „специализирани езикови понятия“, а може да съдържа всяка лексикална единица, която трябва да се „управлява“, за да постигне качествено и последователно съдържание. Терминологичната база е база от данни с термини, отнасящи се до конкретна област (или проект), която е интегрирана в инструмент за компютърноподпомогнат превод.

Сами по себе си терминологичните бази са като речници, но начинът по който са създадени, управлявани и по който функционират, ги прави коренно различни.

Основната им характеристика е, че са персонални и персонализируеми за всяка институция, частна компания и отделен преводач. Техните създатели избират как да ги конфигурират, за кои езикови комбинации, кои термини да бъдат добавени, премахнати, забранени и как тези термини да бъдат класифицирани (Фабер и Араус 2021: 588).

Най-общо казано, към терминологичните бази могат да бъдат включени един или повече от следните компоненти (Райт 2001: 573): дефиниция (описание на значението на термина); източник; контекст на употребата на термина; област, към която принадлежи терминът[14]; граматическа информация (глагол, съществително и др.); етикет за употреба (например: фигуративен, американски английски, официален и др.); автор („създаден от“ – потребителят, който е добавил новия термин; в случай на различни потребители преводачът знае кой термин от кого е създаден / добавен); дата на създаване / модификация („създаден / модифициран на...“); статус („проверен“, „одобрен“, „предпочитан“ или „забранен, да не се използва“); връзки / линкове между отделните термини и дефиниции, както и линкове към външни източници; бележки.

Необходимо уточнение е, че критериите за подбор на термини зависят от изискванията на конкретния проект. Поради тази причина се предлагат следните стъпки за идентифициране на това дали даден термин е подходящ за конкретен проект:

- Терминът принадлежи ли към книжовния език, или не;
- Терминът използва ли се, или не;
- Дали терминът се използва с еднакво значение;
- Съществува ли риск от неправилно използване на термина;
- Дали използването на термина не изисква допълнителни препратки;
- Има ли синоними и какви разлики в значенията носи всеки един от тях.

Могат да бъдат намерени голямо количество свободнодостъпни терминологични ресурси. По долу са изброени някои от тях. Част са собственост на европейски институции, а други – на неправителствени организации.

- **ИАТЕ** (Interactive Terminology for Europe или IATE[15]) е най-известният терминологичен ресурс за преводачите в България, както бе потвърдено от анкетата сред тях. В платформата към юли 2022 г. могат да бъдат намерени общо 8 047 139 термина. Поддържат се всички официални езици на Европейския съюз.

[14] Докато при речниците биват изредени всички или поне най-популярните значения на даден термин, в терминологичните бази се предпочита еднозначността. Всяка база е създадена за конкретна област, клиент или проект. Не се взима под внимание какво означава даденият термин в други области.
[15] http://iate.europa.eu

- **Евровок** (EuroVoc[16]) е многоезиков и мултидисциплинарен тезаурус, обхващащ терминология от областите на дейност на ЕС. Също като ИАТЕ и този ресурс се предлага на 24-те официални езика на ЕС, както и на езиците на трите страни кандидатки (Албания, Северна Македония и Сърбия).
- **Електропедия** (Electropedia[17]) е база от данни на Международната електротехническа комисия (IEC) и предоставя най-големия брой термини и дефиниции в света в областта на електрическата енергия и електрониката, като съдържа повече от 22 000 терминологични записа на английски и френски, организирани по предметни области и с еквивалентни термини на различни други езици.
- **Голям терминологичен речник** (от фр. Le grand dictionnaire terminologique[18] или GDT) е банка от терминологични записи, създадени от Квебекския офис на френски език (от фр. Office québécois de la langue française). Всеки термин се отнася до понятие, свързано със специализирана област на употреба, като термините се представят на френски и английски, а понякога и на други езици.
- **Многоезиковата терминологична база на ООН** (United Nations Multilingual Terminology Database[19] или UNTERM) съдържа термини, свързани с работата на ООН. Информацията се предоставя на шестте официални езика на ООН (арабски, китайски, английски, френски, руски и испански), а има и записи на немски и португалски.

II. Сравнение на терминологичните бази с терминологичните речници

При сравнение на терминологичните бази и терминологичните речници в интернет могат да се определят следните общи черти между тях: а) Възможност за препращане към други термини чрез връзки между тях; б) Възможност за публикуване на изображения и други помощни средства в записа за всеки термин; в) Когато се сблъскат с проблем с терминологията, преводачите могат да получат бърз достъп до основни терминологични бази, както и до множество речници, ако търсената информация е налична; г) Възможност за актуализация на информацията.

Приликите обаче свършват дотук. По време на превод терминологичните бази улесняват, а много често и изобщо елиминират избора на преводача по отношение на

[16] https://op.europa.eu/en/web/eu-vocabularies
[17] https://www.electropedia.org/iev/iev.nsf/Welcome?OpenForm
[18] http://www.oqlf.gouv.qc.ca/
[19] http://unterm.un.org

даден термин. Стандартизирана практика на преводачите е да бъде изпратена предварително подготвена база с термините, одобрени от клиента, които да бъдат използвани в целевия текст. Ако се прави аналогия с терминологичните речници, даден клиент придобива ролята на съставител, защото той е този, който определя какво да се включи във всяка терминологична база. Друга разлика е достъпността. Терминологичните бази са достъпни само в инструментите за компютърноподпомогнат превод (освен в случаите, когато дадена база или част от нея е експортирана, за да се използва с конкретна цел). В повечето случай онлайн речниците са със свободен достъп. Не на последно място е различен начинът на работа. По време на превод в текстовия редактор на програмата за компютърноподпомогнат превод на преводача не му се налага да търси термини. Програмата сама ги визуализира.

Терминологичните речници могат да бъдат разделени на онлайн и печатни. Що се отнася до печатните, въпреки че са изправени пред голяма конкуренция, те далеч не трябва да бъдат отхвърляни. Докато терминологичните бази са индивидуални за всяка компания, проект и преводач, то традиционните терминологични речници и техните онлайн версии носят унификация и утвърждаване на значението на термините. Техните създатели обикновено са специалисти с утвърден авторитет. Последното не може да се каже за терминологичните бази, при които въпреки ясните инструкции за употреба или забрана на ползване на даден термин, не са изключени спорове между клиент и преводач за добавянето или премахването на конкретен термин, неговата употреба или значение.

Алтернатива на терминологичните бази са хартиените издания на терминологични речници. Пример за такива, включващи български език, са (не се цели изчерпателност при изброяването): Английско-български юридически речник, Христо Данов, издателство „Труд“, 1991; Английско-български терминологичен речник по дистанционни изследвания, Румяна Кънчева, АИ „Проф. Марин Дринов“, 2020; Английско-български речник по кардиология, Валентина Минчева, АИ „Проф. Марин Дринов“, 1991; Английско-български аграрен, бизнес и лесотехнически речник, Цветелина Цакова, АИ „Проф. Марин Дринов“, 2012; Английско-български морски речник с колокации, Вяра Петкова, издателство „Стено“, 2006; Многоезичен тематичен речник на европейската интеграция, издателство „Колибри“ от 2007.

Особеното при тези ресурси е, че всеки речник представлява различна тематична област и има различен обхват. Дори и в случаите, когато няколко терминологични речника представляват една дисциплина, съдържанието им и

достоверността им се сравняват трудно. Причините за това също са много разнообразни. Времевият диапазон в годините на публикуване може да бъде ограничаващ при използването им. Ограниченията в тиража също могат да бъдат пречка до достъпността на тези ресурси. Именно заради последната причина двуезиковите ресурси в интернет (включващи български език) също са алтернатива, например: Английско-български онлайн речник[20]; Английско-български онлайн речник на PONS[21]; SA речник[22].

Проблемите с терминологичните ресурси в интернет произтичат (с малки изключения) основно от липсата на информация за тяхната надеждност и достоверност. В повечето случаи не е ясно дали са създадени от експерти. Не е ясно също дали включените лексикални единици отговарят на критериите за термини и критериите за подбор. Разликата в обхвата и подбора на термините, както и на специалистите, които са ги подбирали, са друг много важен фактор.

III. Връзка на терминологичните бази с преводната памет

Преводната памет обикновено е интегрирана в системите за компютърноподпомогнат превод. Всяко преведено изречение се съхранява в нея. Когато преводачът срещне нов сегмент, който съответства на друг от базата, готовият превод се извлича и се предлага за повторна употреба (Тайбех 2008: 97).

Преводната памет може да бъде полезна за обогатяването на терминологичната база. Частичните съвпадения могат да помогнат при намирането на терминология, подобна на тази, използвана в терминологичната база. Процесът по създаване на терминологични бази с помощта на преводна памет не е автоматизиран. Докато добавянето на всички сегменти в паметта е задължителна стъпка, за да може да бъде завършен един превод, добавянето на термини е изцяло ръчен процес, който се основава на преценката или инструкциите на преводача. Някои терминологични бази имат допълнителното предимство да предлагат автоматични предложения за термин (automatic term suggestions). При тях програмата идентифицира възможните нови термини и ги предлага за добавяне на преводача.

[20] https://www.rechnik-bg.com/
[21] https://bg.pons.com/%D0%BF%D1%80%D0%B5%D0%B2%D0%BE%D0%B4
[22] http://notrial.bg/software/windows/sa-dictionary/

IV. Създаване и управление на терминологичните бази в системите за компютърноподпомогнат превод

Всяка съвременна система за компютърноподпомогнат превод включва наличието на функции за управление на терминология. Начините за създаване на терминологична база варират според използваната програма, но общото между тях е, че всяка програма за компютърноподпомогнат превод има настройки, който позволяват на потребителите да създават колкото бази пожелаят. В Традос, например, създаването на нова терминологична база е стъпка от създаването на нов проект. В програмата преводачът има възможност да преведе отделен файл или да създаде проект. Преди да започне самата работа по превод, преводачът или избира една от вече съществуващите бази, или създава нова. Това става в нов прозорец, където стъпка по стъпка преводачът избира дали да добави файлове, онлайн бази и т.н. Допълнителни настройки дават възможността за бързо импортиране (Fast Import), където файлът за импортиране е напълно съвместим с формàта Мултитерм ексемел (MultiTerm XML). По този начин се спестява времето за проверка на записите във файла за импортиране. Функцията за извършване на пълна реорганизация след импортиране (Perform full reorganisation after import) се използва при актуализирането на съществуващи терминологични бази с нови или променени термини. При този метод на работа не е изключено да се получат грешки след импортирането като нежелани дубликати или частичното търсене да не работи правилно.

Терминолозите се подпомагат от „интелигентни“ инструменти като софтуер за извличане на термини, автоматични индексиращи системи и програми за автоматично генериране на текст. През последното десетилетие различни автори изтъкват важни характеристики на това как модерните технологии като виртуална реалност и облачно базирана терминология (cloud-based terminology) (Варга 2012) се прилагат към терминологичната работа.

Освен стандартните функционалности, свързани с терминологичните бази в системите за компютърноподпомогнат превод, преводачите имат възможност да инсталират допълнителни помощни средства (**притурки**). Те не са част от основния софтуер, не всички програми за компютърноподпомогнат превод ги поддържат и много от тях са платени. Повечето се свързват към основната програма чрез плъгин, но някои

могат да работят и самостоятелно. Пример за такива притурки са[23]: Лоджитърм (LogiTerm)[24]; Джуреми (Juremy)[25]; Ердабълюес Апстор (RWS AppStore)[26].

Системите за управление на терминология (Terminology Management Systems или TMS) са предназначени за събиране, поддържане и достъп до терминологични данни. Могат да бъдат както независими от системите за компютърноподпомогнат превод, така и да са част от тях.

Съществуват и програми, посветени на локализацията. Тяхната структура се различава от инструментите за компютърноподпомогнат превод. Чрез тях се автоматизира преводът на уебсъдържание, мобилни приложения и файлово съдържание.

Контролираните езици се използват успешно в много области през последните 20 години като метод за подобряване на четимостта на техническите документи и за улесняване на превода им на езика на техните клиенти. „Контролираният речник е организираното подреждане на думи и фрази, използвани за индексиране на съдържание и/или за извличане на съдържание. Към контролирания речник могат да бъдат добавени предпочитани термини и техни варианти, както и да има дефиниран обхват или да се описва конкретна област“[27]. Предимството на контролираните езици е, че чрез тях авторите имат възможност да създават документи, които са лесно четими и по-последователни по отношение на използвания речник и стил.

V. Умения за работа с терминология и терминологични бази

Профилът на професионалния терминолог[28] обхваща знания и умения, свързани с познаването на принципите на терминологията (теоретични и практически), създаването и локализирането на терминологични ресурси за конкретни цели и групи, отлични езикови познания, изследователски умения и способност за идентифициране на важна информация, умения за използване на терминологични бази, както и отличното владеене на системите за управление на терминология, инструментите за компютърноподпомогнат превод, софтуера за машинен превод и инструментите за извличане на термини[29]. Трябва също да имат познания за онтологии и семантични

[23] Всички примери за такива програми са взети от списъка на Нимзи: https://www.nimdzi.com/tbs/
[24] https://terminotix.com/index.asp?content=category&cat=4&lang=en
[25] https://juremy.com/about
[26] https://appstore.rws.com/
[27] https://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/what.pdf
[28] https://termcoord.eu/why-terminology/31318-2/
[29] http://clearwordstranslations.com/language/en/7-terminology-management-best-practices/

мрежи. Специализирани умения като трансфер на знания и преподаване, познаване на специални предметни области, способност за планиране и управление на процеси и проекти, управление на промените, способността за вземане и прилагане на стратегически решения в координация с други звена в организацията[30] също са необходими.

Все по-голямо количество университети предлагат възможности за специализация и допълнителна квалификация. Примери за такива са: аспекти на превода и устния превод, свързани с терминологията[31], терминологията и професионални нужди[32], развитието на основни умения за превод и постигане на умения в разпознаването на регистри на различни езици[33], разработването на изследвания в области като техническия, литературния и медийния превод, конферентния превод и управлението на терминологията[34].

[30] https://terminorgs.net/downloads/Terminology_skills.pdf

[31] Магистърски курс по терминология на Люксембургския университет https://wwwen.uni.lu/studies/fhse/master_in_learning_and_communication_in_multilingual_and_multicultural_contexts/programme

[32] Онлайн програма за терминология към Института за приложна лингвистика, изследователски център към Университета Помпео Фабра https://www.upf.edu/en/web/terminologiaonline

[33] Магистърска програма по превод и терминологични изследвания в Малтийския университет https://www.um.edu.mt/courses/overview/PMTTFTT-2021-2-O

[34] Център за изследвания по провод към Виенския университет https://transvienna.univie.ac.at/en/about-us/

ТРЕТА ГЛАВА

Полуавтоматично създаване на многоезикови терминологични бази

Една област, която привлича значително вниманието, е автоматичното извличане на термини (Килгариф 2014 и др.; Коста и др. 2016). Най-ранните методи използват само езикова информация като морфо-синтактични модели. По-късно се разработват все по-сложни статистически методи, при които се прилагат различни видове честотни и колокационни алгоритми. На определен етап от развитието на автоматичното извличане на термини се появяват хибридни системи, съчетаващи и двата вида информация: езикова и статистическа.

Решението кои термини да се включат в терминологичната база, е основно предизвикателство и К. Уорбъртън препоръчва то да се основава на корпусни доказателства. Процесът по идентифициране на термините, включването им в терминологична база и добавянето на информация варира в зависимост от това дали подходът, който е избран, е тематичен (според този подход се приема, че термините и техните основни понятия са част от система, и следователно тяхното съществуване и значение могат да бъдат потвърдени само в сравнение с други термини в същата система), или така наречения ad-hoc подход, при който се решава конкретен проблем, отнасящ се обикновено само до едно понятие и неговото назоваване (Уорбъртън 2015: 652).

Технологиите за автоматично извличане не само позволяват на терминолозите да идентифицират повече термини, отколкото би било възможно ръчно, но и използването на тези технологии повишава съответствието между термините в терминологичната база и в корпуса, чието съдържание трябва да отразяват. Също така частично се намалява субективността, защото така се включват термини, които действително се употребяват в корпусите.

Езиковите ресурси за малки езици се намират трудно. При голяма част от междуезиковия трансфер (cross-lingual transfer) се използва английският като изходен език (Сьогард и др. 2015: 1713). „Навлизането на английски заемки се наблюдава и при общоупотребимата лексика, например *тостер*, *стикер*, *бодигард*, както и в младежкия жаргон“ (Благоева, Коева и Мурдаров 2012: 14 – 15).

I. Създаване на терминологични ресурси в областта на компютърната терминология

Определянето на „правилната“ терминология е свързано с разработването на терминологични ресурси и с тяхното управление. Преводачите използват терминология, за да направят преводите си по-точни и последователни.

Отличителна черта на **компютърната терминология** е постоянният напредък и развитието на технологиите, което изисква и осъвременяване и допълване на използваните термини (Хойзъл и др., 2015). Навлизането, разпространението и употребата на компютърна терминология не е нов феномен. Още в началото на новото хилядолетие Л. Кирова констатира, че: „Компютърната лексика има невероятно развитие за период само от двайсетина години. Колкото повече нараства броят на потребителите на компютърна техника, толкова по-определящо става тяхното участие във формирането на езика за работа с нея“ (Кирова 2002).

При термините, които са съставни думи, най-често срещаната синтактична структура е структурата от подчинено прилагателно и главно съществително (Коева 2007: 61) (например *риболовен сезон*), сравнително често се срещат и структурите от главно съществително, предлог и съществително, образуващи подчинена предложна група (например *подобряване на почвата*), всяко едно от съществителните – главно или подчинено може да се пояснява от прилагателни (например *електронен трансфер на фондове*), в някои случаи прилагателните може да са повече от едно (например *европейска парична система*), срещат се и словосъчетания от две съществителни, в които второто е главна част (например *тенис корт*).

Термините също така могат да бъдат класифицирани като неутрални от гледна точка на стилистичната си употреба (например *търсеща система*), професионализми или такива, които се споделят от определена група специалисти в дадена област (например *метатърсачка*), и професионален жаргон (например *паяк*) (Кирова 2004).

Абревиатурите се срещат често в (компютърната терминология). Както се посочва: „заимстваните от английски абревиатури започват да функционират като пълнозначни думи и все по-често се налага да влизат в тълковните речници“ (Кирова 2004). Посочва се, че се наблюдават различни случаи на предаване на чужда абревиатура на български език – в някои случаи се възприема съкращението (*джиесем*, *есемес*), а в други може да се използва както абревиатурата, така и разгърнатото название (*сиди*, *компактдиск*) (Колковска 2010: 2 – 3).

Хартиените **речници на английски и български в областта на компютърната терминология** имат различен обхват и структура, както и за едни и същи термини да бъдат дадени различни дефиниции и др. Отпечатаните речници с ясно показано авторство са по-авторитетни от гледна точка на качеството на включената информация. От друга страна, дори да се преиздават, хартиените речници остаряват много бързо. Проблемът с речниците в интернет пък е несигурността в тяхното качество, липсата на информация как са подбрани единиците, настина ли принадлежат към компютърната терминология. Поради тези причини събирането на едноезикови речници на английски и български в областта на компютърната терминология също може да помогне.

Речниците с компютърна терминология на български език се отличават с малък брой термини, липса на представителност и авторство, некоректно представен в много случаи от гледна точка на българския правопис термин или липса на представяне на български език и дефиниции.

В мрежата и книжарниците могат да бъдат намерени редица речници на английски и български в областта на компютърната терминология. Примери за такива са:

- *Компютърен речник A-Z*[35] (хартиено издание от 2003 г.;
- *Английско-български компютърен речник*: A-Z[36] (хартиено издание от 2017 г.;
- *English-Bulgarian Computer Terms*[37] (база от данни, включваща 729 термина на английски и преводите им на български;
- *Преводно-тълковен речник на някои компютърни термини*[38] (сайт, в който не са посочени източниците на информация и няма цитирана литература;
- Николова Д. *Словарь компьютерных терминов / Речник на компютърните термини*. Шумен: Химера, 2016. 258 с. ISBN 978-619-7218-27-5. Речникът представя популярни термини в сферата на информационните технологии: устройство на компютъра, оборудване; компютърни технологии; програмни продукти; работа с компютър; интернет;

[35] https://m.helikon.bg/43115-Компютърен-речник-A-Z.html [20.07.2022]
[36] https://www.book.store.bg/p5397/anglijsko-bylgarski-kompiutyren-rechnik-a-z.html [20.07.2022]
[37] https://elrc-share.eu/repository/browse/english-bulgarian-computer-terms-processed/f9d2a4e44afb11e9a7e100155d02670615329808b1384ab388941fcd4fba5029 / [20.07.2022]
[38] http://kafence.com/Преводно-тълковен-речник-на-някои-компютърни-термини-201.html [20.07.2022]

Макар че това представяне вероятно не обхваща всички налични речници, представящи компютърна терминология на английски език, се очертават следните тенденции: речниците, създадени от големи издателства или организации, съдържат богата и достоверна информация; наред с това съществуват множество източници в интернет, които предлагат информация, но само специалисти могат да удостоверят в някои случаи дали термините и дефинициите са подходящи.

За български също са достъпни речници за компютърната терминология, които обаче са електронни:

- *Компютърни термини, които трябва да знаем*[39] (кратък списък с термини);
- *Речник на компютърните термини и акроними*[40] (съдържа над 250 термина и техните дефиниции;
- *Речник на компютърни термини по азбучен ред*. *Основни компютърни термини. Програмата и нейното инсталиране на компютър*[41] (речникът се намира в сайт с новини за информационни технологии, отново някои термини и дефиниции не са предадени подходящо на български език, а също така се наблюдават и повторения на грешките от други източници.

Речниците с компютърна терминология на български език се отличават с малък брой термини, липса на представителност и авторство, некоректно представен в много случаи от гледна точка на българския правопис термин или липса на представяне на български език и дефиниции, които също може да не представят адекватно описаното понятие, а в някои случаи има ясни индикации за машинен превод.

Въпреки че някои от изброените речници са полезни, те могат да затруднят преводачите при тяхната работа поради различния си обхват, различната структура, както и най-вече за това, че за едни и същи термини са дадени различни дефиниции (дори преводачите да са наясно за качеството на даден речник, който използват). Още един фактор, който прави съществуващите речници не винаги подходящи за създаването на терминологични бази, е специализираната насоченост на конкретния превод.

Въпросът за **автоматичното извличане на термини** не е нов. Още през 1996 г. К. Кагеура и Б. Умино правят преглед на принципите и методите за автоматично

[39] https://urocibg.eu/компютърни-термини-които-трябва-да-зн/ [20.07.2022]
[40] https://bg.chalized.com/речник-на-компютърните-термини-и/ [20.07.2022]
[41] https://roidmod.ru/bg/lighting/slovar-kompyuternyh-terminov-po-alfavitu-osnovnye-kompyuternye-terminy-programma-i-e-ustanovka-n/ [20.07.2022]

разпознаване на термини (Кагеура и Умино 1996). Те показват две основни тенденции в извличането на термини: едната е извличането на информация (information retrieval), а другата е автоматичното разпознаване на термините (automatic term recognition).

Автоматичното извличане на термини става от езикови корпуси, които трябва да представляват дадена тематична област. Търсенето и колекционирането на подходящи текстове може да не е лесна задача, защото от тях зависи и качеството на извлечените термини. За нуждите на автоматичното извличане на термини също така автоматично или полуавтомаично се събират големи количества документи. Както беше посочено, за някои езици достъпните ресурси в електронна форма може да не са в големи количества. Някои методи за извличане на термини изискват и голям референтен корпус, който да илюстрира общоупотребимата лексика и сравнението с който да позволява адекватната идентификация на термини. Извличането на термини в два и повече езика е свързано с подравняването на паралелни изречения, думи и фрази в паралелни корпуси.

Въпросът за автоматичното извличане на термини не е нов. Още през 1996 г. К. Кагеура и Б. Умино правят преглед на принципите и методите за автоматично разпознаване на термини (Кагеура и Умино 1996). Те показват две основни тенденции в извличането на термини: едната е извличането на информация (information retrieval), а другата е автоматичното разпознаване на термините (automatic term recognition). А. Шаятович и др. разделят методите за извличане на термини на: честотни методи (които се основават на предположението, че по-високата честота на кандидата за термин предполага и по-голяма вероятност той да е действително такъв); контекстни методи (основават се на контекста, в който се появява терминът кандидат); тематично моделиране (тематичната информация може да се използва вместо честотата) и повторно ранкиране (re-ranking) (методите от тази група използват други методи за извличане на термини като характеристики и се стремят да оценят важността на всяка характеристика) (А. Шаятович и др. 2019: 150).

Някои инструменти използват статистически подход към текста, така че те просто търсят текстовите повторения. По този начин обаче има риск да бъдат извлечени невалидни кандидати за термини (известни са като „шум“, noise). Друг подход при извличането на термини е лингвистичният, според който програмата търси структури като „съществително + съществително“ или „прилагателно + съществително“ и т. н. Друг пример за извличане на термини е инструментът на Тием

таун[42] (TM-Town), платформа за сътрудничество между преводачи и клиенти (Фигура 15). Програмата автоматично извлича ключови термини от изходните и целевите текстове. Тези термини са оценяват според алгоритъм, който изчислява уместността на всеки термин в документа. Това става чрез статистически алгоритми за обработка на естествения език, които анализират съдържанието на документите и идентифицират най-подходящите ключови думи. По-високият търм ранк резултат (Term Rank) показва по-висока важност на термина в документите. След извличането на основните термини от целевия текст потребителят съпоставя изходния термин със съответния му превод.

При повечето **инструменти за извличане на термини** от два езика се идентифицират термините кандидати в изходния език въз основа на предварително дефинирани модели. След това се избират еквивалентни термини кандидати в превода на целевия език. Един от алтернативните подходи е генерирането на кандидат термини директно от подравнените думи и фрази в корпуса. Чрез честотна информация (frequency information) и n-грами се определя кои думи или съставни думи са термини. При този подход не се вземат под внимание зависими от езика предварително дефинирани модели, а се свързват фрази, основаващи се на лексикалните съответствия и синтактичното сходство (Лефевер 2009: 496).

Сравнението на инструментите за автоматично извличане на термини показва, че свободнодостъпните инструменти не са много, а тези, които предлагат извличане на термини на български език, са още по-малко. Файффилтърс (fivefilters.org[43]) предлага функционалност за извличане на терминология чрез поставяне на текста за обработване или на интернет връзката към него в определеното за това място. Възможно е извличането на най-много 100 термина в пет различни формàта. Този инструмент поддържа и български език.

В проучване от 2016 А. Зарецкая и др. (2016: 18) предлагат **сравнение на инструментите за автоматично извличане на термини**. Анализират се следните самостоятелни програми, които се определят като „най-популярния тип софтуер“ към момента. По-конкретно, сравняват се следните функционалности: *Двуезиково извличане* (Bilingual extraction) за едновременно извличане на термини от изходен и преведен тест; *Срявняване на контекста в изходния и целевия език* (Source and target context comparison); *Проверка на термин* (Terms validation); *Двуезиково попълване на речник* (Bilingual dictionaries compilation) за създаване на двуезиков речник; *Контекстно*

[42] https://www.tm-town.com/ [20.07.2022]
[43] https://www.fivefilters.org/term-extraction/ [20.07.2022]

извличане на термини (Context extraction); *Поддръжка на различни файлови формàти* (Support various file formats); *Честотно подреждане на термини* (Rank terms by frequency); *Поддръжка на различни езици*; *Определяне на минимален брой срещания на термин* (Specify the minimum number of occurrences) за задаване минималния брой на срещания на даден термин; *Показване на лингвистична информация* (Show linguistic information); *Определяне на минимален и максимален брой преводи* (Specify the maximum number of translations) за броя преводни еквиваленти на даден термин; *Опция за изключване* (Stopword list option) за изключване на думи, които не могат да бъдат термини (например предлози); *Избор на минимален и максимален брой думи в състава на термин* (Choose the minimum and maximum number of words per term); *Статистика* (Term statistics).

От сравнението между различните програми става ясно, че всичките включват статистическа информация. Само три от общо девет предлагат двуезиково извличане. ЕсДиЕл Мултитърм предоставя възможност за срявняване на контекста в изходния и целевия език. Функционалности като контекстно извличане на термини, поддръжка на различни файлови формати, честотно подреждане на термини, определяне на минимален брой срещания на термин и поддръжка на различни езици притежават всички програми без Транслейтед. Показване на лингвистична информация се поддържа от Терминус, ЕсДиЕл и Търмсюит. Определянето на минимален и максимален брой преводи е функционалност на Термюит. Общо 5 от 9 програми (ЕсДиЕл Мултитърм, Симпъл Екстрактор, Киа, Рейнбол и ЯТЕ) предоставят възможност за изключване на термини и за избор на минимален и максимален брой думи в състава на термин.

По модела на направеното проучване от А. Зарецкая и др., през 2022 г. за целите на настоящата дисертация бяха анализирани същите инструменти, за да се провери тяхното развитие. Новите функционалности са: *Многоезиково извличане на термини* (Multilingual term extraction); *Проверка на съответствието между текста и терминологичната база* (Consistency check between the text and the termbase) за проследяване дали се използва терминологията от терминологичната база; *Възможност за ръчно извличане на термини* (Manual terms extraction); *Настройки към проект* (Project settings); *Извличане на доклад за качество* (QA report); *Едноезиково извличане* (Monolingual extraction) за извличане на термини само от изходния или от преводния тест; *Работа с външен софтуер* (Third-party software access); *Определяне на максимален брой извлечени термини* (Select Maximum number of extracted terms);

Двуезиково подравняване (Bilingual alignment) за съотнасяне на най-добрия превод на термин от изходния език на целевия език; *Възможност за търсене с регулярни изрази* (Regex search); *Концептуална структура* (Conceptual structure) за представяне на концептуално дърво.

От описаните нови функционалности СкечЕнджин вече разполага с контекстно извличане, показване на лингвистична информация и опция за изключване на термини; ЕсДиЕл и Рейнбол поддържат определяне на минимален и максимален брой преводи; Транслейтед предлага възможност за избор на минимален и максимален брой думи в състава на термините.

II. Методика за полуавтоматично създаване на терминологична база в областта на компютърната терминология

Съществуват голям брой инструменти за извличането на терминология. Българският не се поддържа от голяма част от тях. Положението е още по-сложно, когато стане въпрос за някоя специализирана област: наличните езикови ресурси за български също не са много. Тези ограничения са причина да се предложи методика за полуавтоматично създаване на терминологична база в областта на компютърната терминология.

Както вече стана ясно, първата стъпка в създаването на една терминологична база е намирането на достоверен източник. Предвид описаните по-горе ограничения и особености при **извличането на термини,** най-подходящият ресурс е **двувуезиковият корпус**.

За целите на дисертацията е подбран паралелният английско-български корпус в областта на компютърната терминология КюТиЛиип (QTLeap V1.2)[44], който съдържа 4 000 двойки въпроси и отговори в областта на отстраняването на проблеми, свързани с информационните технологии (както хардуерни, така и софтуерни). Трябва да се отбележи, че има няколко недостатъка: корпусът е малък, текстовете и термините в него не са разнообразни, броят на специализираните термини не е голям. Лицензът за употреба на корпуса е позволява академична и некомерсиална употреба, изисква цитиране на авторството и същия начин на споделяне (CC-BY-NC-SA). Създаден е в периода 2014 – 2015 година с европейско финансиране. Редактиран е за последно през 2016 година. Корпусът е избран, защото е двуезиков и отговаря на подбраната

[44]http://metashare.ilsp.gr:8080/repository/browse/qtleap-corpus-v12/0176c39ae9cd11e4a2aa782bcb074135ba7d767f645a48dca1d50ee3c9504253/ [20.07.2022]

терминологична област. Трябва да се отбележи, че има няколко недостатъка: корпусът е малък и текстовете и термините в него не са разнообразни. Трябва също да се отбележи, че броят на специализираните термини не е голям, а се срещат по-скоро термини, които чрез разпространяването на технологиите в ежедневието, са навлезли в употреба.

За да бъда избран конкретен инструмент за **извличане на термини от корпус**, той трябва да отговаря на следните изисквания: да поддържа български и изходния език от/на който се превежда (задължително); да бъде безплатен (препоръчително); да бъде самостоятелна програма или вградена функционалност на (безплатна) система за компютърноподпомогнат превод (задължително за хора без допълнителни умения за програмиране); да бъде с интуитивен интерфейс (препоръчително); да притежава настройки за избор на минимален и максимален брой думи, от които да се състои даден термин (препоръчително); да предлага извлечените термини във формат, който позволява лесното им конвертиране. (препоръчително).

След направеното проучване се установи, че на посочените критерии отговаря Файффилтърс. Програмата се справя с извличането на английски термини. За български предлага незадоволителни резултати, въпреки че е езиково независима. Програмата разпознава само имената на продукти и търговски марки, изписани на латиница.

Ключов компонент за това дали дадена програма за извличане на термини ще намери приложение, е дали поддържа български, което обаче към момента (особено за свободнодостъпни инструменти) не е често срещано. Програми като Мултитърм поддържат огромно количество езици, но са платени и тяхното използване е пряко обвързано с лицензите за Традос. По този начин, използването на Мултитърм е ограничено до тези преводачи, които разполагат с лиценз.

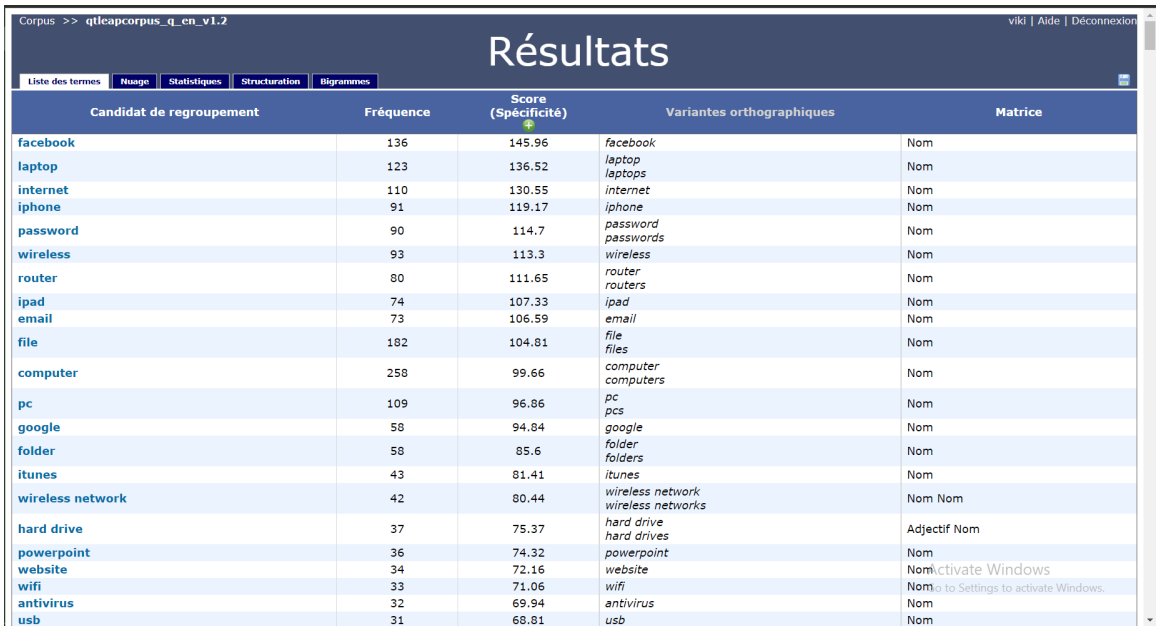
Въпреки наличието на огромен брой инструменти за извличане на терминология, намирането на такъв, който да отговаря на посочените изисквания, е доста трудна задача.

ИАКЕ[45] (Yet Another Keyword Extractor или YAKE) използва множество статистически функции и е езиково независима програма. Също отговаря на заложените критерии и по тази причина бе направен опит за извличане на

[45] http://yake.inesctec.pt/demo/user [20.07.2022]

терминология с нея. Резултатите отново не са добри, защото програмата разпознава само имената на продукти (Facebook, Google Chrome) и няколко български глагола.

Предвид описаното по-горе, бе променен подходът за извличане на термини, като описаните критерии бяха насочени към работа с комбинация от програми, а не с отделен инструмент, с който да се извлекат термините и на двата езика. За целта бе избрана програма, поддържаща извличане на термини на английски език: Термостат Уеб 3.0[46] (TermoStat Web 3.0). Програмата извлича единични термини и термини съставни думи, показва тяхната честота, степента на специфичност и формите на думите, които се срещат в текста (Друан 2003).


Corpus >> qtleapcorpus_q_en_v1.2 — viki | Aide | Déconnexion

Résultats

Liste des termes | Nuage | Statistiques | Structuration | Bigrammes

Candidat de regroupement	Fréquence	Score (Spécificité)	Variantes orthographiques	Matrice
facebook	136	145.96	facebook	Nom
laptop	123	136.52	laptop laptops	Nom
internet	110	130.55	internet	Nom
iphone	91	119.17	iphone	Nom
password	90	114.7	password passwords	Nom
wireless	93	113.3	wireless	Nom
router	80	111.65	router routers	Nom
ipad	74	107.33	ipad	Nom
email	73	106.59	email	Nom
file	182	104.81	file files	Nom
computer	258	99.66	computer computers	Nom
pc	109	96.86	pc pcs	Nom
google	58	94.84	google	Nom
folder	58	85.6	folder folders	Nom
itunes	43	81.41	itunes	Nom
wireless network	42	80.44	wireless network wireless networks	Nom Nom
hard drive	37	75.37	hard drive hard drives	Adjectif Nom
powerpoint	36	74.32	powerpoint	Nom
website	34	72.16	website	Nom
wifi	33	71.06	wifi	Nom
antivirus	32	69.94	antivirus	Nom
usb	31	68.81	usb	Nom

Фигура 2: Част от извлечените термини от корпуса с програмата Термостат

Както се вижда от Фигура 2, програмата извлича термините, показва тяхната честота, специфичност, определя към коя част на речта принадлежат и извлича техните форми, които се срещат в текста. Заложените критерии при автоматичното определяне на термини са: да се извлекат само съществителни (в това число и съставни), да се извлече максималният брой разпознати от програмата термини. Разбира се, в списъка с извлечени английски термини присъстват и думи, които не са термини, и това наложи да бъдат отстранени ръчно. Например такива са *wont, much ram, doesnт*, *someone*, *much space* и *old file*.

[46] http://termostat.ling.umontreal.ca/interfaceTermostat.php [20.07.2022]

Към функционалностите за разработване на програма, насочена към български, може да се добави и възможността за едновременно извличане на двуезикова терминология. Такава възможност има (разработена от латвийската компания Тилде[47]), за която са необходими умения за програмиране, което ограничава възможността за по-широко приложение[48].

След идентифицирането на английските термини, са намерени преводите им от българската част на използвания корпус. За целта английският и българския текст са подравнени с Мейткат (MateCat[49]), програма, която също отговаря на заложените критерии за избора на подходящи инструмент. След като текстовете се визуализират в програмата Мейткат, така че изреченията да се едно до друго, английските термини се търсят като ключова дума, за да може да се намери техният български еквивалент[50].

След определянето на еквивалентните термини за български, резултатите са поместени в екселска таблица. Въпреки че Термостат позволява извличането на файлове в .txt формат, екселската таблица е избрана поради своята достъпност и затова че файловите формати могат да бъдат директно импортирани във всяка програма за компютърноподпомогнат превод. Броят на извлечените английски термини съответства за този на българските. Съкращенията в английски език са предадени непроменени на български.

Беше направена **ръчна проверка и техническо оформление на терминологичната база**. Според изискванията на клиента или техническите възможности на програмата за компютърноподпомогнат превод форматите на терминологичната база може да варират. Съществуват и редица самостоятелни средства, с които извлечената терминологична база може да се преобразува от един формàт в друг. Пример за такива са: ТиБиЕкс Конверт (TBX Convert[51]), Глосари Конвъртър (Glossary Converter[52]), Ексбенч (ApSIC Xbench[53]). За настоящия експеримент бе използвана десктоп програмата Олифант (Olifant[54]).

[47] https://www.tilde.com/
[48] https://github.com/tilde-nlp/taws
[49] https://www.matecat.com/ [20.07.2022]
[50] Подравняването може да се използва за всякакъв вид търсене на преводни еквиваленти с различна цел, например стилистична.
[51] https://www.tbxconvert.gevterm.net/glossary/index.html [20.07.2022]
[52] https://appstore.sdl.com/language/app/glossary-converter/195/ [20.07.2022]
[53] https://www.xbench.net/index.php/download [20.07.2022]
[54] http://okapi.sourceforge.net/downloads.html [20.07.2022]

След конвертирането в желания формат (в този случай в tmx), терминологичната база може да бъде импортирана в проекта за превод. Като инструмент за съотнасяне и за превод бе избран Мейткат (MateCat).

Необходими уточнения са, че:

1) Конвертирането от един формат в друг за създаване на терминологична база не винаги е необходимо. Достатъчно е документът да е оформен както трябва (в две съседни колони, с езиковия код начело на всяка колона и съответно термините на изходния и на целевия език;

2) Предложеният метод не е оптимален и не решава всички въпроси. Показва само един от възможните пътища за решаване на проблема в условията на липса на достатъчно ресурси и програми за обработка на български език. Цел за последващи изследвания е разработването на инструмент, който да може да отговори още по-адекватно на нуждите на преводачите.

III. Описание на характеристиките на създадената терминологична база от данни в областта на компютърната терминология

Новосъздадената терминологична база съдържа извлечените от корпуса термини на английски език и техните еквиваленти на български език. Добавени са три допълнителни полета (за синоними, бележки и части на речта), но друг вид информация като дефиниции, изображения, препратки между термините или към външни източници не се предоставя. Към момента на един изходен термин съответства само един термин от целевия език.

От морфологична гледна точка в базата присъстват прости и съставни думи и всички термини са съществителни. Извлечените термини варират с максимална честота в паралелния корпус от 136 до минимална честота 3.73.

При прегледа на извлечените резултати, може да се направи следният езиков анализ: в английски език има сложни думи, които се предават в български като две отделни думи (например *cellphone – мобилен телефон*); съкращенията от английски или се предават по същия начин (например *wifi – wifi*[55]), или са предадени с цялото им изписване на български (например *pc – настолен компютър*). Забелязват се сложни думи на английски, които се предават на български като съставни (напр. antivirus – антивирусна програма). От български към английски не се забелязват много такива

[55] Необходимо уточнение е, че са възможни правописни грешки и в изходния език, които после се прехвърлят в целевия. Има риск да повлияят на цялостното качество на превода на текста.

примери (*url – URL адрес*). Срещат се термини, образувани чрез калкиране, например *playlist – плейлист*, както и запазване на съкращения като *ram – рам*.

Необходимо уточнение е, че в новосъздадената база не са предложени синоними към термините, както и че няма дефиниции или обяснение за частите на речта. Причината за това е, че не е добра практика при превод на термини да се използват синоними освен ако няма конкретна причина или инструкция за това. Например терминът *рутер* има синоним в български думата *маршрутизатор*, терминът *изтегляне* има синоним *сваляне,* но в зависимост от предназначението на превода може да се използва по-често употребявания термин или терминът, който се предпочита в книжовната реч.. Въпреки че синонимите показват богатството и развитието на езика, те не са взети под внимание при създаване на терминологична база, тъй като заради стандартизираните и еднотипните текстове е задължително спазването и последователната употреба на едни и същи термини за конкретен проект и текст.

Предложената в тази глава методика за разработване на терминологични бази не предлага универсално решение за подобряване и улесняване на работата на преводачите във връзка с уместното използване на терминология. Тя има своите ограничения (ръчния подбор на корпуси и ръчното търсене на преводни еквиваленти) и не претендира за изчерпателност особено ако се сравни с възможностите на програмите за извличане на термини за други езици. Въпреки това е първи опит в тази посока с цел да се подпомогне разработването на специализиран софтуер за извличане на термини и създаването на двуезикови терминологични бази, в които единият език е български.

ЧЕТВЪРТА ГЛАВА

Измерване на качеството в системите за компютърноподпомогнат превод и техните компоненти

Измерването на качеството на превода е изключително трудна задача, защото няма само един идеален превод за даден текст. Например, юридическият превод има много по-различни изисквания по отношение на точността и спазването на специфични законодателни норми в сравнение с рекламен текст или ръководство за употреба. Спецификите на преводния текст зависят не само от преводача, от текста, който ще бъде преведен, или от субективната му интерпретация, но и от причините за превода, потенциалните читатели и публикационната и маркетингова политика, т.е. от фактори, които надграждат превода като езикова процедура. Тук спадат и социалните фактори, както и социокултурните, политическите или идеологическите ограничения, които могат да имат значително влияние върху превода (Хаус 2015: 153). Според С. Колина „Липсата на консенсус за това как може да се определи качеството на един превод, произтича от разногласията относно концепцията за превод и от противоречивия и относителен характер на качеството…“ (Колина 2020: 458). Според К. Чуню и Уонг Так-минг, оценката на качеството на превода изисква разбирането на съдържанието на текста да се използва за определяне на различните видове отношения на еквивалентност и идентифицирането на грешки в превода (Чуню и Уонг Так-минг 2015: 220 – 221). Голяма част от критиките към различните подходи за оценка се отнасят до тяхната зависимост от разбирането за „еквивалентност“. Още в края на 90-те години Х. Хьониг констатира, че приемането на еквивалентността като априорна мярка за качество е неправилно, защото видът и степента на еквивалентност варират в зависимост от условията, при които се извършва преводът (Хьониг 1997: 9).

Друга трудност е, че оценката за качеството на превода се извършва от хора, като по този начин субективният компонент на „човешкия фактор“ е още по-силно изразен (Зехналова и др. 2013: 43). Така оценката на превода не разчита на конкретна теория, а на лични възгледи (Колина 2020: 458). Качеството на превода може да се разглежда и като степента, до която преводът следва договорените спецификации (Дурбан и Мелби 2008[56]). Този поход всъщност следва препоръките на международните стандарти за качество. Затова качеството на превода трябва да бъде дефинирано като

[56] https://www.communicaidinc.com/a-10-strategic-translation.php

относителна (а не абсолютна) адекватност по отношение на рамка, предварително договорена между възложителя и преводача. Преводачът не може да приключи превода, без преди това да е направил проверка на качеството (в зависимост от настройките и проекта за превод) и разрешил евентуалните грешки и несъответствия. Контролът на качеството се извършва на първо място от човека, който е извършил превода, – самия преводач, а след това от редактора, възложителя или от други оператори (Гаудек 2007: 74).

I. Измерване на точността на превода със системите за компютърноподпомогнат превод

Трябва да се отбележи, че досега няма компютърна програма, която да замени професионалния редактор. Езиковите технологии може само да улеснят работата на редакторите. Ако програмата за проверка на качеството на превода не е открила никакви грешки, това не гарантира, че преводът е (напълно) правилен. Не всички грешки обаче винаги са истински „грешки“. Възможно е сегментите, които софтуерът разпознава като съдържащи грешка, всъщност да са правилни. Преводачът (или редакторът) трябва да вземе решение за всеки отделен случай. Тъй като грешките в целевия език могат да окажат значително влияние върху качеството на превода, тяхната оценка включва нарушения в структурата на целевия език и фразеологията, които трябва да се разглеждат като „грешки в превода“, т.е. общи проблеми, които се наблюдават при създаването на текстове (Хансен 2010: 386).

За да се провери измерването на точността на превода с терминологични бази, както и за да се демонстрират възможностите за проверка на качеството в системите за компютърноподпомогнат превод (с фокус върху българския език), са проведени два експеримента: за първия бяха избрани 10 специализирани текста (с обем от около 15 000 думи) от регламентите на ЕС[57], преведени от експерти от английски на български, за превода на които се използва вече готова база; за втория бяха използвани текстове от корпуса (около 20 000 думи приблизително), използван за създаване на терминологичната база, описана в Трета глава.

Що се отнася до първия експеримент, английските текстове и българските им преводи са подравнени с инструмента Традос алайнер (Trados aligner), а терминологията е извлечена ръчно (общо 50 термина). След това термините са

[57] https://eur-lex.europa.eu/homepage.html

представени в основна форма както в изходния, така и в целевия език, защото лематизацията е метод, който се използва при създаването на речници и терминологични бази. Към тестовия корпус са въведени следните допълнителни грешки:

- изтриване на термин (като симулация на случая, в който даден термин не е преведен от преводача);
- частично представяне на термин (когато терминът е съставна дума), състоящо се в изтриването на една от думите, част от термина;
- непоследователен превод (използване на различни термини за предаване на едно и също понятие).

Всички въведени грешки са възможни в реална ситуация.

За проверка на изкуствено въведените грешки при употребата на терминология бяха използвани Традос и мемоКю (най-популярните програми за компютърноподпомогнат превод). За улеснение, всички останали настройки за проверката на качеството и на двете програми са изключени с изключение на проверката на терминологията. С настройките си по подразбиране програмите засичат само термина в основната му форма. Експериментът е извършен от двама преводачи с Традос и мемоКю на различни компютри.

По-долу са показани резултатите от извлечените отчети за качество от тестваните програми. Въведените грешки се откриват с някои изключения. При Традос, например, в изходния език програмата разпознава само „наказателно производство“ и игнорира множественото число на термина, което е и очаквано поведение[58]. Същата ситуация се повтаря и в целевия език. Всеки инструмент за компюърноподпомогнат превод разполага с функционалност за проверка на качеството (**измерване на точността на превода без терминологични бази)**. Контролът на качеството се извършва на първо място от човека, който е извършил превода, – самия преводач, а след това от редактора, възложителя или от други оператори (Гаудек 2007: 74).

За да се провери **измерването на точността на превода с терминологични бази**, както и за да се демонстрират възможностите за проверка на качеството в системите за компютърноподпомогнат превод (особено що се отнася до българския език), са проведени два експеримента: за първия са избрани 10 специализирани текста (15 000 думи приблизително) от регламентите на ЕС[59], преведени от експерти от

[58] Това е често срещан проблем и при други езици като шведски и фински.
[59] https://eur-lex.europa.eu/homepage.html

английски на български, в които се използва вече готова база; за втория са използвани текстове от паралелния корпус (20 000 думи приблизително), приложен за създаване на терминологичната база, описана в Трета глава.

Към тестовите корпуси от двата експеримента са въведени следните допълнителни грешки: **изтриване на термин** (като симулация на случая, в който даден термин не е преведен от преводача); **частично представяне на термин** (когато терминът е съставна дума), състоящо се в изтриването на една от думите, част от термина; **непоследователен превод** (използване на различни термини за предаване на едно и също понятие).

Всички въведени грешки са възможни в реална ситуация. За проверка на изкуствено въведените грешки са използвани системите за компютърноподпомогнат превод Традос и мемоКю. Въведените грешки се откриват с някои изключения. Анализът показва, че Традос разпознава в изходния език термина само в единствено число, което е и очаквано поведение[60]. Същата ситуация се повтаря и в целевия език. Например:

Изходен език: *...this Directive aims to strengthen the trust of Member States in each other's criminal justice systems and thus to improve mutual recognition of decisions in criminal matter**s**.*

Целеви език: *...настоящата директива има за цел да засили взаимното доверие на държавите членки в техните системи за наказателно правосъдие и по този начин да спомогне за постигане на по-широко взаимно признаване на решенията по наказателни производства.*

Вторият експеримент имаше за цел техническа проверка на методиката за създаване и проверка на терминологична база за компютърна терминология, за да се дали може да се използва от системите за компютърноподпомогнат превод без технически затруднения. Проверката на термините е по модела на първия експеримент и обхваща същите изкуствено въведени грешки като изтриване на термин, частично изтриването на една от думите, част от термина и последователно използване на различни термини за предаване на едно и също понятие. Термините бяха коректно визуализирани и резултатът беше същият както при предходния експеримент. Програмата засича липсата на термините или частичното им присъствие във всички изкуствено въведени грешки. Повтаря се обаче ситуацията, описана по-горе, с

[60] Това е често срещан проблем и при други езици като шведски и фински.

неразпознаване на множественото число на термините, което показва, че всяка форма трябва да присъства като запис в терминологичната база

Проведените експерименти проверяват функционалностите за проверка на качеството в системите за компютърноподпомогнат превод за български език, както и поведението на новосъздадената терминологична база. Чрез изкуственото добавяне на грешки, срещани в реална работна среда, бе възможно да се покаже, че терминологичната база може успешно да бъде използвана.

II. Сравнение на ефективността на програмите за оценка на превода

Програмите за проверка на качеството, външни на инструментите за компютърноподпомогнат превод изпълняват същата функция като вградените с тази разлика, че разполагат с много по-широк спектър от допълнителни функционалности. Следните програми са сред най-разпространените: Кюей дистилър[61] (QA Distiller), Ексбенч[62] (Xbench), Верифика (Verifika), Ерърспай (ErrorSpy) и Лингуистик тулбокс или Елтиби[63] (Linguistic ToolBox или Ltb). Основното предимство на тези програми е, че възможните грешки, свързани с терминологията и съгласуването, се отбелязват по различен начин. Функционалности на изброените по-горе програми за откриване на грешки: *Празни сегменти* (Empty segments); *Целевият текст съвпада с изходния текст* (Target text matches the source text) за съответствие между целевия и изходни език; *Несъответствие* в таговете (Tag mismatch); *Несъответствие в цифрите* (Number mismatch); *Граматична грешка* (Grammar); *Несъответствие в линковете* (URL mismatch); *Правописни грешки* (Spelling); *Буквено-цифрово несъответствие* (Alphanumeric mismatch); *Липса на двоен символ* (Unpaired symbols) за наличие на отварящ или затварящ символ, например кавички; *Частичен превод* (Partial translation) за минимален брой непреведени последователни думи; *Повторен празен символ* (Double blanks); *Повтарящи се думи* (Repeated words); *Последователност в изходния текст* (Source consistency) за еднакъв превод на различни изходни сегменти; *Последователност в целевия текст* (Target consistency) за еднакви сегменти в изходния език с различен превод; *Доклад за извършени промени* (Change report); *Обработка на множество файлове едновременно* (Multiple files); *Камелкейс* (CamelCase) за съответствие на съкращения; *Терминология* (Terminology) за последователна употреба

[61] http://www.qa-distiller.com/en
[62] https://www.xbench.net/
[63] http://autoupdate.lionbridge.com/LTB3/

на термините в целия текст; *Списък за проверка* (Checklists) – предварително уговорен списък между преводача и клиента, в който са описани стъпките за гарантиране на качеството; *Силно търсене* (PowerSearch) – режим на търсене, който използва регулярни изрази; *Профили* (Profiles) – персонализирани настройки за проверка на качеството, правописа и граматиката, които са избрани за конкретен потребител; *Доклад* (Report) за визуализация наоткритите от програмата грешки в текста; *Команден панел* (Command line) за работа без графичния потребителски интерфейс на програмата; *Списък от думи, които не трябва да се превеждат* (Do Not Translate List или DNT List).

В сравнение с възможностите за проверка на качеството от системите за компютърноподпомогнат превод специализираните програми разполагат със значително по-голям брой функционалности. Направена е съпоставка на техническите възможности на посочените програми въз основа на техническата им документация, като всяка функционалност е извадена в списък.

Всички програми разполагат с функционалности като правопис, несъответствия в цифрите, пунктуацията и таговете. Всички програми засичат също наличието на празни сегменти в целевия език, липсата на конкретен символ в целевия език и последователното използване на терминология. Други функционалности обаче са присъщи само на конкретна програма, като например възможността за извличане на доклад за извършени промени (само Елтиби разполага с такава). Ексбенч и Еърспай и Кюей дистилър не разпознават граматични грешки; Еърспай не регистрира частичен превод, повторен празен символ, обработка на множество файлове едновременно и не предлага списък за проверка. Кюей Дистилър не открива повтарящи се думи, ЕлТиБе не разполага с опция за силно търсене. Освен Кюей дистилър нито една програма не предлага работа с команден панел, а Ексбенч и Еърспай не предоставят възможност за персонализирани настройки. Всички тези различия трябва да се имат предвид при употребата на програмите за проверка на качеството, за да може да се подбере най-подходящата за целите на всеки проект за превод.

По време на работата със самостоятелните програми за апроверка на качеството и системите за компютърноподпомогнат превод бе забелязано още нещо: всяка от тях степенува откритите грешки по различен начин. Например правописна грешка е много по-сериозен пропуск в сравнение с двоен празен символ. В Традос грешките се делят на: грешка (Error), предупреждение (Warning) и забележка (Note), като към грешка

принадлежат липсващи или сгрешени цифрови изрази, към предупреждение – различна пунктуация в края на изречението, а към забележки – двоен празен символ. В Ексбенч разделението е между основни грешки (Basic) – непреведен сегмент или несъответствие в текста, грешки в текста (Content) – пунктуационни несъответствия, двойни празни символи, пропуснати или сгрешени цифри и други, списък с грешки (Checklists) и проверка на правописа (Spell-checker). Кюей дистилър определя сериозността на грешката според цифрова скала. Например: двоен празен символ би получил Тежест 1 (Severity 1), пунктуационна грешка би получила Тежест 2 (Severity 2), докато пропуск или несъответствие в превода би получил Тежест 5 (Severity 5).

III. Класификация на грешките

Фактът, че всяка програма отдава различна важност на грешките, които открива е причина да се предложи следната **класификация на грешките**, в опит да се анализират грешките и техните реални последици за качеството на превода. класификацията е унифицирана и се основава на възможните негативни ефекти, които дадената грешка може да окаже на качеството на превода.

Скала	Описание	Вид на грешката
1	Оказва слабо или никакво влияние върху качеството на превода; рядко може да доведе до объркване	<u>Несъответствие в пунктуацията:</u> - ограждащи символи (скоби, кавички) – липса на единия знак, липса на двата знака; липса на единия или и на двата знака и наличие на друг знак; наличие на различни знаци около една дума; - различен, липсващ или излишен пунктуационен знак в края на изречението; - излишни или липсващи празни символи (спейсове, табулации); - излишен (липса на) нов ред.

2	Може да доведе до объркване, в зависимост от контекста	<u>Графични несъответствия</u> - несъответствие в цифровите изрази: липсващи, объркани или разместени цифрови изрази (несъответствия в числа, дати, часове, дроби, номерация и др.). При определени ситуации обаче, програмите могат да сметнат за грешка нещо, което е вярно (превръщане на градуси от Целзий към Фаренхайт, локализиране в различни метрични системи и др.); - несъответствие в интернет линкове, имейли и др.; - несъответствие в комбинациите от цифри и букви.
3	Риск за сериозна подмяна на съдържанието на текста	<u>Несъответствия в пълнотата на превода</u> - празни сегменти; - сегментите в целевия език и в езика източник съвпадат. В различните инструменти се разглежда и като текст, който не е преведен; - несъответствие в главните букви в началото на сегмента или изречението; - изпуснати, добавени, разменени букви; - наличие или липса на специфични символи; - наличие на знаци от друга азбука в думата.

4	Неприемлива грешка	<u>Правопис:</u> - повторени думи; - изпуснати думи (могат да бъдат засечени само когато се отнасят до съкращения или думи, намиращи се между кавички, скоби или в комбинация с числа); - несъответствие в съкращения (за да бъдат засечени подобни несъответствия, е необходимо да бъдат въведени предварително в настройките); - липса на преведени части от изречението (подобна липса се отразява най-често, като бъде сравнена дължината на изреченията в %); - непоследователна употреба на конкретни думи и термини в текста. <u>Граматика</u> Само някои инструменти са снабдени с подобна функция, и то за най-разпространените езици. Избраните за експеримента програми не притежават подобна функция и тя няма да бъде тествана.

Таблица 1: Класификация на грешки при превод и тяхното влияние за правилното пренасяне на съдържанието

Грешките варират по важност спрямо последствията при правилното пренасяне и разбиране на съдържанието при превод от един език на друг.

IV. Сравнение на резултатите от автоматичните средства за измерване на качеството

Беше направено сравнение на начина на работа на различни програми и на тяхната успеваемост при откриването на грешки при превода от английски на български. За целта бяха използвани специализираните програми Кюей дистилър и

Ексбенч и функционалностите за проверка на качеството на Традос 2019 и мемоКю 9.2. Бяха използвани текстовете с изкуствено въведени грешки, използвани за първия експеримент.

Грешки със степен на важност 1 се откриват винаги и от всички програми. Понеже става въпрос за липсващ ограждащ символ или пропуснати празни символи, подобни грешки са с минимални последствия за качеството на превода.

Грешки със степен на важност 2 също се откриват винаги (несъответствие между цифри, линкове и комбинации от цифри). Тук е необходимо да се подчертае, че от контекста може да зависи дали са реални, или фиктивни грешки. Към тази група принадлежат и значителна част от намерените „фалшиви грешки“:

Изходен език: *It will be launch on March 17, 2020 at **2** pm Bulgarian time, as the local authorities…*

Целеви език: *Той ще бъде отворен на 17 март 2020 г. в **14.00** ч. българско време като преди това местните власти…*

Пример за „фалшива грешка“ със стойност 3. Програмата е засякла липсата на скоби в целевия език и го сигнализира като пропуск:

*Изходен език: …territory of the Member State**(s)** shall not…*
*Целеви език: …територията на държава/и-членка**/и** не*

При откриването на грешки, характеризиращи се със стойност 4, се забелязват различия в поведението на отделните програми. Ексбенч и Кюей дистилър не отчитат проблеми с правописа, докато Традос и мемоКю се справят значително по-добре. Въпреки това се отразяват като грешки правилни думи: *правоприлагане*, *дихлоробензен*, *ароматизатори*, *договорка*, *киберсигурност* и др.

Безспорното и изключително ценно при превода е предимството на тези програми, когато става въпрос за грешки от тип 4 от скалата – възможността да се открива несъответствие при употребата на конкретни думи, което може да доведе до загуба или объркване при интерпретацията на значението. Става въпрос за последователната употреба на едни и същи думи или термини в целия текст на целевия език.

Поведението на програмите при работа с български език заслужава отделно внимание. Най-общо, не се отчитат някои специфични пунктуационни правила и знаци (по точно, програмите не се справят с правилното отчитане на българските кавички), които не присъстват в други езици. За програмите българските кавички са неправилни поради различието им с тези в изходния език:

Изходен език: *“Almost 90 % of the inspected products do not contain substances…has an obligation to provide it when requested,”*

Целеви език: *„Почти 90 % от проверените продукти не съдържат вещества…има задължението да я предоставя при поискване“*

Безспорно е предимството на програмите за проверка на качеството. За оптимални резултати обаче винаги са необходими персонализирани настройки и човешка преценка. Независимо от факта, че тези програми намират „фалшиви“ грешки, основното им предимство е възможността да открият несъответствия в превода (например, дали в целия текст се използват едни и същи термини).

За да се постигнат най-добрите възможни резултати с тези програми, е задължително да се зададат конкретни настройки за всеки превод. Предлагаме и допълнителна схема за оценяване на инструментите за проверка на качеството (Таблица 2).

Първо ниво	интерфейс и основни характеристики	- лесно ли се използва инструментът - самостоятелен ли е, или е част от програма за компютърноподпомогнат превод - какви файлове може да обработва - кои езици поддържа - позволява ли персонализация
Второ ниво	често използвани функции	- как се справя с общите проблеми (пропуснати знаци, цифри и т.н.) - настройки по подразбиране

		- непоследователност при употребата на терминология - количество на фалшивите грешки
Трето ниво	персонализирани настройки	- как се справя с локализацията - как се справя с настройките за конкретен език в сравнение с друг - персонализирани настройки за конкретен език/клиент – формат на дати, елемент, който не е за превод и др. - разчитане на специални символи - количество на „фалшивите“ грешки при персонализирани настройки
Четвърто ниво	допълнителни предимства	- аудиокоманди - редакция на преводния файл - други предимства (например директни връзки между различни програми)

Таблица 2: Допълнителна схема за оценяване на програмите за проверка на качеството на превода

Програмите за автоматична проверка на качеството (било то вградени в системите за компютърноподпомогнат превод или самостоятелни) са полезни, защото, до известна степен улесняват работата на преводачите. Въпреки това, не трябва да се забравя, че те не могат да заместят хората. Последното се доказва от присъствието на „фалшивите“ грешки.

V. Международни стандарти за качеството на превода

Всички средства за измерване на качеството, последователността, в която трябва да се преглеждат и обработват файловете за превод, уменията и квалификацията на преводачите, комуникацията с клиентите и договорите при какви условия ще се

извършва превода, са подвластни на международните стандарти за качество. Те се приемат с цел унифициране на работните процеси, терминологията, длъжностите и определяне на уменията на преводачите и гарантирането на качеството на превода.

Към днешна дата актуалните международни стандарти при превод на документи са: ISO 9001:2015[64] – управление на качеството на системите; ISO 17100:2015[65] – преводачески услуги; ISO 27001:2013[66] – системи за информационна сигурност; ISO 18587:2017[67] – редакция на машинен превод.

Въпреки че стандартите за превод съществуват от известно време, едва през 2006 година организациите започват официално да работят заедно за стандартизиране на практиките и създаване на официални документи за стандарти, както и на схеми за сертифициране за областта на превода. Крайната им цел е защитата и информираността на потребителите чрез гарантиране на определено ниво на качество при превода. Има няколко елемента, общи за всички стандарти за превод: необходимото или договорено ниво на постигане на качество, необходимостта от ясни договорени спецификации на проекта, изискване за споразумение между клиент и доставчик на преводаческа услуга, процес на управление на проекти, технически капацитет, както и някои други условия, свързани с предоставянето на превод (Бендана и Мелби 2012: 81).

[64] https://www.iso.org/iso-9001-quality-management.html
[65] https://www.iso.org/standard/59149.html
[66] https://www.iso.org/standard/54534.html
[67] https://www.iso.org/standard/62970.html

Заключение

Основната цел на дисертацията е да анализира технологиите и средствата, използвани за създаването на терминологични ресурси в контекста на системите за компютърноподпомогнат превод и да предложи (без претенции за изчерпателност) унифицирана методика за тяхното създаване.

За тази цел бяха анализирани системите за компютърноподпомогнат превод, техническите им характеристики, техните компоненти, както и измененията, които тези инструменти налагат на работните процеси, образованието и обучението на преводачите. Изводът е, че преводната памет, поне към момента, е най-ценният компонент в системите за комптърноподпомогнат превод заради съхранението на данни от предишни преводи. Терминологичната база също има важно значение за преводачите, защото спестява нуждата от търсене на подходящите термини и подпомага избора на термин. Технологията, която изцяло преобръща представите за преводаческата дейност, е машинният превод и неговата употреба ще се увеличава все повече.

Представя се процесът по създаване и управление на терминологичните бази в системите за компютърноподпомогнат превод. Основните начини за създаване и управление на терминологични бази са вградени в системите за компютърноподпомогнат превод. Освен вградените възможности съществуват редица разработени притурки, които могат да се използват както заедно със системите за компютърноподпомогнат превод, както и самостоятелно. Използват се и отделни програми за работа с терминология, които работят независимо от системите за компютърноподпомогнат превод.

Сравнението между най-разпространените програми за проверка на качеството показва, че имат съотносими възможности, но разликата в настройките им може да доведе до различни резултати. Това е поводът да се създаде класификация на грешките при превод, която отчита степента на сериозност на грешките във връзка с последиците за качеството на превода.

Представени са редица проучвания, които изследват уменията и предпочитанията на преводачите в чужбина. Създадена е нова анкета, насочена специално към професионалните преводачи в България. Анализът на резултатите показа, че (с малки изключения) българските преводачи са добре запознати с инструментите за компютърноподпомогнат превод и се стремят да ги използват, макар

че не всички ги намират за полезни и достъпни (подобно е и отношението към преводната памет и терминологичните бази). Преводачите очакват функционалностите на системите за компютърноподпомогнат превод да бъдат все по-интуитивни и по-опростени. Невронният машинен превод създава известни притеснения, че ако продължава да се развива и усъвършенства, може да представлява конкуренция за работата на преводачите.

В дисертацията се предлага методика за полуавтоматично създаване на многоезикова терминологична база от данни в дадена област и за дадени езици. В конкретния случай терминологичната база е двуезикова, езиците са български и английски, като посоката на превода е от английски към български, а областта е компютърната терминология. Първата стъпка при създаването на нова терминологична база е подборът на достоверен източник: паралелен корпус, който представя дадена тематична област достатъчно пълно. Следващите стъпки са: избор и работа с програма за извличане на термини от изходния език (в случая английски), ръчно извличане на преводните еквиваленти на целевия език от подравнените изречения на корпуса, проверка и техническо оформяне на базата.

При анализа на създадената терминологична база в областта на компютърната терминология се показва, че изходният корпус оказва съществено влияние върху качеството и пълнотата на извлечените термини, което, впоследствие може да повлияе на превода. От една страна, доколкото методиката се основава на паралелни езикови корпуси и автоматично извличане на термини, може да се твърди, че се предлага унифициран начин на работа при създаването на терминологични бази. От друга страна, ръчният подбор на преводни еквиваленти осигурява проверка на качеството на подбраните термини, като трябва да се посочи, че това в някои случаи може да отнеме много време.

Като цяло в изследването се предлага описание, анализ и сравнение на съвременните средства, които се използват при превод: програмите за компютърноподпомогнат превод, техните компоненти: преводна памет, терминологична база, машинен превод, инструментите за автоматично извличане на термини и за проверка на качеството на превода. Всичко това е насочено към постигането на основната цел – да се разработи методика за полуавтоматично създаване на терминологични ресурси за дадена специализирана област.

Списък с публикациите, свързани с темата на дисертацията

Translation Quality Assessment Tools and Processes in Relation to CAT Tools. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop* (HiT-IT 2019), pages 89–97, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Computer Terminology in Bulgarian Related with the Human Translation. В: *Сборник с доклади от Международната годишна конференция на Института за български език „Проф. Любомир Андрейчин“* (София, 2020), Том II, стр. 250–257, издателство на БАН „Проф. Марин Дринов“.

Функции и начин на работа на инструментите за проверка на качеството на човешки преводи. *Сборник с доклади от Международната годишна конференция на Института за български език „Проф. Любомир Андрейчин“* (София, 2021), Том II, стр. 225–233, издателство на БАН „Проф. Марин Дринов“.

БИБЛИОГРАФИЯ

Аренас 2020: Arenas, A. G. Pre-editing and post-editing. – In: E. Angelone, M. Ehrensberger-Dow, & G. Massey (Eds.), The Bloomsbury Companion to Language Industry Studies (1 ed., pp. 333–360). (Bloomsbury Companions). Bloomsbury Academic [https://research.rug.nl/en/publications/pre-editing-and-post-editing] (посетен на 20.07.2022).

Бейкър 2017: Baker, W. *Controlled vocabularies in the digital age: are they still relevant?*, Dissertation at University of North Texas, US [https://digital.library.unt.edu/ark:/67531/metadc1011802/m2/1/high_res_d/BAKER-DISSERTATION-2017.pdf] (посетен на 20.07.2022).

Боукър и Сиро 2019: Bowker L. and J.B. Ciro. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*, 37–54.

Булте и Тецкан 2019: Bulté, B. and A. Tezcan. Neural Fuzzy Repair: Integrating Fuzzy Matches into Neural Machine Translation. 10.18653/v1/P19-1175, [https://www.researchgate.net/publication/334745562_Neural_Fuzzy_Repair_Integrating_Fuzzy_Matches_into_Neural_Machine_Translation] (посетен на 20.07.2022).

Гарсия 2015: Garcia, I. Computer-aided translation: systems. – In: In S.-W. Chan (Ed.), *Routledge Encyclopedia of Translation Technology,* University of Western Sydney, Australia, pp. 68–87. Retrieved from http://UWSAU.eblib.com.au/patron/FullRecord.aspx?p=1843560 (посетен на 20.07.2022).

Гаудек 2007: Gouadec, D. Translation as a profession. University of Rennes, Benjamins translation library, 0929-7316 ; v. 73, 2007 [https://benjamins.com/catalog/btl.73] (посетен на 20.07.2022).

Де Калуве и Ван Сантен 2003: De Caluve, J. and A. Van Santen. Phonological, morphological and syntactic specifications in monolingual dictionaries. – In: P. van Sterkenburg (Ed.), *A Practical Guide to Lexicography*, Institute for Dutch Lexicology, Leiden, pp. 71–82. [https://benjamins.com/catalog/tlrp.6.10dec] (посетен на 20.07.2022).

Друан 2003: Drouin, P. Term extraction using non-technical corpora as a point of leverage. – In: *Terminology*. 9. 10.1075/term.9.1.06dro, [https://www.researchgate.net/publication/228683045_Term_extraction_using_non-technical_corpora_as_a_point_of_leverage] (посетен на 20.07.2022).

Дьо Шривер 2003: De Schryver, G.-M. Lexicographers'dreams in the electronic-dictionary age. – In: *International Journal of Lexicography*, 16(2): 143–199. [https://academic.oup.com/ijl/article-abstract/16/2/143/925134] (посетен на 20.07.2022).

Зарецкая и др. 2015: Zaretskaya, A., G. Corpas Pastor and M. Seghiri. Translators' requirements for translation technologies: a user survey. – In: *New Horizons in Translation and Interpreting Studies*, pp. 247–254, Genebra, Switzerland. Tradulex. [https://www.researchgate.net/publication/282658144_Translators'_Requirements_for_Translation_Technologies_a_User_Survey] (посетен на 20.07.2022).

Зарецкая и др. 2015: Zaretskaya, A., G. Corpas Pastor, G and M. Seghiri. Integration of Machine Translation in CAT Tools: State of the Art, Evaluation and User Attitudes. – In: SKASE *Journal of Translation and Interpretation*, vol. 8, no. 1. [https://www.researchgate.net/publication/283667119_Integration_of_Machine_Translation_in_CAT_Tools_State_of_the_Art_Evaluation_and_User_Attitudes] (посетен на 20.07.2022).

Зарецкая и др. 2016: Zaretskaya, A., G. Corpas Pastor, G and M. Seghiri and C. Hernani. Nine Terminology Extraction Tools: Are they useful for translators?. – In: *Multilingual*, 27(3), April/May, pp. 14–20. [https://wlv.openrepository.com/handle/2436/622550] (посетен на 20.07.2022).

Зехналова и др. 2013: Zehnalová, J., O. Molnár and M. Kubánek. A Comprehensive Survey of Multilingual Neural Machine Translation. – In: *Tradition and Trends in Trans-Language Communication*, Olomouc, Univerzita Palackého. [https://arxiv.org/pdf/2001.01115.pdf] (посетен на 20.07.2022).

Имре 2013: Imre, A. Term Bases Reloaded. – In: *Philologia 14*, Studia Universitatis „Petru Maior". pp. 204–210. [https://www.researchgate.net/publication/287911571_Term_Bases_Reloaded] (посетен на 20.07.2022).

Кагеура и Умино 1996: Kageura, K. and B. Umino. Methods of automatic term recognition. – In: *Terminology* vol 3(2), pp. 259–289, John Benjamins Publishing Co. [http://www.iro.umontreal.ca/~felipe/IFT6010-Automne2011/resources/tp3/gabriel_bc/Kageura_Umino_1996.pdf] (посетен на 20.07.2022).

Кирова 2002: Кирова, Л. Еволюция на българската компютърна терминология и компютърен жаргон. – В: *LiterNet*, № 2 (27), [https://liternet.bg/publish3/lkirova/evolution.htm] (посетен на 20.07.2022).

Кирова 2004: Кирова, Л. Компютърната лексика – актуални процеси и тенденции. – В: *LiterNet* № 5 (54), [https://liternet.bg/publish3/lkirova/lex-procesi.htm] (посетен на 20.07.2022).

Кис 2005: Kis, Á. Terminusalkotás: a terminológiai helyzet és a terminológiai szerep. Mindent fordítunk, és mindenki fordít" Értékek teremtése és közvetítése a nyelvészetben, 105–112.

Коева 2007: Koeva, S. Multi-word term extraction for Bulgarian. – In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing Information Extraction and Enabling Technologies* - ACL '07. [https://doi.org/10.3115/1567545.1567556] (посетен на 20.07.2022).

Колковска 2010: Колковска, С. Лексикални неологизми в българския език, възникнали от чужди инициални абревиатури. – В: *Български език*, кн. 4, с. 21–30. [http://ibl.bas.bg/neolex/wp-content/uploads/2013/Issues/Publications/20%20Sia%20Leksikalni%20neologizmi%20ot%20initsialni%20abreviaturi.pdf] (посетен на 20.07.2022).

Коста и др. 2016: Hernani, C., A. Zaretskaya, G. Corpas Pastor and M. Seghiri. Nine Terminology Extraction Tools: Are they useful for translators? – In: *Multilingual,* 27(3), April/May, pp. 14–20. [https://wlv.openrepository.com/handle/2436/622550] (посетен на 20.07.2022).

Лефевер 2009: Lefever, E., L. Macken and V. Hoste. Language-independent bilingual terminology extraction from a multilingual parallel corpus. – In: *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 496–504, Athens, Greece [https://aclanthology.org/E09-1057.pdf] (посетен на 20.07.2022).

Лукашик 2012: Łukasik, M. Terminological dictionary as a comprehensive cognitive and linguistic tool. – In: *Language in Different Contexts*: Research papers = Kalba ir kontekstai, Volume 5 (1), pp. s.98-108. [https://www.researchgate.net/publication/294836193_Terminological_dictionary_as_a_comprehensive_cognitive_and_linguistic_tool] (посетен на 20.07.2022).

Мичъл-Шуитевьордер 2020: Mitchell-Schuitevoerder, R. *A Project-Based Approach to Translation Technology*, London and New York, Publisher: Routledge.

П. Нюмарк 1988: Newman, P. *A textbook of translation. Shanghai foreign language education press*. First published by Prentice HaH International.

Попова 2016: Попова, М. За някои видове терминографски параметри. Институт за български език – БАН, София

[https://ispan.waw.pl/ireteslaw/bitstream/handle/20.500.12528/705/Lexikografia-2016-70-78.pdf?sequence=1&isAllowed=y] (посетен на 20.07.2022).

Попова 2019: Попова, М. Някои въпроси на терминологията във връзка с обучението. Journal: Български език и литература, Issue Year: 61/2019, Issue No: 2, pp. 149–155, [https://www.ceeol.com/search/article-detail?id=758654] (посетен на 20.07.2022).

Райт 2001: Wright, S. E. Data Categories for Terminology Management. – In: Wright S. E. and G. Budin (Eds), *Handbook of Terminology Management: Volume 2: Application-Oriented Terminology Management*, pp. 552–571, John Benjamins Publishing Company.

Сьогард и др. 2015: Søgaard et al. Inverted indexing for cross-lingual NLP. – In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 1713–1722, Beijing, China, July 26-31, Association for Computational Linguistics [https://www.aclweb.org/anthology/P15-1165.pdf] (посетен на 20.07.2022).

Тайбех 2008: Tayebeh, M. Translation Memories and the Translator. – In: *International Journal of Translation*. 20, pp. 97-106. [https://www.researchgate.net/publication/271602313_Translation_Memories_and_the_Translator] (посетен на 20.07.2022).

Тръмбъл и Стивънсън 2002: Trumble, W. R., & Stevenson, A. (Eds.), *Shorter Oxford English Dictionary*. Fifth Edition, North Carolina, publisher: Oxford University Press.

Уорбъртън 2015: Warburton, K. Terminology management. In Sin-wai, C. (Ed), *The Routledge Encyclopedia of Translation Technology,* pp. 644–661, New York, Routledge, [https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.4324/9781315749129&type=googlepdf] (посетен на 20.07.2022).

Фабер и Араус 2021: Faber, P. and Araúz P. Designing Terminology Resources for Environmental Translation. – In: Meng, Ji. and S. Laviosa (Eds.), *The Oxford Handbook of Translation and Social Practices*, pp. 587–615, Oxford Academic [https://www.researchgate.net/publication/350213272_Designing_Terminology_Resources_for_Environmental_Translation] (посетен на 20.07.2022).

Хансен 2010: Hansen, G. Translation ‘errors’. - In: Y. Gamber and Van Doorslaer L. (Eds.) *Handbook of Translation Studies. Volume I*. Amsterdam: John Benjamins, pp. 383–388.
[https://books.google.bg/books?hl=bg&lr=&id=BTwzAAAAQBAJ&oi=fnd&pg=PA385&dq=translation+error+classification&ots=bhrc4Gmjv2&sig=3l5Dpv8GDPsvv1SXo9Z8pp3ioxs

&redir_esc=y#v=onepage&q=translation%20error%20classification&f=false] (посетен на 20.07.2022).

Хаус 2015: House, J. *Translation Quality Assessment. Past and Present*. London, Routledge.

Хьониг 1997: Hönig, H.G. Positions, Power and Practice: Functionalist approaches and translation quality assessment. – In: *Current Issues in Language and Society* 4(1): pp. 6–34. [https://www.tandfonline.com/doi/abs/10.1080/13520529709615477] (посетен на 20.07.2022).

Шаятович и др. 2019: Šajatović, A., M. Buljan, J. Šnajder and B. Bašić. Evaluating Automatic Term Extraction Methods on Individual Documents. – In: *Proceedings of the Joint Workshop on Multiword Expressions and WordNet*, pp. 149–154, Florence, Italy. Association for Computational Linguistics. [https://aclanthology.org/W19-5118/] (посетен на 20.07.2022).