# Тема 15. Машинний переклад

Історія машинного перекладу починається в 1933 році. Радянський вчений Петро Троянський звертається до Академії Наук СРСР з винайденою «машиною для підбору і друкування слів при перекладі з однієї мови на іншу».

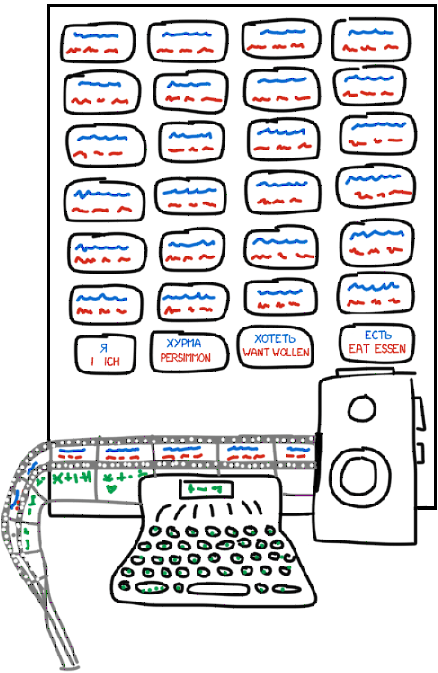


Рис.15.1. Машина Троянського

У проекті машина Троянського представляла собою стіл з похилою поверхнею, перед яким була закріплена фотокамера, поєднана з друкарською машинкою. Клавіатура друкарської машинки складалася зі звичайних клавіш, які дозволяли кодувати морфологічну і граматичну інформацію. Стрічка друкарської машинки і плівка камери повинні були бути пов'язані один з одним і подаватися синхронно. На самій же поверхні столу знаходилося так зване «глосарне поле» - пластина з надрукованими на ній словами, що вільно рухалася (рис.15.1).

Кожне зі слів супроводжувалося перекладами на трьох, чотирьох і більше мовах. Всі слова повинні були бути надані в початковій формі і розташовані на дошці таким чином, щоб найбільш часто використовувані слова були ближче до центру - як літери на клавіатурі. Оператор машини повинен був зрушити глосарне поле і зробити фотознімок слова і його переклади, одночасно набравши на друкарській машинці граматичну та морфологічну інформацію, що відноситься до слова.

У підсумку виходило дві стрічки: одна зі словами відразу на декількох мовах, а друга - з граматичними поясненнями до них, наприклад, «іменник, множина, родовий відмінок». Клавіші друкарки були модифіковані для зручності, кожна однозначно кодувала одну з властивостей. Стрічка друкарської машинки і плівка камери подавалися паралельно, на виході формувався набір кадрів зі словами і їх морфологією (рис.15.2):

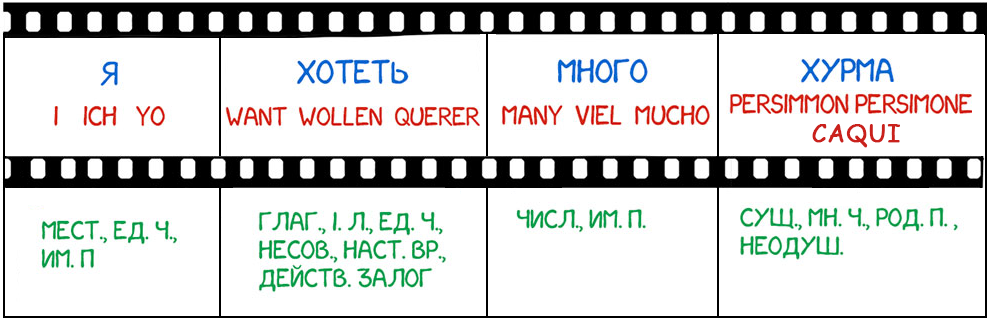


Рис.15.2. Сформована плівка машини Троянського

Коли весь вхідний текст був набраний таким чином, матеріал передавали до носіїв мови - ревізорам, які повинні були звірити дві стрічки і скласти по них тексти на своїх мовах. Далі матеріали повинні були бути передані до редакторів, які знають обидві мови. Їх завданням було довести текст до літературного виду.

Головна ідея винаходу - поділ процесу перекладу на три основних етапи:

1. Людина перетворює слова в реченні в базову форму.
2. Машина перекладає слова на іншу мову.
3. Отриманий результат знов редагує людина.

Перший і останній етапи в сучасній термінології називалися б «pre-editing» і «post-editing». Найвитратнішими за часом процеси (кодування вихідного тексту і синтез з цієї інформації текстів на інших мовах) вимагають від операторів лише знання рідної мови.

Таким чином, переклад здійснювався спочатку між природною мовою та його логічною формою, потім між логічними формами двох мов, а після цього текст в логічній формі вихідної мови вивірявся і приводився до природної форми. Машина Троянського вперше на практиці реалізувала «проміжну мову» (interlingua, інтерлінгва), про створення якої мріяли ще Лейбніц і Декарт.

В СРСР винахід визнали «непотрібним», Троянський помер, 20 років намагаючись її доопрацювати. Ніхто в світі так і не знав про машину, поки його патенти не знайшли в архівах два інших радянських вчених в 1956 році. Сталося це не випадково.

Вони шукали відповідь на виклик холодної війни, оскільки 7 ​​січня 1954 року в штаб-квартирі IBM в Нью-Йорку відбувся Джорджтаунський експеримент. Комп'ютер IBM 701 вперше в світі автоматично перевів 60 речень з російської мови на англійську (рис.15.3). «Дівчина, яка не розуміє ні слова на мові СРСР, набрала російські повідомлення на перфокартах. Машинний мозок зробив їх англійський переклад і видав його на автоматичний принтер з шаленою швидкістю - два з половиною рядки на секунду», - повідомлялося в прес-релізі компанії IBM.



Рис.15.3. Комп'ютер IBM 701 для машинного перекладу тексту

Газети рясніли радісними заголовками, але ніхто не говорив, що приклади для перекладу були ретельно підібрані і протестовані, щоб прибрати будь-яку неоднозначність. Для повсякденного використання ця система підходила не краще від кишенькового розмовника, але гонка озброєнь була запущена. Канада, Німеччина, Франція і особливо Японія теж долучилися до гонки за машинний переклад.

Дослідження, які проходили в 60-х роках в СРСР і США, в основному були зосереджені на мовній парі російська-англійська. В основному предметом перекладу були наукові та технічні документи, наприклад статті з наукових журналів. Недоліки перекладу не заважали сприймати загальний зміст статей. Якщо в статті піднімалися питання, що пов'язані з інтересами безпеки, тоді стаття відправлялася перекладачеві для більш детального перекладу; якщо ні, викидалася за непотрібністю.

Великого удару було нанесено по дослідженням в області машинного перекладу після публікації звіту ALPAC в 1966 році. Звіт було складено урядовою комісією США і представлено Спеціальним комітетом з прикладної лінгвістики (ALPAC) Національної академії наук США. В нього входили сім вчених, зібраних урядом США в 1964 році. Уряд США було стурбований тим, що прогрес був не співмірним зі значними витратами на розробку проекту. В результаті було встановлено, що машинний переклад є дорожчим, повільнішим і менш точним у порівнянні з людським перекладом. Як підсумок визначалося, що найближчим часом комп'ютерний переклад не досягне такої ж якості, що переклад, виконаний людиною.

Звіт рекомендував більше сфокусуватися на розробці словників та розмовників, ніж на машинному перекладі, через що роботи в цьому напрямку практично на десятиліття уповільнюються.

Незважаючи на це, саме спроби вчених, їх дослідження та напрацювання згодом ляжуть в основу сучасного Natural Language Processing. Пошуковики, спам-фільтри, персональні асистенти - все це з'явиться завдяки тому, що кілька країн сорок років поспіль намагалися шпигувати один за одним.

На рис.15.4 наведено розвиток технологій машинного перекладу у часі

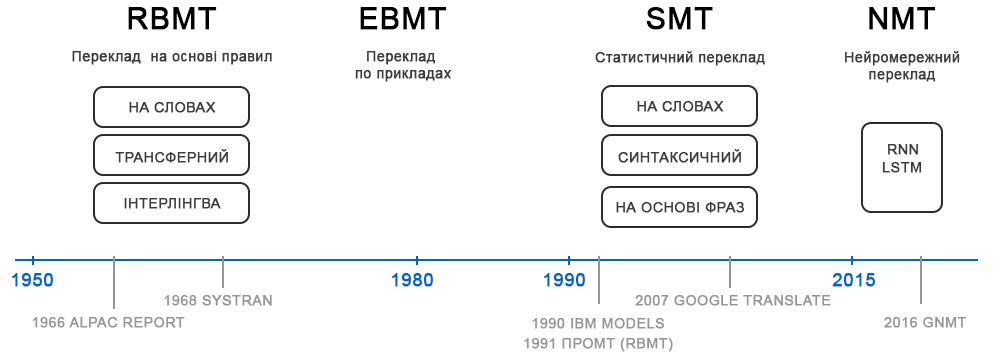


Рис.15.4. Розвиток технологій машинного перекладу

## 15.1. Машинний переклад на основі правил Rule-based Machine Translation (RBMT)

У системах машинного перекладу, створених до початку 1980-х років, використовувався переважно прямий підхід до перекладу. Вчені досліджували роботу лінгвістів-перекладачів і намагалися відтворювати це на комп’ютерах (рис.15.5).



Рис. 15.5. Прямий підхід до машинного перекладу на основі правил

Системи з прямим підходом до перекладу працюють з парами мов, використовуючи:

* Двомовні словники (наприклад, EN -> RU).
* Набір лінгвістичних правил під кожну мову (наприклад, іменники жіночого роду закінчуються на -а/-я).

Вхідний текст перекладається слово в слово, без докладного аналізу синтаксичних структур вхідного тексту і кореляції значень. Після цього в тексті перекладу робляться деякі елементарні поправки відповідно до морфологічних і синтаксичних правил вихідної мови. За бажанням системи доповнювалися модулями типу списків імен, коректорами орфографії і транслітератор.

Це найпримітивніший підхід машинного перекладу, проте, в деяких комерційних системах він досі використовується. Самими яскравими представниками RBMT-систем є PROMT і Systran. «Его отец вербовать других как его устройств…», Аліекспрес досі так перекладає. Ці системи повільно покращувалися, але переклад залишався низької якості.

### 15.1.1. Системи дослівного перекладу Direct Machine Translation

Такий спосіб машинного перекладу є простим та зрозумілим для будь-кого. Текст поділяється за словами, кожне слово перекладається, дещо правиться морфологія, узгоджуються відмінки, закінчення і решта синтаксису (рис.15.6). Спеціально навчені лінгвісти пишуть правила під кожне слово. На виході отримується зазвичай неякісний переклад.

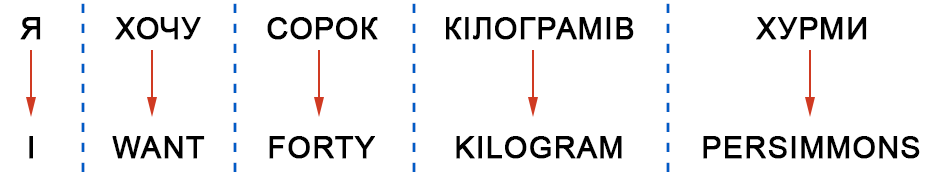


Рис.15.6. Дослівний переклад

В сучасних системах такий підхід не використовується взагалі.

### 15.1.2. Трансферні системи Transfer-based Machine Translation

З часом домінуючим методом машинного перекладу стає непрямий підхід , який має більш складну архітектуру. Системи перекладу аналізують синтаксичну структуру тексту, зазвичай створюючи абстрактне (проміжне) представлення про зміст вхідного тексту і створюючи на його основі текст на вихідній мові. Системи ідентифікують структуру слова (морфологію) і речення (синтаксис) та вирішують проблеми омонімії (семантики).

В цих системах застосовується попередня обробка. Текст розбирається на синтаксичні конструкції - підмет і присудок, шукаються визначення і все інше. Після цього в систему вже не потрібно закладати правила перекладу кожного слова, а відбувається маніпулювання цілими конструкціями. Теоретично, можна домогтися більш-менш непоганої конвертації порядку слів у мовах (рис.15.7).

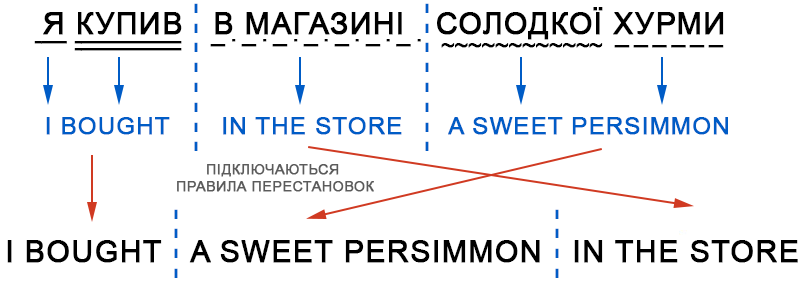


Рис.15.7. Конвертація порядку слів у дослівному перекладі

На практиці переклад виходить фактично дослівним. З одного боку простіше: можна задати загальні правила узгодження за родом і відмінком. З іншого складніше: поєднань слів є значно більше, ніж самих слів. Кожен варіант не врахуєш.

### 15.1.3. Інтерлінгвістичні системи Interlingual Machine Translation

Інтерлінгва – це принцип машинного перекладу, який використовує проміжну (семантичну) модель тексту в якості загального посередника для всіх мовних пар. Інтерлінгва представляє незалежну від конкретної мови модель, що описує не граматику, а зміст тексту. З інтерлінгви може бути згенерований переклад на будь-яку мову.

При машинному перекладі з використанням інтерлінгви процес перекладу складається з двох основних стадій (рис.15.8):

1. Текст мовою оригіналу перекладається в абстрактну модель, що визначає зміст тексту.
2. Абстрактна модель переводиться в текст вихідної мови.

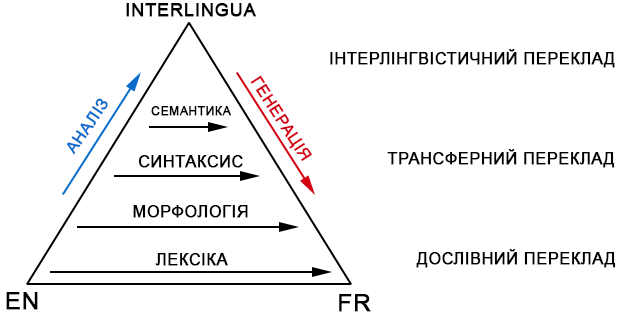


Рис.15.8. Принцип інтерлінгва

Для здійснення принципу інтерлінгви потрібно аналізатор для кожної вхідної мови і генеруюча програма для кожної вихідної мови. Для аналізу вхідного тексту необхідно проведення глибокого семантичного розбору, який передбачає широке знання лексики.

Правила інтерлінгви є єдиними і покривають всі мови світу, тим самим перетворюючи переклад в технічне завдання перегону туди-сюди. Конвертацією інтерлінгви у потрібну мову займаються спеціальні парсери. Найбільш очевидна перевага такого методу полягає в тому, що при перекладах з використанням кількох мовних пар непотрібно створювати перехідний компонент для кожної пари.

Часто інтерлінгву плутають з трансферними системами, адже там теж є конвертація. Однак в трансферних системах правила конвертації пишуться під дві конкретні мови, а в інтерлінгвістичних між кожною мовою і інтерлінгвою. Якщо в інтерлінгвістичну систему додати третю мову, то можна перекладати між всіма трьома мовами, в трансферній системі такого зробити не можна.

У реальному житті все виявилося не так. Створити універсальну інтерлінгву вручну виявилося вкрай складно. Багато вчених займалися цим, але нічого не вийшло. Однак, завдяки багатьом наробкам з'явилися методи морфологічного, синтаксичного і навіть семантичного аналізу. Сама ідея проміжної мови буде затребуваною значно пізніше, через 30 років.

Системи на основі правил RBMT є примітивними, тому, зараз рідко використовуються. Серед плюсів RBMT відзначають морфологічну точність (не плутає слова), відтворюваність (всі перекладачі отримують однаковий результат) і можливість підлаштуватися під предметну область (враховуються спеціальні терміни економістів або програмістів).

Навіть, якщо б вченим вдалося створити ідеальну RBMT, а лінгвістам закласти в неї всі правила правопису, в будь якій мові присутні виключення. Неправильні дієслова в англійській, плаваючі приставки в німецькій, суфікси в російській, зрештою ситуації, коли «в нас так не говорять, треба ось так». Спроба врахувати всі нюанси обертається мільйонами витрачених даремно годин праці.

Множина правил не вирішує головної проблеми - омонімії. Одне і те ж слово може мати різне значення залежно від контексту, а значить відрізняється і його переклад.

Мови розвивалися не завжди на основі граматик і правил, про які люблять розмірковувати лінгвісти. Вони більше залежать від того, який вплив на мову справили загарбники та імперії. За сорок років холодної війни вчені так і не змогли знайти виразного рішення RBMT, тому цей напрямок на сьогодні вважається не перспективним.

## 15.2. Машинний переклад на прикладах Example-based Machine Translation (EBMT)

У 80-і роки в Японії був великий ажіотаж, пов'язаний з машинним перекладом. Там не було холодної війни, але були свої причини: мало хто в країні знав англійську. Це обіцяло великі труднощі зважаючи на прогнозовану глобалізацію, через що японці були вкрай мотивовані знайти робочий метод машинного перекладу.

З заявою щодо створення комп'ютерів п'ятого покоління Японія планувала стрибнути вище всіх в області техніки і програмування. Тому, проект, що пов'язаний зі створенням програм для перекладу з/на англійську, зацікавив багато відомих компаній.

Англо-японський переклад на основі одних тільки правил є вкрай складним, будова мов відрізняється, майже всі слова доводиться переставляти і додавати нові. У 1984 році вченому Макото Нагао з університету Кіото приходить ідея не намагатися кожного разу перекладати заново, а використовувати вже готові фрази.

Наприклад, потрібно перекласти речення «я йду в магазин». Десь в сховищах вже є переклад схожої фрази «я йду в театр» і словник з перекладом слова «магазин». Отже, можна обчислити різницю і перекласти лише одне слово в наявному прикладі, не дивлячись на інші конструкції. І чим більше є прикладів - тим краще буде переклад (рис.15.9).

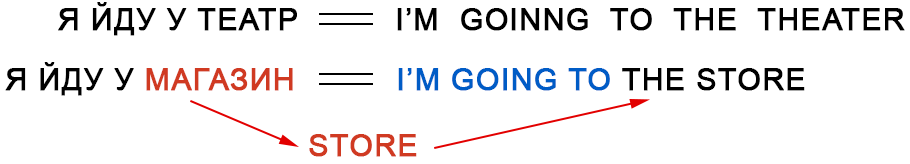


Рис.15.9. Принцип перекладу на прикладах

Історична важливість методу була в тому, що вчені всього світу вперше прозріли: можна не витрачати роки на створення правил і винятків, а просто взяти багато наявних перекладів і надати їх машині. Це був важливий крок до революції, до винаходу статистичного перекладу залишалося п'ять років.

## 15.3. Статистичний машинний переклад Statistical Machine Translation (SMT)

На рубежі 1990 року в дослідницькому центрі IBM вперше показали систему машинного перекладу, яка нічого не знала про правила і лінгвістику. Вчені надали комп'ютеру дуже багато однакових текстів на двох мовах, і змусили його самому розбиратися в закономірностях (рис.15.10).



Рис.15.10. Паралельний корпус текстів

Ідея була простою і одночасно красивою: береться одне речення на двох мовах, розбивається за словами і кожне слово зіставляється з його перекладом. Операція повторюється мільйони разів. Машина підраховує скільки разів слово *das Haus* перекладалося як *house, building, construction* тощо. Якщо найчастіше це було *house*, його і будуть використовувати (рис.15.11). В цьому випадку не було задано ані правил ані словників. Машина сама все знайшла, керуючись чистою статистикою і логікою «люди переводять ось так, значить і я буду». Так народився статистичний переклад.

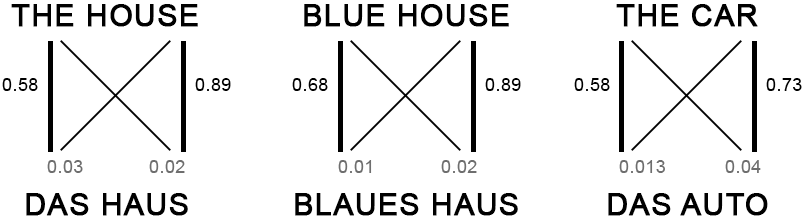


Рис.15.11. Визначення відповідності слів перекладу за статистикою

Точність таких перекладачів виявилася помітно вищою за всі попередні, а розробка не вимагала участі лінгвістів. Більше текстів - покращується переклад.

Одна проблема: як машина здогадається, що для *das Haus* парою є саме *house*, а не будь-яке інше слово з речення? Порядок слів є різним, звідки можна дізнатися як саме треба розділити і знайти потрібні слова? Відповідь: ніяк.

На початку роботи машина з рівною часткою ймовірності вважає, що слово *das Haus* перекладається як кожне зі слів наявного речення. Коли вона зустрічає *das Haus* в інших реченнях, то кількість перекладів *das Haus* як *house* починає збільшуватися з кожним разом. Це називається «алгоритмом вирівнювання слів» (word aligment) і є типовою задачею для машинного навчання.

Машині потрібні мільйони і мільйони речень на двох мовах, щоб набрати статистику за кожним словом. І де стільки взяти? Наприклад, в Європарламенті і раді ООН ведуть конспекти засідань на мовах всіх країн-членів, їх і беруть. Зараз ці документи є у відкритому доступі: [UN Corpora](https://catalog.ldc.upenn.edu/LDC2013T06) і [Europarl Corpora](http://www.statmt.org/europarl/).

Деталі статистичного перекладу в Google Translate. Google не тільки показує ймовірності, але і рахує зворотну статистику (рис.15.12).

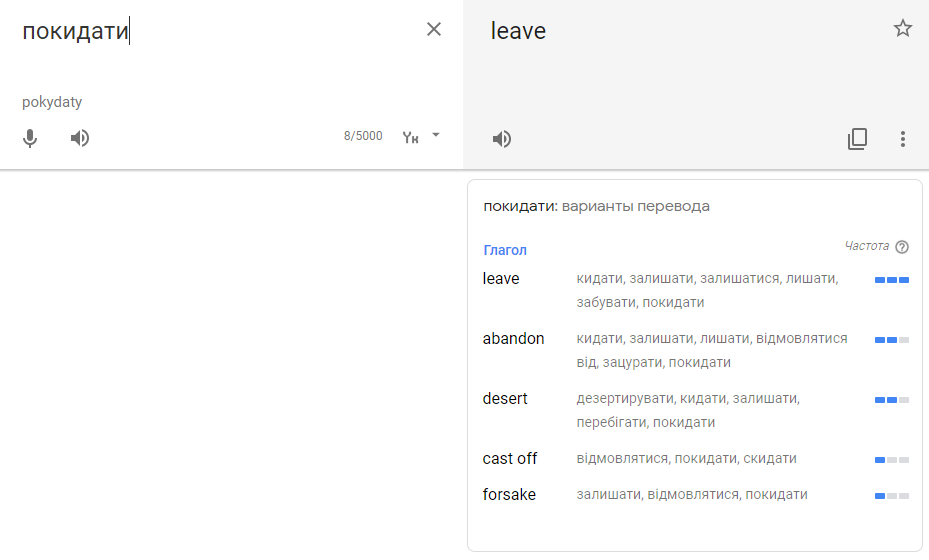


Рис.15.12. Статистика вживання слів перекладу в Google Translate

### 15.3.1. Статистичний переклад за словами Word-based SMT

Перші системи статистичного перекладу розпочинали з розподілу за словами. Це здавалося логічно і просто. Першу винайдену модель статистичного перекладу в IBM назвали IBM Model1.

#### Model 1: множина слів

Класичний підхід – речення поділяються на слова і підраховується статистика. Жодного обліку порядку або перестановок. З властивостей, Model 1 вміла хіба що перекладати одне слово в кілька. *Der Staubsauger* (порохотяг) легко перекладався в *Vacuum Cleaner*, але назад переклад був непередбачуваним (рис.15.13).

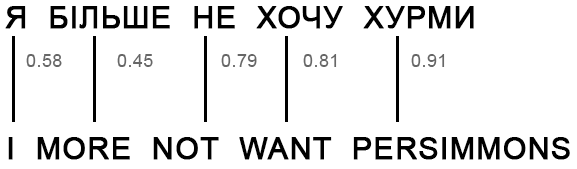


Рис.15.13. Модель статистичного перекладу Model 1

#### Model 2: облік порядку слів у реченні

Відсутність знань про порядок слів у мовах стало проблемою для Model 1, оскільки в деяких випадках порядок є доволі важливим. Тому, в Model 2 стали запам'ятовувати на якому місці з'являється слово перекладеного речення. Додали проміжний крок - після перекладу машина намагалася переставити слова місцями так, що це виглядало більш природно (рис.15.14).

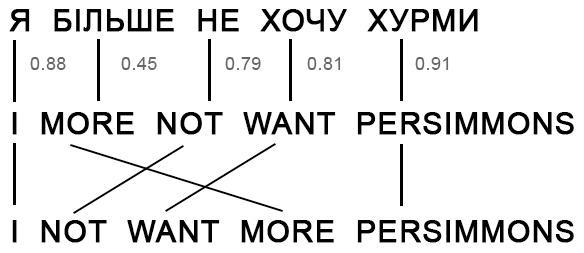


Рис.15.14. Модель статистичного перекладу Model 2

Стало краще, але все ще не достатньо.

#### Model 3: додавання відсутніх слів

Часто при перекладі з'являються нові слова, яких не було в оригінальному тексті. В німецькій мові раптово вилазять артиклі, в англійському часто вставляють дієслово *do*. «*Я не хочу хурми*» → «*I do not want persimmons*». Щоб вирішити цю проблему в Model 3 додано два проміжних кроки (рис.15.15):

1. Уставляння маркерів (NULL-слів) на ті місця, де машина передбачає необхідність нового слова.
2. Підбір потрібного артикля, частки або дієслова під кожен маркер.

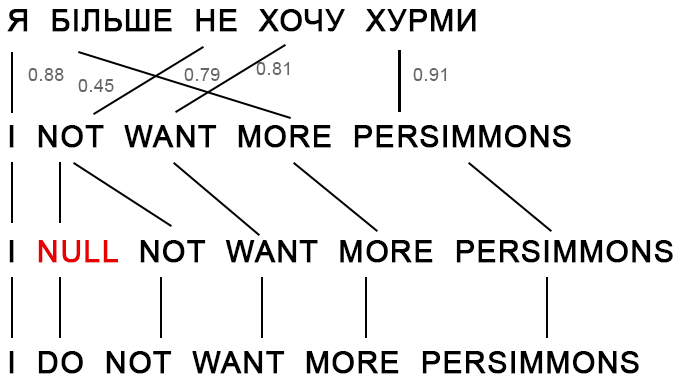


Рис.15.15. Модель статистичного перекладу Model 3

#### Model 4: перестановки слів

Model 2 хоча й враховувала порядок слів у реченні, але нічого не знала про перестановки слів між собою. Часто при перекладі потрібно, наприклад, поміняти іменник і прикметник місцями. Тому, в Model 4 стали враховувати ще й так званий «відносний порядок». Якщо при перекладі два слова постійно змінювалися один з одним - модель це запам'ятовувала.

#### Model 5: виправлення помилок

В Model 5 додали параметри для навчання і виправили проблеми, коли два слова конфліктували за місце в реченні.

Незважаючи на всю революційність, Word-based системи як і раніше погано справлялися з відмінками, родом і омонімією. Кожне слово вони перекладали єдиним, на їхню думку, вірним способом. Зараз такі системи не використовуються, їх замінив більш просунутий метод - переклад по фразах.

### 15.3.2. Статистичний переклад по фразах Phrase-based SMT

Метод взяв за основу все принципи перекладу за словами: статистика, перестановки і лексичні прийоми. Але для навчання текст розділявся не лише на слова, а й на цілі фрази. Точніше N-грами або фраземи - набори з N слів поспіль, що перетинаються (рис.15.16). Машина вчилася перекладати стійкі поєднання слів, що помітно покращило точність.



Рис.15.16. Розділення тексту на фраземи

Особливість методу полягала в тому, що «фрази» не завжди були зрозумілими синтаксичними конструкціями. Як тільки в переклад намагалася втручатися людина, яка знає про лінгвістику і будову речень, якість перекладу різко падала. Піонер комп'ютерної лінгвістики Фредерік Йелинек одного разу пожартував з цього приводу: «Кожен раз, коли з команди йде лінгвіст, якість розпізнавання зростає».

Крім покращення точності, переклад по фразах надав більше свободи в пошуку двохмовних текстів для навчання. Для Word-based перекладу була дуже важливою точна відповідність перекладів, що виключало будь-які літературні твори і вільні переклади. Phrase-based прекрасно навчався навіть на них. Розробники навіть почали парсити новинні сайти на різних мовах і покращували переклад цими текстами (рис.15.17).

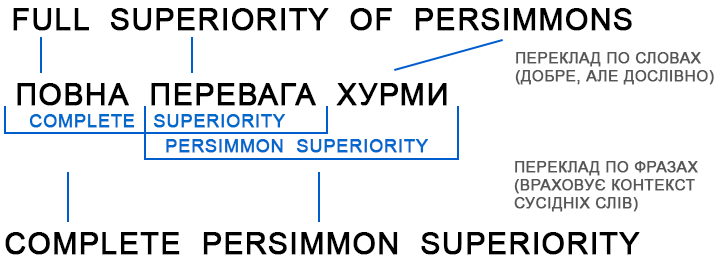


Рис.15.17. Статистичний переклад по фразах

З 2006 року цей підхід почали використовувати всі: Google Translate, Yandex, Bing та інші якісні онлайн-перекладачі, вони працювали саме як Phrase-based до 2016 року. Але траплялися випадки, коли одне речення Google перекладав на відмінно, літературно переставляючи слова, а на іншому починав формувати повну нісенітницю. Така особливість перекладу по фразам.

Якщо Rule-based підхід стабільно надавав передбачуваний, хоч і жахливий результат, то статистичні методи іноді дивували. Наприклад, Google Translate міг перекласти «three hundred» як «300», вважаючи це за вірне. Такі випадки назвали статистичними аномаліями.

Phrase-based переклад став настільки популярним, що під терміном «статистичний машинний переклад», скоріше за все розуміють його. До 2016 року у всіх дослідженнях Phrase-based переклад схвально називають [the State of Art](http://www.aclweb.org/anthology/D16-1161). Тоді ніхто навіть не підозрював, що в лабораторіях Google вже працюють з нейромережами, щоб знову змінити ставлення до машинного перекладу.

### 15.3.3. Статистичний переклад на основі синтаксису Syntax-based SMT

До приходу нейромереж, про синтаксичний переклад багато років говорили як про «майбутнє перекладачів», але досягти успіху він так і не встиг.

Прихильники синтаксичного перекладу вірили в об'єднання підходів SMT і старого трансферного перекладу за правилами. Потрібно навчитися робити досить точний синтаксичний розбір речень - визначати підмет і присудок, залежні члени тощо, а потім побудувати дерево. Маючи таке дерево, можна навчити машину правильно конвертувати фігури однієї мови в фігури іншої, виконуючи решту перекладу за словами або фразами (рис.15.18). Тільки робити це тепер машинним навчанням. В теорії це вирішило б проблему порядку слів назавжди.

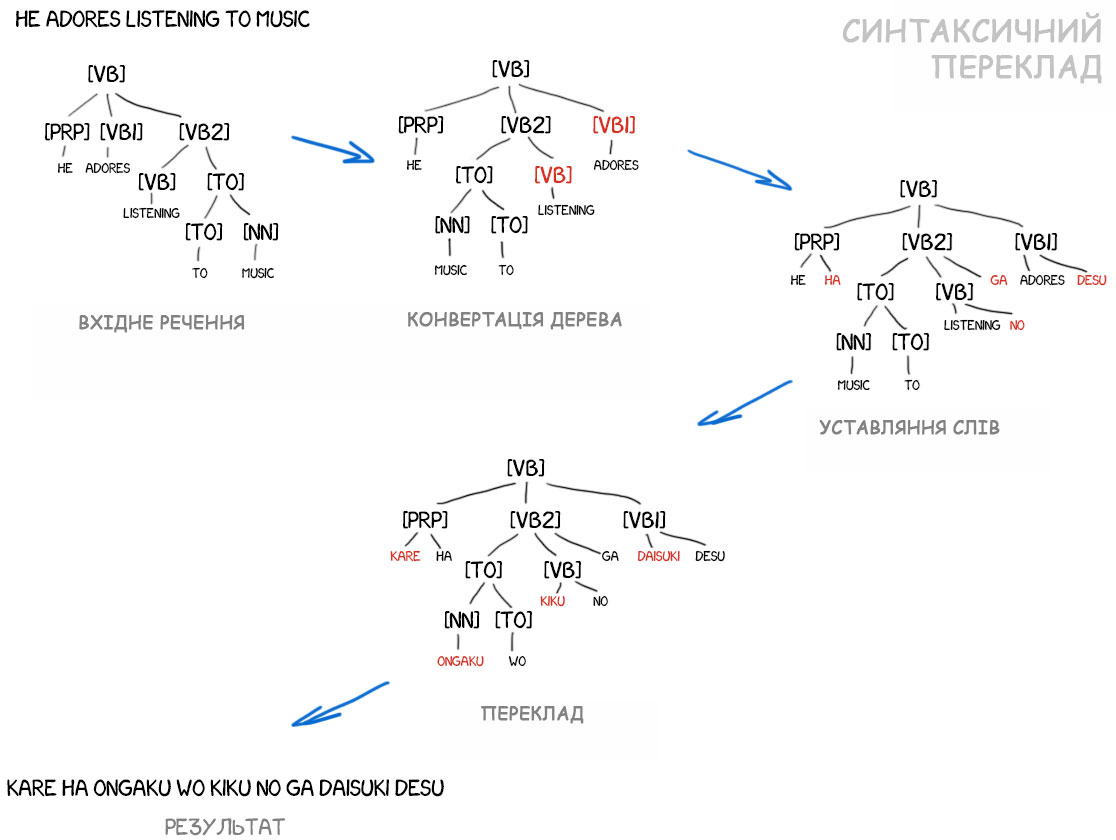


Рис.15.18. Статистичний переклад на основі синтаксису

Проблема в тому, що хоча людство і вважає проблему синтаксичного розбору давно вирішеною (для багатьох мов є готові бібліотеки), за фактом він працює досить погано.

## 15.4. Нейронний машинний переклад Neural Machine Translation (NMT)

У 2014 році виходить стаття з коротким описом ідеї застосування нейромереж глибокого навчання до машинного перекладу. В Інтернеті її взагалі ніхто не помітив, а в лабораторіях Google почали активно працювати.

У вересні 2016 року Google Brain опублікувала статтю про свою систему нейронного машинного перекладу (Neural Machine Translation, NMT). Йшлося про такі системи ще в нульових. На відміну від попередніх систем, NMT алгоритм розглядає речення не як сукупність слів і фраз, а як цілісну одиницю тексту. В результаті кількість помилок при онлайн перекладі різко скоротилася, а точність підвищилася.

У липні 2017 року Google представила нову архітектуру нейромереж - Transformer. Такі нейромережі ефективніше зчитували зв'язки між словами, що стояли в реченні далеко один від одного. Наприклад, у реченні «I arrived at the bank after crossing the…» переклад слова bank залежить від того, що знаходиться в кінці фрази:

I arrived at the bank after crossing the road («Я перетнув дорогу і опинився у банку»).

I arrived at the bank after crossing the river («Я перетнув річку і опинився на березі»).

Такі речення коректно перекладалися і раніше. Але нейромережі архітектури Transformer аналізують контекст ефективніше.

Анонс архітектури Transformer перегорнув наступну сторінку машинного перекладу.

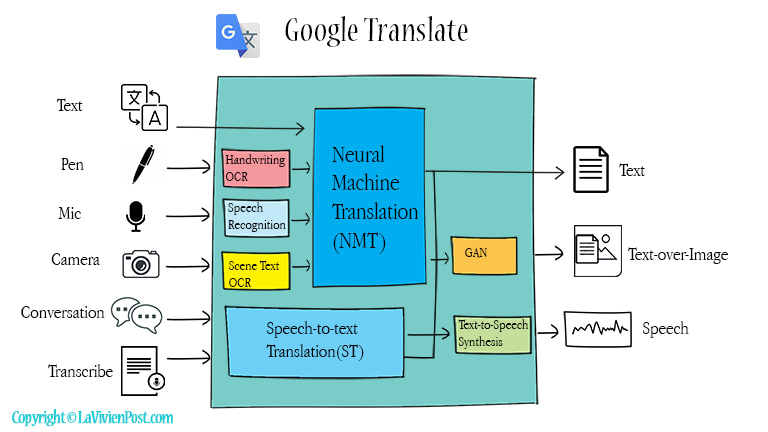
Neural Machine Translation (NMT) – це підхід до автоматизованого перекладу текстів з однієї мови на іншу з використанням штучних нейронних мереж.

* В основі лежать глибокі нейронні мережі, навчені перекладати текст. Це дозволяє досягти вищої якості, порівняно з традиційними методами.
* Використовуються архітектури sequence-to-sequence із механізмами уваги. Вони дозволяють моделі враховувати контекст вихідної пропозиції.
* Модель навчається на великих паралельних корпусах – текстах та їх перекладах – для витагування перекладацьких патернів.
* Завдяки глибоким нейромережам, модель може обробляти все речення повністю, а не за окремими фразами.
* NMT дозволяє досягти природнішого, людського перекладу, зберігаючи сенс.

Технологія активно застосовується в онлайн-перекладачах від Google, Facebook, Microsoft та інших. Завдяки нейронним мережам, машинний переклад виходить на новий рівень якості, наближаючись до можливостей людини.

## Google Translation

Google Translate – це програма для перекладу з однієї мови на іншу в реальному часі.



Перекладач Google працює на основі найсучасніших технологій ШІ з нейронними мережами.

* Переклад тексту в текст
* Розпізнавання рукописного тексту та переклад
* Розпізнавання тексту, машинний переклад і синтез зображень
* Переклад мовлення в текст і синтез мовлення
* Транскрибація мови в текст.

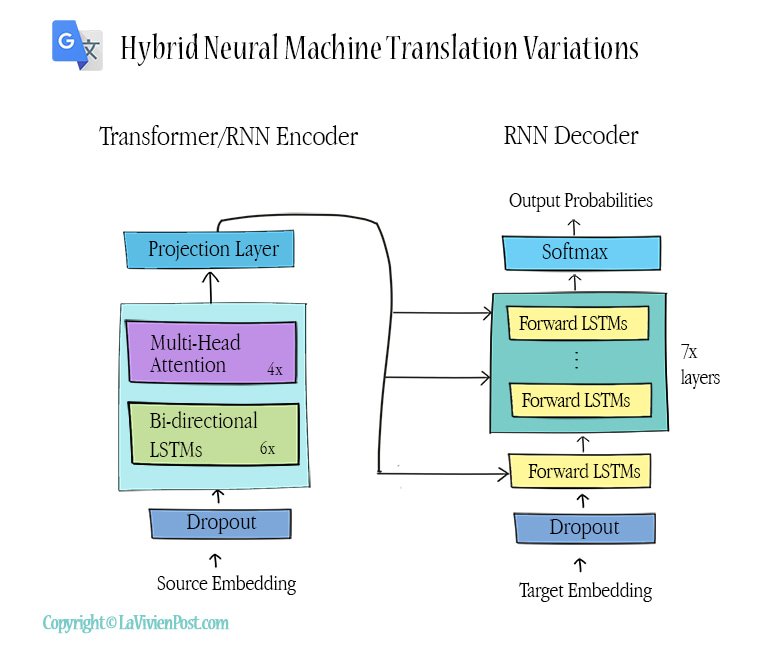
Це є складовими частинами Google Translate. Деякі компоненти працюють разом як каскадна система. Деякі каскадні системи перебувають у процесі заміни наскрізними моделями.

### Хронологія технологій Google Translate.

* 2006 рік. Google запустив Google Translate, який використовував статистичний машинний переклад.
* 2011 рік. Google Brain започаткував нейронні мережі в технологіях обробки природної мови.
* 2015 рік. Google запустив розпізнавання рукописного тексту.
* 2016 рік. Представлено Neural Machine Translation від Google.
* 2016 рік. Застосування WaveNet – згорткова нейронна мережа, яка, після навчання на великому обсязі даних, може робити висновки про нові дані.
* 2017 рік. Модель трансформера з механізмом уваги для заміни RNN і CNN.
* 2018 рік. Гібридні моделі Transformer і RNN в GNMT
* 2019 рік. Translatotron - перша модель прямого перекладу мови в мову.
* 2020 рік. Переклад із зображення, моделі прямого перекладу із зображення на зображення.
* 2022 рік. Translatotron 2 – покращена модель перекладу прямої мови в мову.

### Google Neural Machine Translation (GNMT)

Google Translation використовує нейромережну модель власної розробки Google Neural Machine Translation (GNMT). GNMT вперше представлена у 2016 році і з того часу вона постійно вдосконалюється. GNMT використовується в різних продуктах Google, включаючи Google Translate, Google Assistant та Google Search.



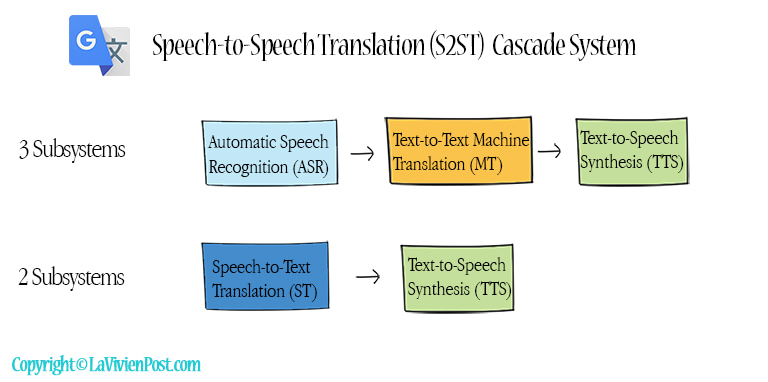
Модель розроблено на основі глибоких нейронних мереж та архітектур енкодер-декодер.

* Навчається на величезних обсягах паралельних текстів різними мовами з відкритих джерел.
* Підтримує переклад між більш ніж 100 мовами.
* Використовує механізми уваги для врахування контексту під час перекладу.
* Реалізована на TPU (Tensor Processing Units) – спеціалізованих чіпах Google для прискорення нейромереж.
* Постійно вдосконалюється із використанням машинного навчання.
* Доступна на сайті translate.google.com та в інших сервісах Google.

GNMT – це високотехнологічна нейронна модель перекладу, розроблена інженерами Google для свого популярного сервісу Google Перекладач. GNMT забезпечує надійний високоякісний переклад тексту в текст. Він також є посередником в перекладі мовлення та перекладі зображень.

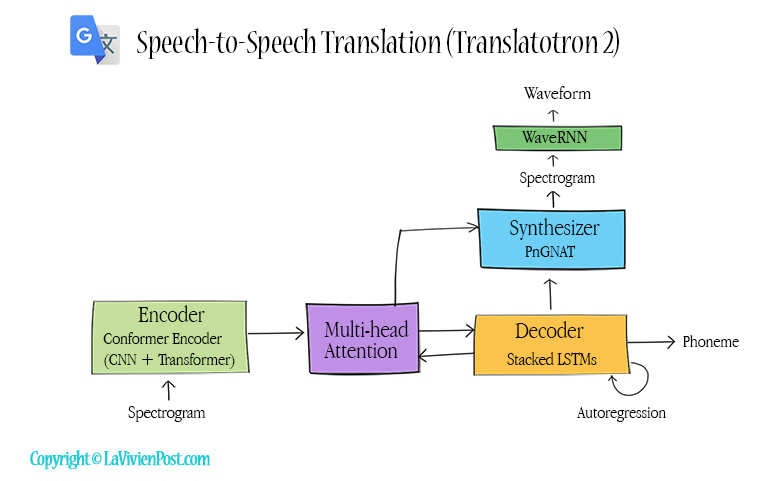
### Синхронний переклад (S2ST)

У минулому для перекладу розмов використовувалася каскадна система з трьох компонентів: автоматичного розпізнавання мовлення (ASR), машинного перекладу тексту в текст (MT) і синтезу тексту в мовлення (TTS). ASR і машинний переклад були об’єднані в єдину наскрізну систему перекладу мови в текст (ST).



У 2019 році Google представив перший прямий переклад мовлення під назвою Translatotron. Translatotron міг перекладати мову з однієї мови на мову іншою без проміжного представлення тексту. Ця система також змогла зберегти голос мовця джерела.

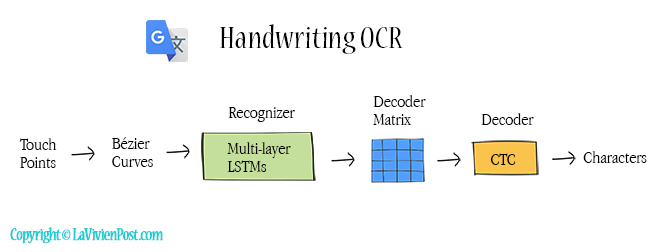
У 2022 році Google витіснив Translatotron2. Він перевершив Translatotron і наблизився до каскадної системи.



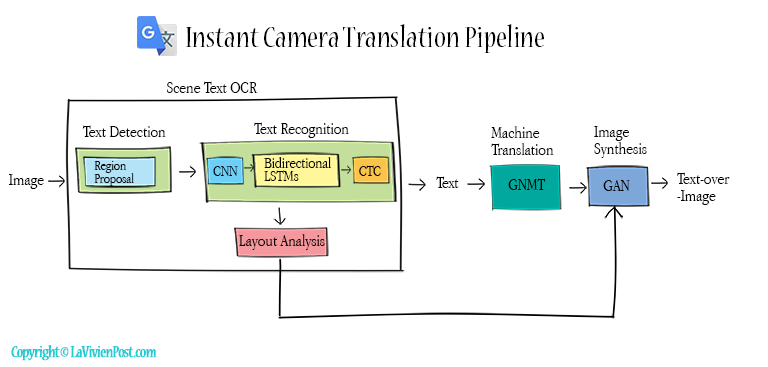
Синтезатор виконує синтез мовлення за допомогою системи Non-Attentive Tacotron (NAT) на основі тривалості. Синтезатор залежить від виходу як з декодера, так і від уваги. Після того, як синтезатор генерує спектрограми, нейронний вокодер, такий як WaveNet або WaveRNN, перетворює спектрограми на сигнали у часовій області.

### Оптичне розпізнавання символів (OCR)

Google Translate забезпечує рукописне введення та миттєвий переклад з камери. Технологією обох є оптичне розпізнавання символів ( OCR). OCR — це набір технологій для автоматичного виявлення та розпізнавання текстів на папері чи іншому носії та перетворення їх у цифровий формат, який комп’ютер може редагувати, шукати та зберігати.



Миттєвий переклад з камери дозволяє наводити камеру на документи чи публічні вивіски та перекладати їх. Конвеєр включає OCR, машинний переклад і синтез зображень. Виявлення тексту використовує мережу речень регіону (конволюційна модель) для виділення тексту на зображенні та розміщує рамки навколо тексту. Тексти ізольовані від фонових об'єктів. Потім розпізнавання тексту використовує мережі на основі NLP (мережі CNN і LSTM) і CTC для перетворення витягнутої інформації для побудови осмислених речень.



Коли OCR ідентифікує символи, воно також визначає метадані, такі як розмір, шрифт, колір і ракурс камери. Цей процес називається аналізом компонування.

Розпізнані тексти завантажуються в GNMT для перекладу. У подальшому перекладені тексти відображаються поверх оригінального тексту в тому самому стилі, що відповідає оригінальному шрифту, розміру та довжині. Він також використовує генеративні змагальні мережі (GAN) для перебудови пікселів під ними, щоб зробити переклад виглядати природним. GAN — це тип моделі глибокого навчання, який використовується для створення синтетичних зразків даних, подібних до навчального набору даних.

## Facebook AI Translate

Модель машинного перекладу Facebook AI Translate може перекладати текст між будь-якою заданою парою мов зі 100 мов. Перекладач Facebook працює без використання англійської мови як посередника та може досягти приблизно 90% точності.

Перекладач заснований на архітектурі Transformer, яка дозволяє моделювати довгострокові залежності між словами та фразами у тексті. У Facebook Translate застосовуються методи навчання з підкріпленням для покращення якості перекладу. Це дозволяє моделі машинного перекладу взаємодіяти з навколишнім середовищем та отримувати зворотний зв'язок на основі якості перекладу, що допомагає їй покращувати свої результати.

Навчальні дані Facebook для моделі ШІ складалися з приблизно 7.5 мільярдів пар речень, розподілених між 100 різними мовами. Дані були зібрані з Інтернету за допомогою ряду веб-сканерів, а мови, присутні в зібраних даних, були ідентифіковані за допомогою мовної моделі під назвою FastText. Після збору даних за допомогою інструменту LASER 2.0 витягували значення різних зразків речень і зіставляти речення різними мовами на основі їх значення. LASER 2.0 розроблено Facebook і використовує алгоритми неконтрольованого навчання створення вставок. Вбудовані речення містять інформацію про зв’язки між різними реченнями на основі таких характеристик, як частота використання та те, наскільки речення здаються одне до одного. Після цього LASER 2.0 може створювати пари речень, які мають дуже подібні значення.

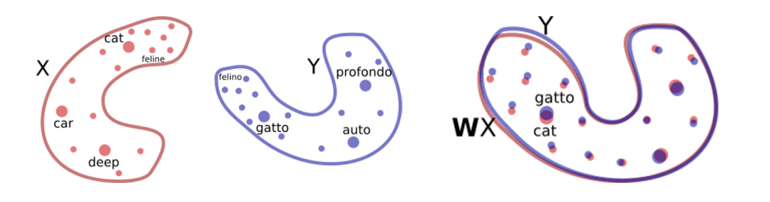


Рис.15.27. Значення різних зразків речень

Навчальні дані не просто об’єднувалися на основі значень речень. Самі мови були згруповані разом. Мета полягала в тому, щоб розробити систему, яка не вимагала б використання англійської як посередника між двома мовами. Інженери Facebook провели навчання, зосередившись на поєднанні мов, які зазвичай перекладаються один на одного та з них. Було створено чотирнадцять різних мовних груп на основі таких змінних, як культура, лінгвістична подібність і географія. Наприклад, одна з лінгвістичних груп, створених дослідниками, містила найпоширеніші мови по всій Індії, які включають мови урду, тамільську, гінді та бенгальську. Це було зроблено для того, щоб загальноприйняті мови отримували якісний переклад.

Метод навчання, орієнтований на мовну групу, дає цікаві результати. Виявлено, що отримана модель перекладу має більшу точність, ніж існуючі моделі для певних мовних пар. Наприклад, під час перекладу між англійською та білоруською мовами штучний інтелект зміг застосувати певні шаблони, які він вивчив під час перекладу російською мовою, оскільки білоруська мова має мовну схожість із російською. Подібним чином покращилася робота з перекладу між іспанською та португальською мовами, оскільки іспанська мова є другою за поширеністю мовою, і для цього завдання було набрано значний обсяг навчальних даних.

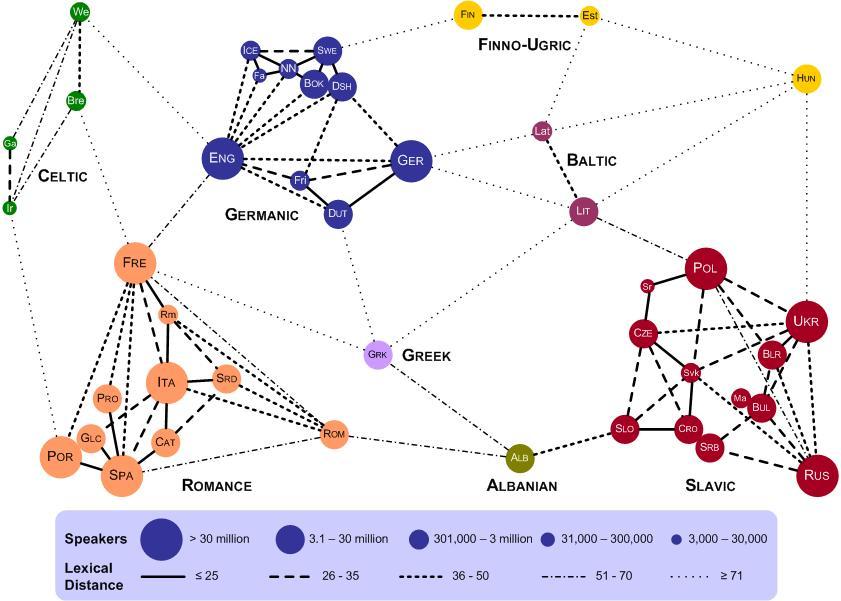


Рис.15.29. Лексична відстань серед мов Європи

Існує приблизно шістдесят мов, які система перекладу ще не охоплює, і точність моделі для мов без великої кількості навчальних даних потрібно покращити, перш ніж вона буде готова до використання. Багатьом мовам у Південно-Східній Азії та Африці бракує даних, необхідних для підготовки надійної моделі. Дослідницькій групі потрібно буде визначити якийсь спосіб компенсації цієї нестачі даних. Дослідницькій групі також потрібно визначити, як контролювати будь-які расистські, сексистські чи інші профанні моделі, які могла засвоїти модель. Хоча дослідницька група використовувала фільтр ненормативної лексики, фільтр працює переважно на даних англійською мовою.

Важливо відзначити, що Facebook Translate продовжує розвиватися та покращуватися з часом, і Facebook постійно працює над розширенням та вдосконаленням своєї системи машинного перекладу.

## SeamlessM4T

Meta, материнська компанія соціальної медіа-платформи Facebook, представила нову модель перекладу на основі штучного інтелекту SeamlessM4T. Компанія рекламувала SeamlessM4T як «першу все-в-одному багатомовну мультимодальну модель перекладу та транскрипції ШІ». На відміну від традиційних моделей перекладу, SeamlessM4T є мультимодальним, тобто він може розуміти та перекладати як усну мову, так і текст, пропонуючи високу універсальність.

SeamlessM4T може виконувати широкий спектр завдань перекладу, включаючи переклади з мовлення в текст, з мовлення в мовлення, з тексту в мовлення та з тексту в текст, обслуговуючи 100 мов залежно від конкретного завдання.

Цей комплексний підхід усуває потребу в окремих моделях, що призводить до зменшення кількості помилок і затримок, що підвищує ефективність і якість перекладів. Він має потенціал кардинально змінити спосіб спілкування людей, долаючи мовні бар’єри.

Meta оприлюднила SeamlessM4T за дослідницькою ліцензією, запрошуючи дослідників і розробників використати його потужність і зробити внесок у його еволюцію. Також, Meta випустила метадані SeamlessAlign, що містить приголомшливі 270 000 годин ретельно видобутого мовлення та вирівнювання тексту. Цей набір даних є цінним ресурсом для спільноти дослідників ШІ, допомагаючи в розробці майбутніх мовних моделей.

SeamlessM4T є частиною ширшої місії Meta зі створення універсального перекладача. За останні роки Meta випустила кілька проектів AI, пов’язаних із мовами, зокрема No Language Left Behind (NLLB), модель машинного перекладу тексту в текст, що підтримує 200 мов, а також системи перекладу мовлення.

No Language Left Behind (NLLB) — це інноваційний проект в галузі штучного інтелекту, який надає відкритий доступ до моделей, здатних забезпечувати високоякісний переклад 200 мовами, включаючи низькоресурсні мови (астурійська, луганда, урду та ін.). Мета проекту – дати людям можливість отримувати та поширювати веб-контент рідною мовою, а також спілкуватися де і з ким завгодно незалежно від мовних уподобань.

SeamlessM4T може похвалитися унікальною можливістю перемикання кодів. Це означає, що він може легко розпізнавати та перекладати кілька мов, коли вони використовуються як взаємозамінні в одному реченні, що є важливою функцією для користувачів, які регулярно перемикаються між мовами.

## Microsoft Translator

У сервісі машинного перекладу Bing Translator Microsoft використовується нейромережна модель під назвою Microsoft Neural Machine Translation (MNMT).

Основні характеристики цієї моделі:

* Заснована на глибоких нейронних мережах та механізмах уваги (аналогічно Transformer).
* Навчена на величезних паралельних корпусах Організації Об'єднаних Націй та Євросоюзу для якісного перекладу.
* Підтримує переклад між більш ніж 100 мовами по 6000 мовних пар.
* Дозволяє перекладати не лише текст, а й документи, веб-сторінки, а також голосову та відео-мову.
* Реалізовано на графічних процесорах для високої швидкості перекладу.
* Постійно покращується на основі відгуків користувачів та нових даних.
* Доступна через веб-інтерфейс Bing Translator та API для використання у програмах.

## Baidu Translate

Baidu Translate, розроблений китайським технологічним гігантом Baidu, є онлайн-службою перекладу, яка дозволяє користувачам перекладати текст і документи між різними мовами. Він використовує передові технології штучного інтелекту і машинного навчання для надання точних і ефективних послуг перекладу.

Сервіс пропонує можливості перекладу для більш ніж 100 мов, включаючи такі широко поширені мови, як англійська, іспанська, французька, китайська, японська та німецька, а також менш поширені мови, такі як суахілі, ісландська та бенгальська.

Однією з ключових особливостей Baidu Translate є його здатність виконувати переклад у реальному часі. Це означає, що користувачі можуть вводити текст або мовлення, а система забезпечить миттєвий переклад. Наприклад, мандрівники можуть використовувати мобільну програму Baidu Translate для захоплення тексту з вивісок, меню або документів за допомогою камери свого смартфона, і програма швидко перекладе текст на потрібну мову.

Baidu Translate підтримує переклад мови в мову, що робить його потужним інструментом для вербального спілкування. Користувачі можуть говорити в мікрофон свого пристрою, а Baidu Translate перетворить їхнє мовлення на потрібну мову в режимі реального часу.

Baidu Translate пропонує кілька інших розширених функцій, які сприяють його загальній функціональності. Наприклад, сервіс надає мовний словник і посібник з вимови, що дозволяє користувачам вивчати нові мови або вдосконалювати свої мовні навички. Словник містить визначення, синоніми та приклади речень, а посібник із вимови допомагає користувачам правильно вимовляти слова та фрази. Цей комплексний аспект вивчення мови виділяє Baidu Translate серед звичайних служб перекладу, позиціонуючи його як цінний освітній ресурс.

Ефективність і точність Baidu Translate постійно вдосконалюються завдяки відгукам і внескам користувачів. Користувачі можуть залишати відгуки про переклади, допомагаючи Baidu удосконалювати та вдосконалювати свої алгоритми. Baidu також активно співпрацює з перекладацькою спільнотою, співпрацюючи з лінгвістами та експертами для подальшого покращення якості своїх перекладів. Використовуючи колективні знання та досвід своїх користувачів і мовних професіоналів, Baidu Translate прагне до постійного вдосконалення та гарантує, що він залишається на передньому краї ландшафту технологій перекладу.

Baidu Translate пропонує ряд додаткових функцій, які покращують його функціональність і зручність використання. Однією з важливих особливостей є режим офлайн-перекладу, який дозволяє користувачам завантажувати мовні пакети та використовувати Baidu Translate без підключення до Інтернету.

Baidu Translate пропонує API (інтерфейс прикладного програмування), який дозволяє компаніям інтегрувати служби перекладу у власні програми, веб-сайти або продукти. Цей API дозволяє компаніям надавати своїм клієнтам безперебійні можливості перекладу, дозволяючи їм взаємодіяти з глобальною аудиторією та розширювати охоплення.

Останніми роками Baidu Translate також почав працювати в галузі синхронного перекладу. Синхронний переклад передбачає переклад усної мови в режимі реального часу під час конференцій, зустрічей або заходів. Система Baidu Translate на базі штучного інтелекту може аналізувати та інтерпретувати мову однією мовою та створювати відповідні переклади в режимі реального часу, дозволяючи учасникам розуміти та спілкуватися один з одним.

## Контрольні питання

1. Які принципи покладено у машинний переклад на основі правил Rule-based Machine Translation (RBMT)?
2. Перелічіть проблеми, що виникають при прямому дослівному перекладі.
3. Яка країна вперше застосувала машинний переклад на прикладах Example-based Machine Translation (EBMT)?
4. Назвіть ідею, яку було взято за основу при розробленні статистичного машинного перекладу Statistical Machine Translation (SMT).
5. Перелічіть моделі, якими доповнювали можливості статистичного машинного перекладу.
6. Які властивості людини відтворюються при нейронному машинному перекладі Neural Machine Translation (NMT)?
7. Перелічіть типи нейронних мереж, що використовуються для машинного перекладу.
8. Які властивості Google Translate сприяють його популярності?
9. Який підхід для автоматичного перекладу застосовано у Facebook перекладачі?
10. Перелічіть особливості, які застосовують для синхронного машинного перекладу.

## Використані джерела

### Історія розвитку машинного перекладу <https://vas3k.ru/blog/machine_translation/>

1. Основні дати машинного перекладу <https://www.langust.ru/etc/history.shtml>
2. Розвиток машинного перекладу <https://magazine.skyeng.ru/machine-translation-evolution/>
3. Історія і досвід використання машинного перекладу <https://habr.com/ru/company/yandex/blog/427271/>
4. Машинне навчання це весело! Частина 5: Переклад мов з допомогою глибокого навчання <https://medium.com/@ppleskov/машинное-обучение-это-весело-часть-5-51f40fbae4b0>
5. LSTM – нейромережі довгої короткотермінової пам’яті <https://habr.com/ru/company/wunderfund/blog/331310/>
6. Google долає бар’єр між людським та машинним перекладом <https://www.reg.ru/blog/google-preodolevaet-barer-mezhdu-chelovecheskim-i-mashinnym-perevodom/>
7. Візуалізація нейронного машинного перекладу <https://habr.com/ru/post/486158/>