# СТАТИСТИЧЕСКИ МАШИНЕН ПРЕВОД: ПРОБЛЕМИ И ПОДХОДИ

**ПРЕСЛАВ И. НАКОВ**

## Увод

Интересът към машинния превод – първата и една от най-важните задачи на изкуствения интелект – датира от 40-те години на миналия век и възниква още с появата на първите компютри. Това е периодът непосредствено след края на Втората световна война, по време на която в САЩ, във връзка с необходимостта от декодиране на съдържанието на прихванатите немски съобщения, силно развитие получават теорията на информацията и криптографията. Създаденият в резултат мощен математическо-статистически апарат съвсем естествено се разглежда като средство за постигане на автоматичен превод. През 1949 г. Уорън Уейвър от фондация Рокфелер пише [*Weaver,1955*]: „*Пред мен стои текст, написан на руски, но аз ще си мисля, че всъщност е на английски, но е кодиран с някакви странни символи. Всичко, което трябва да направя, за да разчета закодираната информация, е да го декодирам.*" Макар привидно наивно, това разсъждение се оказва изключително полезно и днес лежи в основата на съвременния статистически машинен превод.

Идеята за автоматичен превод бързо буди ентусиазъм. През 1954 г. в САЩ се създава първият руско-английски прототип, което подсилва очакванията за бърз напредък, включително и за други двойки езици. Изследванията в България също не закъсняват и през 1964 г. се създава специална група за машинен превод между руски и български език под ръководството на проф. Александър Людсканов в Института по математика на БАН. Междувременно, във връзка със Студената война и започналото съревнование между СССР и САЩ в овладяването на Космоса, обемът на превежданата научно-техническа документация от руски на английски език нараства значително, а изследванията в областта на автоматичния превод се радват на богато финансиране. През 1966 г. обаче настъпва драматичен обрат: по поръчка на правителството на САЩ, Американската академия на науките изготвя доклад за състоянието на изследванията в областта на компютърната лингвистика и на машинния превод в частност, който се оказва силно скептичен [*Hutchins,2003*]. В резултат настъпва дълъг оздравителен период със силно ограничено финансиране първо в САЩ, а после и в световен мащаб. Едва през 1975-1985 г. започва постепенно възраждане първо в Европа, Япония и СССР, а след 1985 г. – и в САЩ.

Нещата рязко се променят през 90-те години на миналия век, когато компютрите набират достатъчно изчислителна мощ, за да бъде проправен пътят на *статистическия* подход [*Brown&al.,1993*], който е доминиращ и до днес. Благодарение на последвалия значителен теоретичен и практически напредък, както и поради видимото подобрение в качеството, днес машинният превод отново се радва на значителен изследователски интерес, отчасти мотивиран и от големи икономически очаквания – така например, само годишните разходи за преводи на Европейската комисия възлизат на над един милиард евро.

## Защо машинният превод е труден

Постигането на качествен автоматичен машинен превод е една от най-важните, но и най-трудни, задачи на изкуствения интелект (ИИ). Счита се, че е „ИИ-пълна" (по аналогия с класа на NP-пълните задачи в теорията на алгоритмите) в смисъл, че изисква решаване на всички фундаментални задачи на изкуствения интелект. От гледна точка на компютърната лингвистика качественият машинен превод предполага правилен автоматичен анализ и генериране на естествен език на морфологично, синтактично, семантично и прагматично ниво с отчитане на контекста, използване на знание за света и познаване на културата на носителите на езика. Изисква още генериране на текст, разрешаване на различни видове многозначност,

разпознаване на собствени имена, транслитерация, анализ на местоимения, и изобщо дълбоко разбиране на смисъла на превеждания текст. Наред с това, много от предизвикателствата са чисто лингвистични: машинният превод, както и преводът изобщо, се усложняват заради обективните разлики между езиците.[1] Без да имаме претенция за изчерпателност, по-долу ще посочим някои от тези разлики.

**Словоред:** Основният словоред в повечето европейски езици е подлог-сказуемо-допълнение (ПСД), например англ. "*I like beer.*" ('*Аз обичам бира.*'). В други езици като турски, японски и хинди той е подлог-допълнение-сказуемо (ПДС), например тур. „*Ben bira seviyorum*" (буквално „*Аз бира обичам*."). В ирландски пък основният словоред е сказуемо-подлог-допълнение (СПД). Някои езици с падежи, напр. чешки, имат относително свободен словоред. Такъв е и българският – макар че няма падежи, той позволява всичките 6 варианта (основният словоред е ПСД, но са възможни и алтернативи с различна информационна структура): *„Аз обичам бира."*, *„Аз бира обичам."*, *„Бира аз обичам.", „Бира обичам аз.", „Обичам аз бира.", „Обичам бира аз."* В немски език словоредът може да бъде ПСД (напр. „*Ich mag Bier.*") или ДПС (напр. „*Bier mag ich*."), като спрегнатият глагол трябва задължително да бъде на втора синтактична позиция (при сложни глаголни времена инфинитивите се струпват в края на изречението), а подлогът и прякото допълнение могат да си разменят свободно местата и обикновено заемат позициите непосредствено преди/след глагола. Словоредът може да се различава по много други начини. Например, в турски език предлогът следва, а не предшества думата, за която се отнася, а относителните изречения предшестват глагола, вместо да го следват. На френски език въпрос, на който се отговаря с „да/не", най-често се формулира чрез промяна на словореда (напр. „*Aimez-vous la bière?*", т.е. буквално "*Обичате-вие бирата*?"), на български може да се използва въпросителна частица, поставена на подходящо място, например „*Вие обичате ли бира?*" и „*Вие бира обичате ли?*" (възможни са и алтернативи с друг фокус, напр. „*Вие ли обичате бира?*", „*Вие обичате бира ли?*", „*Вие бира ли обичате?*"), а на английски обикновено се използва спомагателен глагол ("*Do you like beer?*"). В турски също може да се използва въпросителна частица, която обаче се вмъква между основата и окончанието на глагола „*Bira seviyor musunuz?*" („*Обичате ли бира?*") Сравнете със съобщителното изречение „*Bira seviyorsunuz."* („*Вие обичте бира.*") В български език прилагателните обикновено предшестват съществителните, за които се отнасят, напр. *бяла къща*, докато в романските езици най-често ги следват, напр. фр. *maison blanche*. Друг основен източник на разлики в словореда е местоположението на отрицанието (виж по-долу), на наречията (сравнете „*Само веднъж*." на нем. „*Nur einmal*." и англ. „*Once only*") и на обстоятелствените пояснения. Например, в случай на няколко обстоятелствени пояснения на немски, те трябва да бъдат подредени в следния ред: (1) за време, (2) за начин и (3) за място. Сложните комбинации от такива разлики могат силно да затруднят получаването на правилния словоред при автоматичен превод.

**Отрицание:** Друг източник на разлики е отрицанието. В повечето европейски езици то се изразява с помощта на специална отрицателна частица, най-често непосредствено предшестваща глагола, например бълг. „*Не искам бира*.", исп. „*No quiero cerveza*.", итал. „*Non voglio birra*." В немски позицията на отрицанието обикновено е в края на изречението, например „*Ich will Bier nicht*.", освен в случай на сложно глаголно време, когато *nicht* се поставя непосредствено преди глаголните форми в края на изречението: „*Ich bin nicht spazieren gegangen*." ('*Не съм ходил да се разхождам*.'). В турски отрицанието се вмъква между основата и окончанието на глагола: „*Bira istemiyorum*.". В английски пък обикновено се изисква използването на спомагателния глагол *do*: „*I do not want beer*." Във френски отрицанието се състои от две частици, ограждащи спрегнатия глагол, напр. „*Je ne veux pas de bière*.", като *pas* може да се заменя с други отрицателни частици, напр. „*Je ne veux plus de bière*." ('*Не искам повече бира.*'). В руски език отрицанието обикновено изисква родителен падеж: сравнете „*Есть человек, есть проблема*." и „*Нет человека, нет проблемы*." По подобен начин, на френски в положително изречение се използва частичен член „*Je veux de la bière*.", докато в

[1] Понякога точният превод е дори невъзможен, както се вижда и от началото на заглавието книгата на *Bond* [*2005*]: „*Да преведеш непреводимото...*".

отрицателно – само *de*: „*Je ne veux pas de bière*.”. Отрицанието може да има друга форма, когато фокусът му не е глаголът, а допълнението. Например, „*Нямам пари*.” на немски се изразява като „*Ich habe kein Geld*.”, а на английски като „*I have no money.*” (макар че е възможно и „*I do not have money*.”). На български „*Нямам пари*.” използва специална форма за отрицанието на *имам – нямам*, вместо *„*не имам*”. Специална форма (*yok*) има и в турски: сравнете „*Benim param var*.” (буквално „*Мои пари има*.”, т.е. ‘*Имам пари*.’) с отрицателното “*Benim param yok*.” (букв. „*Мои пари няма*.”, т.е. ‘*Нямам пари.*’). В руски такива изречения са безглаголни и използват *нет* вместо *не*: „***У меня*** *нет денег*.” Подобен начин на изказ има и в турски: „***Bende*** *para yok*.” (буквално „*У мен пари няма*.”, т.е. ‘*Нямам пари в себе си*.’). Освен това, поради липсата на глагол *съм* в турски има безглаголни изречения, които се отричат с помощта на специалната форма *değil*: “*Ayşe evde değil.*” (буквално „*Айше вкъщи не (е)*.”, т.е. ‘*Айше не е вкъщи*.’). Накрая, макар че в български език двойното отрицание е нормално (напр. „*Не искам никаква бира*.”), на английски то не е допустимо (напр. “*I do not want* ***any*** *beer*.”, букв. „*Аз не искам* ***някаква*** *бира*.”, т.е. ‘*Не искам никаква бира*.’).

**Род:** В някои езици, напр. български и немски, съществуват три граматични рода: мъжки, женски и среден. В съвременните романски езици обаче липсва среден род, а в някои езици, напр. турски и малайски, изобщо няма граматичен род. Английският език е интересен в това отношение: той прави разлика между мъжки и женски род в личните местоимения за хора и животни (*he*, *she*), но не и за неодушевени предмети (*it*). В езици като български, руски, немски и френски прилагателните се съгласуват по род със съществителните; в английски език обаче не се изменят. В езици като български, руски и немски род има само в единствено число. Във френски обаче има две различни местоимения за трето лице множествено число: *ils* (‘*те*’, м.р.) и *elles* (‘*те*’, ж.р.). Така е и в испански, където има и различни местоимения за първо лице множествено число: *nosotros* (‘*ние*’, м.р.) и *nosotras* (‘*ние*’, ж.р.). Такива родови несъответствия между езиците създават редица проблеми, особено при превод от език с по-бедна система от граматични родове към такъв с по-богата. Но дори когато броят на родовете и начинът им на употреба съвпадат, остава проблемът с несъответствията между родовете на отделни думи в различните езици, напр. *бира* е в женски род в български, но *das Bier* е от среден род в немски език.

**Число:** Различните езици изразяват граматичното число по различен начин, което е източник на проблеми при превод. В съвременния български език, подобно на повечето европейски езици, има две граматични числа: единствено (напр. *вълк*) и множествено (напр. *вълци*, *вълка*). В старобългарски обаче е имало и двойнствено число – за чифт предмети[2], напр. за очи, уши, ръце, крака, крила, обувки и др. Днес това двойнствено число е напълно запазено само в словенски език (измежду славянските езици)[3], където има двойнствени форми както за прилагателните и съществителните, така и за глаголите, напр. „*En volk hodi*.” (ед. ч., ‘*Един вълк ходи.’*), „*Dva volkova hodita*.” (дв. ч., „*Два вълка ходят*.”), „*Trije volkovi hodijo*.” (мн. ч. ‘*Три вълка ходят.’*). Забележете, че в български се употребяват различни форми на множественото число, в зависимост от това дали е зададен конкретен брой, напр. *един вълк*, *два/три/пет вълка*, но по принцип – *вълци*, също *много вълци*. В словенски, руски и други славянски езици формата за множествено число при пет и повече предмета е различна от тази за два, три и четири, напр. на словенски се казва *pet volkov* (‘*пет вълка*’), а на руски – *пять волков* (сравнете с *два/три/четыре волка*). Забележете още, че на български се казва *много вълци* (небройна форма), а на руски – *много волков* (бройна форма). В турски език пък вместо бройна форма се използва формата за единствено число, напр. *bir*/*iki/üç/dört kurt* (‘*един*/*два/три/четири вълка*’), а граматичното множествено число, напр. *kurtlar* (‘*вълци*’), се използва, само когато не се подразбира от контекста. За сравнение, нещата са доста по-прос-

[2] Употребата на двойнственото число всъщност е по-сложна. Така например, макар да е ясно, че имаме две ръце, на словенски се казва „*Umij si roke!*” („*Измий си ръцете!*”), което е в множествено число. Сравнете с „*Umij si obe roki!*” („*Измий си и двете ръце!*”), което е в двойнствено число.

[3] В български език все още са запазени някои форми на двойнственото число, които могат да се употребят в случай на двойка предмети. Например, можем да кажем *криле* (т.е., „*две крила”*), но не и **три криле*, докато нормалната форма за множествено число *крила* може да се употреби както за два, така и за повече предмети. Други такива двойнствени форми са *обуща* (вместо *обувки*) и *нозе* (вместо *крака*).

ти в английски език, където има само една форма за множествено число (напр. *wolf* ‘*вълк*’ става *wolves*), която се употребява, когато става дума за повече от един предмет, независимо дали е специфициран конкретен брой или количество. Накрая, макар в европейските езици множественото число да се изразява най-често морфологично (т.е. чрез промяна на окончанието), в други езици това може да става с различни механизми, напр. в малайски – с повторение на думата, напр. *serigala* (‘*вълк*’) става *serigala-serigala* (‘*вълци*’).

**Падеж:** Много езици използват падежи за маркиране на граматичната функция на думите в изречението, което обуславя някои свободи в словореда. Така например на немски можем да кажем както „*Das Parlament verabschiedet den Antrag*.” (‘*Парламентът утвърждава предложението*.’), така и „*Den Antrag verabschiedet das Parlament*.”, без да има двусмислици, защото и в двата случая подлогът е в именителен падеж, а прякото допълнение – във винителен. Ако не разпознаем правилно падежа при втория словоред, можем да направим погрешен превод на английски като “*The request approves the Parliament*.” (‘*Предложението утвърждава парламента*.’) Сравнете с правилните преводи на български „*Предложението (го) утвърждава парламентът*.” или „*Парламентът утвърждава предложението*.”. Виждаме, че правилното членуване с пълен/кратък член за мъжки род в български също изисква разпознаване на подлога и допълнението. Поради отсъствието на падежи обаче, лесно можем да стигнем до двусмислици като следните вестникарски заглавия: „*Дете ухапа куче*”, *„Баба блъсна влак”*, „*7 месечно бебе застреля мъж и се самоуби*”, *„Вълк-великан застреля ловец в Родопите*”, „*Разсеяни шофьори избиват американците*” и др. Падежите често се обуславят от конкретни глаголи и могат да се използват без предлози в случаи, когато на български се изисква предлог, напр. „*Работя като учител*.” на руски се превежда с творителен падеж: „*Работаю учителем*.”. Така при превод към език с падежи следва да разпознаем граматичната функция на думите, за да можем да ги преведем правилно. В български, за разлика от всички останали славянски езици, няма падежи (освен при местоименията, напр. *той-го,него,му*, и при въпросителни думи в мъжки род, напр. „*Кого видя? Не видях никого*.”). Запазен е обаче звателният[4] (напр. *Българийо*), който съществува и в други славянски езици (напр. в чешки и сръбски), но липсва в руски. За славянските езици са характерни творителен и предложен падеж, които липсват в немски. В други езици падежната система може да бъде много по-различна. Така например, в турски притежание се изразява, като притежателят се маркира с родителен падеж (както в руски, немски и английски), а притежаваният предмет – с притежателен падеж (без аналог в останалите европейски езици). Например, *книгата на моя приятел* на турски е *arkadaş**ım**ın kitab**ı***, т.е. „*приятел-мой*-род.падеж *книга*-прит.падеж”. (Сравнете с английското *my friend’s book*, където има родителен падеж за притежание *'s*, но няма притежателен падеж.) Това означава, че при превод от български на турски трябва да се разпознава, кога един израз е притежателен (напр. *книгата на моя стол* не е такъв), както и кой е притежателят и кой е притежаваният предмет.

**(Не)определителен член:** Различните правила и начини на членуване са друг източник на проблеми при автоматичен превод. В български, както и в някои други балкански езици като румънски и турски, определителният член е задпоставен и се пише слято с думата, за която се отнася, например бълг. *човекът*/*човека* (в мъжки род има различни форми за подлог и за допълнение), тур. *adamı* (използва се само за членуване на пряко допълнение във винителен падеж), рум. *omul*. Така е и в скандинавските езици, например швед. *mannen*. В повечето европейски езици обаче определителният член е предпоставен и е отделна част на речта, например англ. *the man*, фр. *l’homme*, нем. *der Mann*. В немски определителният член може да се изменя освен по род и число и по падеж, напр. *den Mann* (винителен), *dem Mann* (дателен), *des Mannes* (родителен). В славянските езици, с изкл. на българския, няма определителен член. Същевременно, в някои български диалекти (македонски и родопски) определителният член има три форми в зависимост от разстоянието между говорещия и предмета, за който става дума, напр. *човеков* (близо: ‘*този човек тук*’), *човекот* (междинно разстояние),

[4] В съвременния български език има тенденция звателният падеж да се избягва: макар че все още е частично запазен за мъжки род (напр. „*Иване, ела тук!”*), употребата му за женски род обикновено се счита за обидна.

*човекон* (далече: '*онзи човек там*')[5]. Накрая, в български[6], руски и останалите славянски езици няма неопределителен член, какъвто има в повечето европейски езици, напр. ит. *un иото*, порт. *um homem*, исп. *un hombre*, англ. *a man*, нем. *ein Mann*. Правилата за употреба на членните форми се различават между различните езици, например на български казваме *президентът Буш*, но на английски няма членуване: *President Bush* (както при *др. Сталин*). Повечето езици не позволяват членуване на лични имена, но португалският го допуска, например „*Vi o José.*", т.е. „*Видях Жозе.*". На български казваме „*Обичате ли бира?*", но на френски членуването е задължително: „*Aimez-vous la bière?*" На английски не се членува, когато се правят общи твърдения и обобщения, но на френски това е задължително и т.н.

**Лични местоимения:** Различните езици разделят света по различен начин от гледна точка на личните местоимения. В английски език, например, е невъзможно да се направи разлика между *ти* (за неформално единствено число), *вие* (за множествено число) и *Вие* (форма на учтивост). На всички тях съответства *you*[7], което представлява голяма трудност пред автоматичен превод от английски на български език: например „*What do you think?*" може да се преведе като „*Какво мислиш/те ти/вие/Вие?*" При превод на италиански биха съществували четири възможности, защото там има отделни форми на учтивото *Вие* за единствено и за множествено число: *Lei* и *Loro*. Във френски език пък съществуват две различни форми, съответстващи на българското *те*: *ils* за мъжки род и *elles* за женски род. Така е и в испански, където освен това има отделни форми на *ние* за мъжки и за женски род: *nosotros* и *nosotras*. В малайски също има различни форми на *ние*, но в зависимост от това дали се включва слушателят (*kita*) или – не (*kami*). В малайски има и две форми на *аз*: *aku* (неформално), и *saya* (формално). От друга страна, в малайски (както и в турски) няма граматичен род и има само едно местоимение за трето лице единствено число (*dia*), което създава проблеми при превод на български език, където трябва да се избере измежду *той*, *тя* и *то*. В малайски *dia* може да означава и *те*, ако от контекста е ясно, че се има предвид множествено число (макар че има специална форма: *mereka*). По подобен начин, в немски *sie* може да означава както *тя*, така и *те*. В Латинска Америка пък в испански и португалски са отпаднали местоименията и съответните глаголни форми за второ лице множествено число (съответно *vosotros* и *vós*: т.е. *вие*): заместени са от учтивите форми за множествено число (*ustedes* и *vocês*: т.е. *Вие*), които съвпадат с формите на глагола за трето лице множествено число (т.е. *те*). В бразилски португалски по същия начин е отпаднала формата за второ лице единствено число (*tu*): като вместо нея се използва учтивата форма *você*, която е в трето лице.

**Глаголи:** Глаголните времена са друга важна характеристика, по която се различават езиците. Така например, в български няма сегашно продължително време, и при превод на английски на „*Аз уча английски.*" трябва да се избира между "*I am studying English.*" ('*Уча английски /в момента/.*') и "*I study English.*" ('*Уча английски /по принцип/.*'). В италиански и френски пък има два вида минало свършено време: едно предимно за разговорната реч и друго – за писмената. Така е и в немски, в който освен това няма разлика между минало свършено и минало несвършено време. Това е сериозен проблем при превод към език, който прави такава разлика. Но дори когато дадени глаголни времена са еквивалентни (в даден контекст), може да има разлика в начина на образуването им. Така например, стандартният български език използва глагола *съм* като единствен спомагателен, напр. „*Чел съм го този вестник.*", докато в македонските диалекти такъв може да бъде и *имам* (както е било в старобългарски), например „*Го имам читано весников.*" Друг проблем са сложните съгласувания на времената. Например, на български казваме „*Ако мога, ще отида.*", т.е. имаме сегашно време в подчиненото изречение и бъдеще – в главното. На италиански такава комбинация от времена не е възможна и трябва да се използва сегашно-сегашно („*Se posso vado.*", буквално

---

[5] Сравнете с испанските показателни прилагателни: "*este hombre*" ('*този човек*', близо до говорещия), "*ese hombre*" ('*онзи човек*', близо до слушателя) и "*aquel hombre*" ('*онзи човек*', който е далеч и от двамата).

[6] Според някои лингвисти, в български има неопределителен член, който се изразява лексикално с формата *един* (напр. „*Видях една жена на улицата.*"), или с ∅ (напр. „*Видях жена на улицата.*").

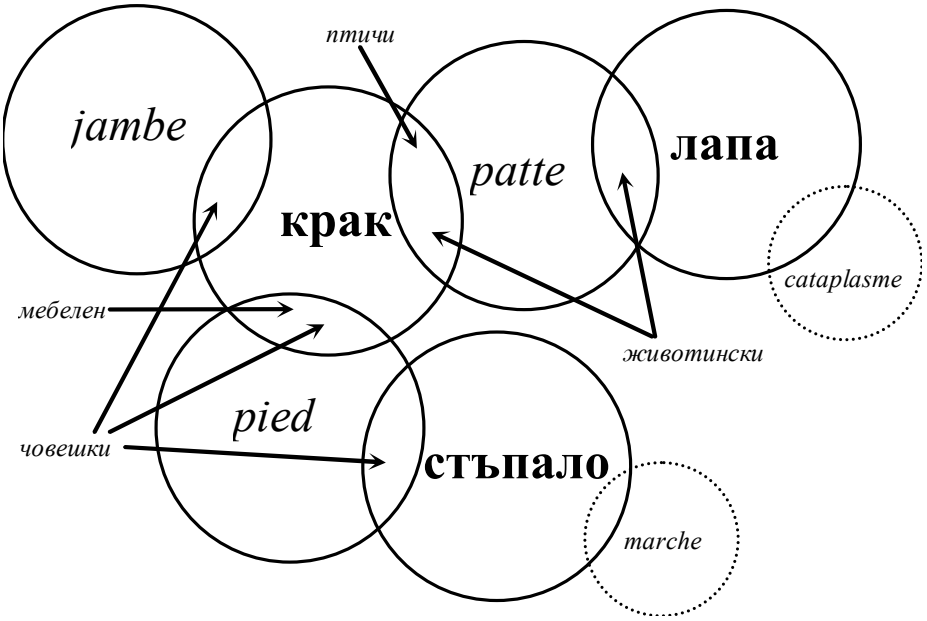
[7] В разговорния американски английски изразът *you guys* (буквално „*вие момчета*") се използва като лично местоимение и означава *вие* – множествено число, неформално. Употребява се за лица както от мъжки, така и от женски пол.

„*Ако мога, отивам.*“) или бъдеще-бъдеще (“*Se potrò andrò.*”, буквално *„*Ако ще мога, ще отида.*“). По подобен начин, за разлика от български, на английски има съгласуване на времената в минало време, например „*Каза, че ще дойде.*“ се превежда не с бъдеще време, а с бъдеще време в миналото като „*He said he would come.*“ (вместо *„*He said he will come.*“). Друг пример, на испански език: сравнете фрагментите: (1) „*Maria iba a México. Durante su visita, aprendió español.*“ (‘*Мария отиде в Мексико. По време на престоя си научи испански.*’) и (2) „*Maria fue a México. Cuando regresó, comenzó a hablar español.*“ (‘*Мария отиде в Мексико. Когато се върна, започна да говори испански.*’). На български и в двата случая се използва минало свършено време, докато на испански в първия случай има минало несвършено време, а във втория – минало свършено време. Други разлики произтичат от наклоненията. Така например, в български език има преизказно наклонение, което се употребява за събития, на които говорещият не е бил пряк свидетел. То често се използва във българските вестници, например „*Срещал се с оперативно интересни лица.*“ (сравнете със „*се срещаше*“), но няма аналог в повечето европейски езици и не може да се преведе точно. В немски за тази цел се използва подчинителното наклонение, което по принцип има друга основна функция. В някои случаи глаголно време в един език може да се изрази просто с наречие в друг, например на български се казва „*Бебето току-що яде.*“, докато на френски подчертаното наречие се превежда с минало непосредствено време: „*Le bébé vient de manger.*“ Накрая, но не на последно място, проблеми при превод могат да произтекат от самия глагол. Така например, в руски глаголът *съм* изобщо липсва в сегашно време, напр. „*Я болгарин*.“ (‘*Аз съм българин*.’). Подобно е положението в турски, където глаголът *съм* също липсва (не само в сегашно време), но съответното окончание си остава и се прикача към последната дума в изречението, например „*Ben bulgarım.*“. На другата крайност са някои индиански езици като Маках, в които практически всички думи са глаголи. В други езици глаголи са определени думи, които сме свикнали да приемаме като съществителни, например в корейски – думата за *нос*, а в Оджибве (индиански език) – дните от седмицата.

**Изпускане на подлога:** В някои езици, напр. български, руски, испански, италиански, турски и др., в които окончанието позволява лесно възстановяване на подлога, той може да се изпусне, например вместо „*Ние го разбираме.*“, можем да кажем „*Разбираме го.*“ Това не е възможно при езици с многозначни глаголни окончания като английски, френски и немски. Интересно е, че японският също допуска изпускане на подлога, без глаголната форма еднозначно да подсказва лице или число. Изпускането на подлога създава проблеми при превод, например, за да преведем изречението „*Каза, че идва в петък*.“ на английски език, трябва да изберем първо кой върши действието: *той*, *тя* или *то*. Освен това, може да бъде свързано с разлика в словореда в някои езици (напр. бълг. „*Ние го разбираме*.“, но „*Разбираме го*.“), но не и в други (напр. в македонските диалекти се казва „*Го разбираме.*“). В португалски словоредът се променя като в български: казва се „*Nós o comprendemos.”*, но „*Comprendemo-lo*.“. В испански обаче е като в македонските диалекти: “*Nosotros lo comprendemos*.” става “*Lo comprendemos*.”.

**Структурни разлики:** В някои случаи, при превод на друг език се налага смяна на тематичната роля, което изисква правилно определяне на логическия субект и на логическия обект на действието на семантично ниво. Например, в английското изречение “*You like her.*”, т.е. ‘*Ти я харесваш.*’, подлог е субектът, докато на испански подлог става обектът на действието: „*Ella te gusta.*“, т.е. ‘*Тя ти харесва*.’. На български са възможни и двете конструкции. По подобен начин „*Имам много работа.*“ се превежда на турски като „*Çok işim var.*“, което буквално означава „*Много работа-моя има.*“ Някои езици пък правят разлика между одушевени и неодушевени предмети. Така например, в руски има различни окончания във винителен падеж, съвпадащи с формата за родителен или именителен падеж, в зависимост от това дали предметът е одушевен или – не: например, „*Я вижу медведя.*“ (‘*Аз виждам мечка.*’, т.е. вин. пад. = род. пад.), но „*Я вижу дом.*“ (‘*Аз виждам къща.*’, т.е. вин. пад. = им. пад.). В испански пред одушевените предмети се употребява така нареченото „персонално *a*”, което не е предлог; горните два примера се превеждат съответно като “*Veo a un oso.*” и “*Veo una casa.*”. Друг чест проблем при превод е, че някои глаголи приемат пряко допълнение в

един език и непряко – в друг. Например, „*Тя влезе в стаята*.” се превежда на английски без предлог: „*She entered the room.*” Друг източник на разлики може да бъде позицията на предлога в относителните изречения, например лозунгът на Барак Обама „*A Change We Can Believe In*” буквално се превежда като „*Промяна ние можем повярваме в*”, докато правилното на български е „*Промяна, в която можем да повярваме*”. Забележете още, че в английски език е възможно изпускане на относителното местоимение. Сравнете още как паралелни начини на изказ в български като „*Искам да дойда.*” и „*Искам да дойдеш.*” се превеждат по различен начин на испански, съответно като „*Quiero llegar.*” (с инфинитив) и „*Quiero que tú llegues.*” (с подчинително наклонение). Така е и в другите романски езици, като португалски допуска и паралелен превод, тъй като позволява инфлектен инфитив: „*Quero chegar*.” и „*Quero chegares*.” (но е възможно и „*Quero que tu chegues*.” – с подчинително наклонение)




***Фигура 1. Различен светоглед в български и френски език по отношение на „крак” и сродните му понятия.***

**Непреводими думи, реалии, лексикални дупки:** Както показва Фигура 1, естествените езици често правят различно разделяне на света по отношение на различни понятия. По подобен начин, на английското *uncle* в български съответстват три по-специфични думи: *вуйчо* (‘*брат на майката*’), *чичо* (‘*брат на бащата*’) и *свако/калеко/лелинчо* (‘*съпруг на сестрата на бащата/майката*’). На японски пък няма дума за брат, а вместо това има *отоото* (‘*по-малък брат*’) и *ониисан* (‘*по-голям брат*’). По подобен начин българската дума *риба* се превежда на испански като *pez*, в смисъл на живия организъм, но като *pescado*, в смисъл на суровината за готвене (в български език има такава разлика между *свиня* и *свинско*). При превод от език с по-общи думи към такъв с по-специфични, се налага задължителен избор на по-специфично съответствие. Контекстът често помага (напр. при избор между *pez* и *pescado*), но в някои случаи може да няма достатъчно информация (напр. при избор между *чичо*, *вуйчо* и *свако*) и думата става непреводима. Друг широк клас непреводими думи са реалиите, които са свързани със специфичната култура на носителите на езика, например *таратор*, *мартеница*, *кукери*, *читалище*, *мутра*, *чалга*, *бурканбанк*, *менте* и др. В повечето европейски езици няма аналог за български думи като *църцори* (нещо средно между *тече* и *капе*) и за роднински връзки като *бате* и *кака*, които не са реалии. Накрая, но не на последно място, българският език позволява лексикализации, комбиниращи в една дума движение с начина, по който то се извършва, напр. „*Иван преплува канала.*” В испански това не е възможно, в резултат на което възниква „лексикална дупка” по отношение на глагола *преплувам* и при превод се налага допълнително използване на деепричастие: “*Ivan cruzó el canal nadando*.”, т.е. ‘*Иван пресече канала плувайки.*’ На испански пък има глагол *madrugar*, означаващ *ставам рано*, който липсва в български (лексикална дупка).

**Устойчиви словосъчетания, мултилексемни и идиоматични изрази:** Някои изрази влизат в устойчиви словосъчетания, например на български се казва *вземам решение*, но на английски език решението обикновено „се прави": *to make decision*. Някои мултилексемни изрази пък са свързани с превод на една и съща дума по различен начин в зависимост от относителната ѝ позиция в изречението, например на английски се казва „*John is as tall as Mary.*", докато на испански имаме „*Juan es tan alto como María.*" (т.е. на първото *as* съответства *tan*, а на второто – *como*). На български съответствието е „*Иван е (толкова) висок колкото Мария.*", т.е. *толкова* може изобщо да се изпусне. Друг клас трудно преводими изрази са идиомите. Например, английският израз *to kick the bucket* (буквално „*да ритнеш букетчето*") може да се преведе на български с изрази като *да гушнеш чимширчето*, *да се усмихнеш между букетчетата* и *да ритнеш камбаната/кофата*, никой от които не е буквален превод. На испански съответният израз е *estirar la pata* (буквално „*да опънеш лапата*") и най-близките до него български изрази отново не са буквален превод: *да навириш петалото*, *да изпънеш жартиера*. Някои идиоматични изрази са специфични за даден език и често изискват познания за културата на носителите му, напр. *да минеш като през турски гробища*. Това важи и за по-нови изрази като *да го отнесеш като депутат кюфте*, *да си живееш като Симеончо през фашистко* и др.

**Други многозначности:** Основен проблем при автоматичния превод е, че многозначността в даден контекст се разрешава трудно, а понякога това е и невъзможно. На лексикално ниво понякога различни думи могат да имат една и съща форма, напр. англ. *book* означава *книга* като съществително, но *резервирам* като глагол. Същия тип многозначност е налице в рекламното съобщение „*Ако мълчиш, си няма жертва. Ако говориш, няма жертва.*" На семантично ниво една дума може да има няколко различни значения, например англ. *ball* може да означава както *бал*, така и *топка*. Съществен проблем при превод на испански пък е, че има два глагола, съответстващи на *съм*[8]: *ser* и *estar*. Например при превод на „*Мария е красива.*" трябва да се избира между „*María es linda.*" (т.е. красива е по принцип) и „*María está linda.*" (т.е. красива е в момента). На синтактично ниво, може да има различни видове референтна многозначност, например в изречението „*Ей го оня с коня, дето върви пеша и го носи.*", не е ясно за кого се отнасят местоимението *го* и подчиненото изречение *дето върви пеша и го носи*. При превод на английски обаче трябва да се вземе съответно решение: например, *го* трябва да се преведе като *it* (т.е. коня) или *him* (т.е. човека). В много случаи подобна многозначност е „невидима" за хората, но е реална за компютъра, който няма достатъчно знания за света. Сравнете изречнията „*Четох книга за еволюцията през последните 10 милиона години.*" и „*Четох книга за еволюцията през последните 10 минути.*" В пресата, нарочно или несъзнателно, такива многозначни примери се срещат в изобилие, особено в заглавия. Например: „*Катаджия иска нафта от шофьор, нямал пари*" (Кой нямал пари?), „*МВР погва баровци с джетове и яхти*" (Кой кара джетовете и яхтите?), „*Ако бирата пречи на работата Ви, оставете я.*" (Кое да оставим?), „*Летърман призна за секс с колежки след изнудване*" (Признанието ли е след изнудване или сексът?), „*Откриха екзотични животни без разрешително*" (Животните ли нямат разрешително или за откриването им не е имало разрешение?), „*Еврократите се отърваха от удръжки заради България*" (Благодарение на България ли са се отървали или причината за евентуални удръжки е щяла да бъде България?), „*Минаваме на живот под наем*" (Т.е. наемаме жилища или си вземаме живот под наем?), „*Втора страна от ЕС търси спасение от МВФ*" (Търси някой да я спаси от МВФ или вика МВФ на помощ?), „*Бандити застреляха търговец по време на грабеж*" (Кой е извършвал грабежа: бандитите или търговецът?), „*Либия предложи на Украйна да отмени визовия режим*" (Кой ще отмени визовия режим: Либия или Украйна?), „*Косовските сърби все по-застрашени от албанците*" (Застрашени са от страна на албанците или са по-застрашени отколкото са албанците?), „*Софиянци няма да напускат града заради земетресенията, показа анкета*" (Няма да напускат *въпреки* или *поради* земетресенията?), „*100 полицаи срещу телефонната мафия, 13 в ареста*" (Кого са пратили в ареста?), „*Абхазия ще отбележи 16-*

[8] На български също има два глагола: *съм* и *бъда*, но разликата в значенията е по-малка. Например, можем да кажем както „*Аз ще съм щастлив.*", така и „*Аз ще бъда щастлив.*"

*годишнината от освобождаването на Сухуми от грузинските войски*" (Грузинските войски окупатор или освободител на Сухуми са били според Абхазия?), „*Жалко е България да се свързва с корупция и купуване на гласове в Западна Европа*." (В Западна Европа ли така гледат на България или България купува гласове там?), „*Вашингтон отправи покана към отбора за борба с мафията на Бойко Борисов*." (Бойко Борисов ще се бори ли срещу мафията или е неин шеф?), „*Ирландия гласува за Лисабонския договор*" (Подлага договора на гласуване или гласува положително за него?), „*Барак Обама ще съобщи номинацията си за Върховния съд на САЩ по-късно днес*" (Барак Обама ще се кандидатира ли или ще номинира някой друг?), „*Половината далавери с европари се разследват след сигнал от ОЛАФ*" (Половината от разследванията са в следствие на сигнал на ОЛАФ или заради сигнал на ОЛАФ се разследват половината известни иначе от други източници далавери? Колко сигнала на ОЛАФ има: един или повече?), „*US съд отреди Куба да плати $1,2 млрд. на имигрант*" (На един конкретен имигрант или на всеки имигрант? Имигрантът в Куба или в САЩ се е заселил?), „*В България кандидатурата пази от затвора*" (Кандидатурата пази някого да не влезе в затвора или тя самата седи в затвора и пази оттам?) Накрая: *"Това, което се случва около СДС, е тест за демократичните институции. Прави се опит за приватизация на една партия с неполитически средства", бе коментарът на бившия външен министър и лидер на СДС Надежда Михайлова.* (Тя бивш или настоящ лидер на СДС е?)

## Човешкият превод

Преди да видим как компютърът би могъл автоматично да превежда, например от български на английски език, нека първо помислим как го прави човек. Едно възможно обяснение би могло да бъде следното:

1. Човек *чете* и *разбира* българския текст.
2. Човек *генерира* съответен английски превод.

Съгласно това предположение, човек би следвало да изгражда някакво междинно мислено представяне на смисъла (семантиката) на българския текст, от което след това да генерира съответен превод. Дали при превод на друг език, напр. френски, това вътрешно представяне ще се промени или ще се запази (във втория случай става дума за *интерлингва*, виж Фигура 3)? Когнитивните специалисти и лингвистите не могат да дадат еднозначен отговор.

Всъщност, не е необходимо човек да е прочел и разбрал *целия* текст, за да започне да превежда, например при симултанен превод. Вместо това, обикновено преводът се извършва на части – изречения, малки фрагменти или дори отделни думи – като при нужда се взема предвид информацията в останалата част на текста.

Много учени изобщо отричат необходимостта от дълбок анализ и цялостно разбиране при превод. Така например, през 1984 г. Макото Нагао дефинира следните два фундаментални принципа на превода [*Nagao,1984*]:

(1) *Човек не превежда едно просто изречение чрез дълбок лингвистичен анализ.*

(2) *Човек превежда, като първо разделя по подходящ начин изречението на фразови фрагменти, след което ги превежда с други фразови фрагменти, които накрая обединява по подходящ начин в цяло изречение. Преводът на всеки фразов фрагмент се извършва по принципа на аналогията, с използване на подходящи примери.*

Накрая, изглежда естествено, че за да може да превежда от български на английски език, човек трябва да владее и двата езика. Всъщност, това не е абсолютно необходимо: съгласно горните принципи, ако има достатъчно подходящи примерни преводи, които да използва за справка, човек може да превежда по аналогия прости изречения и без да владее никой от двата езика. Така например, ако разполага със следните примери

Пример 1:

- англ: *(He buys) a notebook.*
- яп.: *(Kare ha) nouto (wo kau).*

Пример 2:

- англ: *I read (a book on international politics).*
- яп.: *Watashi ha (kokusaiseiji nitsuite kakareta hon) wo yomu.*

дори и да не владее нито японски, нито английски, човек лесно може да съобрази, че изречението

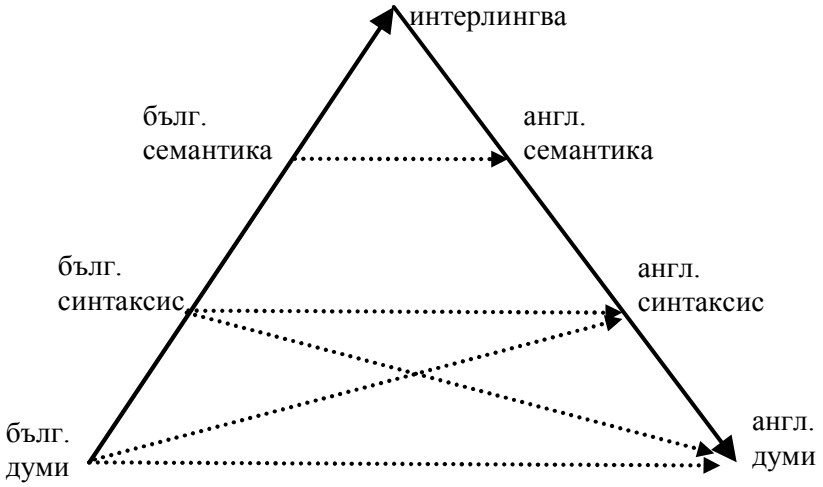
He buys *a book on international politics*.

би следвало да се превежда на японски като

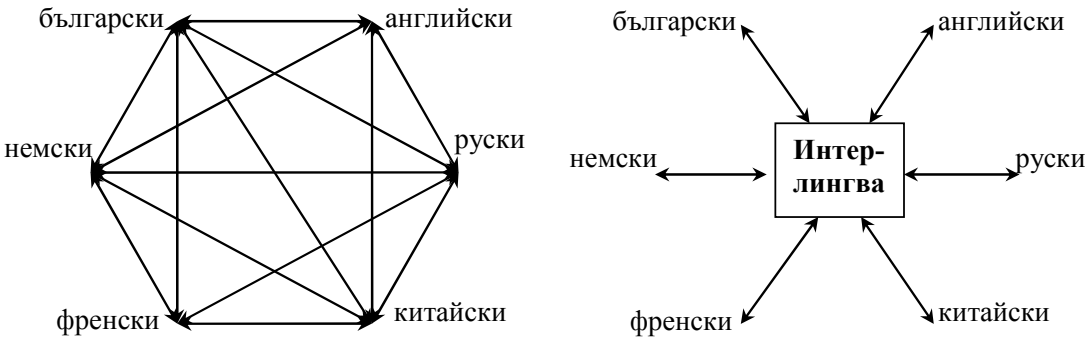
(Kare ha) *(kokusaiseiji nitsuite kakareta hon)* (wo kau).

Така формулирана, задачата за машинния превод прилича на задача на математическата лингвистика, но е по-сложна, тъй като машината не владее нито един от двата езика, докато при математическата лингвистика човек най-често разсъждава върху съответствия между някакъв непознат език и своя роден.

## Нива на трансфер и интерлингва




***Фигура 2. Нива на трансфер:*** *В идеалния случай българският текст първо се анализира на ниво дума, а след това и на синтактично и семантично ниво, след което се строи интерлингва, от която се генерира съответно английско семантично и синтактично представяне, и накрая – съответният английски превод. Повечето съвременни системи за машинен превод съкращават анализа по някоя от пунктираните линии, най-често от някое от най-долните две нива.*




***Фигура 3. Използване на интерлингва:*** *В общия случай, за да се превежда в двете посоки между 6 езика, са необходими общо 30 системи за машинен превод, докато при използване на интерлингва са достатъчни 12. А при добавяне на седми език ще са нужни съответно 12 и 2 допълнителни системи. Дефинирането на подходяща интерлингва обаче засега остава нерешена задача.*

Фигура 2 показва схематичен триъгълник с различни възможни нива на трансфер при превод от български на английски език. В идеалния случай българският текст първо се анализира на ниво дума, а след това и на синтактично и семантично ниво, след което се строи интерлингва, от която се генерира съответно английско семантично и синтактично представяне, и накрая – съответен английски превод. Както видяхме по-горе, повечето съвременни системи за машинен превод съкращават анализа по някоя от пунктираните линии, най-често от някое от най-долните две нива. Използването на семантично представяне – както езиково зависимо, така и езиково независимо (т.е. интерлингва) – засега остава извън възможностите на съвременните технологии за машинен превод, макар и да са правени някои опити в това отношение върху силно ограничени обеми данни.

Досегашното развитие на машинния превод недвусмислено показва, че всяко успешно изкачване нагоре по триъгълника от Фигура 2 води до по-добри резултати и до малка „революция". За момента са овладени едва първите две нива, при това не напълно: (1) превод на ниво думи – изцяло, и (2) използване на синтаксис – донякъде. Включително са овладени и някои междинни нива – превод с цели фрази и превод с йерархични фрази. Несъмнено стремежът към семантика и, евентуално, интерлингва ще продължи, защото потенциалните ползи са несъмнени. Така например, както показва Фигура 3, в общия случай, за да се превежда в двете посоки между 6 езика, са необходими общо 30 системи за машинен превод, докато при използване на интерлингва са достатъчни 12, а при добавяне на седми език към системата ще са нужни съответно 12 и 2 допълнителни системи. Дефинирането на подходяща интерлингва обаче засега остава нерешена задача.

## Статистически машинен превод: модел на канала с шум

С цел простота на изложението по-долу ще предположим, че преводът се извършва изречение по изречение, без отчитане на съдържанието на останалата част от текста. Така работят повечето съвременни системи за машинен превод.

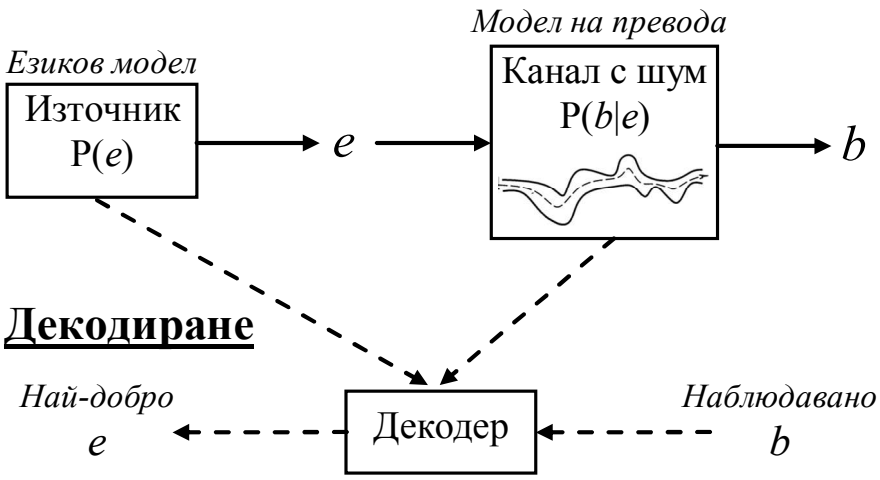
За по-голяма определеност ще предположим, че искаме да превеждаме от български на английски език. Така задачата за статистическия машинен превод може да се формулира по следния начин: *по дадено българско изречение b да се намери най-добрият му съответен английски превод e*. Ако означим с P($e|b$) вероятността $e$ да бъде превод на $b$, получаваме следната формулировка: при дадено $b$ да се намери $e$, за което вероятността P($e|b$) е максимална. Това ни води до уравнение (1), където с $\hat{e}$ сме означили търсеното максимално вероятно $e$. По правилото на Бейс, уравнение (1) може да се запише като (2), което от своя страна е еквивалентно на (3), тъй като знаменателят на (2) не зависи от $e$.

$$\hat{e} = \arg\max_e \mathrm{P}(e \mid b) \quad (1)$$

$$= \arg\max_e \frac{\mathrm{P}(b \mid e)\mathrm{P}(e)}{\mathrm{P}(b)} \quad (2)$$

$$= \arg\max_e \mathrm{P}(b \mid e)\mathrm{P}(e) \quad (3)$$

Това е моделът на *предаване на информация по канал с шум*, приложен към статистическия машинен превод. Съгласно този модел, българското изречение $b$ се разглежда като повреден вариант на английския оригинал $e$. Моделът обяснява процеса на генериране на българското изречение $b$ като двустъпков процес: първо се генерира английско изречение съгласно *модела на източника* P($e$), а след това то се предава по канала с шум и се поврежда съгласно *модела на канала* P($b|e$). Нашата задача е при зададени P($e$), P($b|e$) и българско изречение $b$ да намерим най-вероятното съответно английско изречение $e$. Моделът е показан графично на Фигура 4.




***Фигура 4. Преводът като предаване на информация по канал с шум.*** *Горната част на схемата показва генеративен модел, описващ процеса на получаване на българското изречение* ***b****: първо, източникът генерира случайно английско изречение* ***e*** *с вероятност* ***P(e)****, след което* ***e*** *се изпраща по канал с шум, който с вероятност* ***P(b|e)*** *преобразува* ***e*** *в* ***b****. Задачата за превод от български на английски език се разглежда като декодиране, т.е. търсене на най-вероятното* ***e*** *по дадени* ***b****,* ***P(e)*** *и* ***P(b|e)****, както е показано в долната част на схемата.*

На пръв поглед не е ясно какво се постига с преобразуването на (1) в (3) – започва се с P(*e*|*b*), а в крайна сметка се стига до P(*b*|*e*), което изисква превод в обратната посока, който е също толкова труден, както и в правата посока. В допълнение обаче се получава и допълнителен множител P(*e*), чието значение ще стане ясно по-долу.

Уравнение (3) всъщност дава трите основни компоненти на една система за статистически машинен превод:

- **езиков модел** P(*e*)
- **модел на превода** P(*b*|*e*)
- **декодер**, който по зададено *b* търси най-вероятното *e*

**Езиковият модел** P(*e*) показва колко е вероятно да бъде казано дадено английско изречение *e*. Моделът следва да дава висока вероятност на граматично правилни изречения (напр. „*I eat an apple*.") и ниска вероятност на неправилни (напр. „*Apple is eats I*."), както и на граматично правилни, но малко вероятни изречения (например в семантично отношение: „*The colourless green idea sleeps furiously*.", т.е. „*Безцветната зелена идея спи яростно*." – пример на Чомски от 1957 г.).

**Моделът на превода** P(*b*|*e*) следва да дава висока вероятност за двойката (*b*,*e*), ако *b* би могло да бъде превод на *e*, и ниска вероятност – в противен случай. Забележете, че моделът на превода се интересува единствено от това дали е вероятно двете изречения да са превод едно на друго, без значение дали *e* е граматично правилно построено изречение. Последното е задача на езиковия модел P(*e*). Разделянето на двете задачи в отделни модели позволява от една страна езиковият модел да се опрости значително, а от друга – създава възможност двата модела да се обучават поотделно, включително от различни изследователски групи. Възможно е дори обучаване върху различни учебни данни: както ще видим по-долу, моделът на превода изисква паралелен двуезичен текст, който се намира трудно, докато езиковият модел използва само текст на английски език, който е леснодостъпен в големи обеми.

**Декодерът** е търсещият алгоритъм, който по зададено *b* се опитва да намери най-вероятното *e*, т.е. това, което максимизира произведението P(*e*).P(*b*|*e*). Така например, ако превеждаме *жената котка* на английски, това произведение ще бъде голямо за *the cat woman*, но малко за *woman cat the* (поради малък първи множител), за *Batman forever* (поради малък втори множител) и за *dog an a for* (заради двата множителя). Виж Таблица 1.

***Таблица 1. Как се превежда на английски език „жената котка”?*** *Добрият превод трябва да бъде едновременно граматично правилен и смислово верен.*

| | **P(e)** | **P(f\|e)** | **P(e).P(f\|e)** |
|---|---|---|---|
| *woman cat the* | 👎 | 👍 | 👎 |
| *woman the cat* | 👎 | 👍 | 👎 |
| *a cat woman* | 👍 | 👎 | 👎 |
| *cat woman* | 👍 | 👎 | 👎 |
| *the woman* | 👍 | 👎 | 👎 |
| *Batman forever* | 👍 | 👎 | 👎 |
| *dog an a for* | 👎 | 👎 | 👎 |
| ***the cat woman*** | 👍 | 👍 | 👍 |

Моделът на канала с шум отразява факта, че човек превежда по-лесно от чужд език на своя собствен, отколкото в обратната посока: макар че „човешкият модел на превода” вероятно е еднакво добър в двете посоки, „човешкият езиков модел” е много по-добър за родния език, което прави преводите към този език по-гладки. Това се взема предвид от институции като Европейската комисия, която изисква официалните й документи да бъдат превеждани от преводачи, за които езикът, към който се превежда, е роден. Преводачи, за които роден е само езикът, от който се превежда, се считат за недостатъчно квалифицирани. По подобен начин при ръчно оценяване на качеството на система за машинен превод най-често се използват два отделни 5-степенни критерия: *адекватност* (каква част от информацията се съдържа в превода: цялата, повечето, много, малко, никаква) и *гладкост* (доколко гладък е изказът на превода: безупречен, добър, “нероден”, несвободен, неразбираем).

След 2002 г. обаче доминиращият начин на оценяване е автоматичен и използва *BLEU* (*Bi-Lingual Evaluation Understudy*) – специална оценка, предложена от екип на IBM, която измерва в каква степен генерираният от машината превод съвпада на ниво последователности от думи с дължина 1, 2, 3 и 4 (*n*-грами) с един или повече еталонни човешки превода [*Papineni&al.,2002*]. Пресмятанията се извършват по следната формула:

$$BLEU = BP.\exp\left(\sum_{n=1}^{4}\frac{\log p_n}{n}\right)$$

където $p_n$ е $n$-грамна точност спрямо еталоните, а $BP = \exp(\max(0,1-r/c))$ е наказание за прекалено къс превод ($c$ и $r$ са съответно дължините на на машинния и на най-късия измежду еталонните човешки преводи).

## Езиковият модел P(e)

Както споменахме по-горе, езиковият модел P($e$) показва колко е вероятно да бъде казано дадено английско изречение $e$. Ако разгледаме $e$ като последователност от думи $w_1w_2\ldots w_n$, можем да представим съответната вероятност като (4), което е еквивалентно на (5) съгласно правилото на веригата. Тъй като нямаме достатъчно данни, за да научим всички вероятности от (5), можем да направим приближението (6), което представлява Марковски модел от ред 2.

$$\begin{aligned} P(e) &= P(w_1w_2...w_n) && (4)\\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1w_2)P(w_4 \mid w_1w_2w_3)P(w_5 \mid w_1w_2w_3w_4)... && (5)\\ &\approx P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)P(w_4 \mid w_3)P(w_5 \mid w_4)... && (6)\end{aligned}$$

Марковският модел от ред $n$ приближава вероятността на дадено изречение като произведение от условни вероятности на всяка дума при условие предходните $n-1$. Така например, при $n = 2$ вероятността на “*I eat an apple*” се пресмята така:

P("*I eat an apple*") = P(*I* | <S>) . P(*eat* | *I*) . P(*an* | *eat*) . P(*apple* | *an*) . P(</S> | *apple*)

В горната формула <S> и </S> са специални символи съответно за начало и за край на изречение и са необходими, когато искаме да моделираме последователности с различна дължина.

Условната вероятност $P(w_i \mid w_{i-1})$ се пресмята най-общо като отношение на броя на срещанията $C(w_{i-1}w_i)$ на последователността от думи „$w_{i-1}\ w_i$" и на срещанията $C(w_{i-1})$ на думата $w_{i-1}$ в голям обем текстове на английски език:

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i)}{\sum_{w_i} C(w_{i-1}w_i)} = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$$

На практика се използват по-сложни формули, които умеят да се справят с нулеви числители/знаменатели, както и с редки думи. Най-простото, но и най-неефективно решение е добавяне на някаква константа $\varepsilon$ ($\varepsilon > 0$), например $\varepsilon = 1$, към всички $C(w_{i-1}w_i)$, преди използването им във формулата. Така получаваме следния вариант:

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i) + \varepsilon}{\sum_{w_i}\left[C(w_{i-1}w_i) + \varepsilon\right]} \approx \frac{C(w_{i-1}w_i) + \varepsilon}{C(w_{i-1}) + \varepsilon \mid V \mid}$$

където $|V|$ е броят на различните думи в текста.

Обзор на най-популярните алтернативи, с подробна теоретична и практическа оценка, може да бъде намерен в [*Chen&Goodman,1999*].

# Моделът на превода P(b|e)

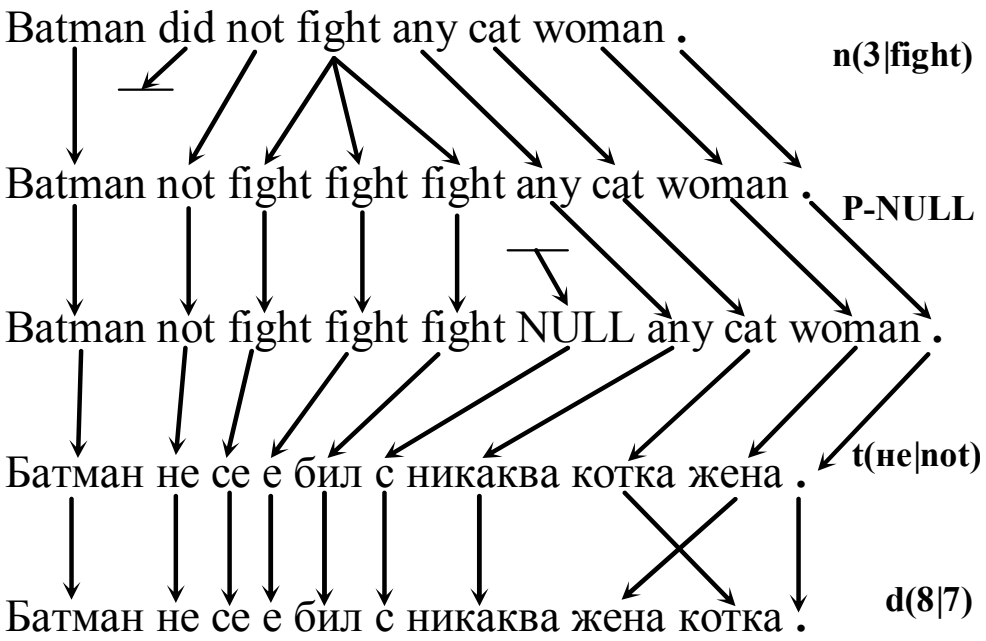
Моделът на превода P(*b*|*e*) се учи от двуезичен паралелен корпус от български изречения и техния английски превод. Директното моделиране на P(*b*|*e*) в общия случай е невъзможно, тъй като е почти изключено да сме видели изреченията *e* и *b* в учебния корпус, особено в случай на по-дълги изречения. Това налага представяне на P(*b*|*e*) с помощта на вероятности за превод на по-малки фрагменти.

### *Превод дума-по-дума*

Някои възможни решения се дават от класическите модели за статистически машинен превод, разработени от IBM през 1991 г. В статия от 1993 г. [*Brown&al.,1993*] са публикувани 5 модела, по-късно станали известни[9] като *модели на IBM 1, 2, 3, 4* и *5*. С цел простота на изложението по-долу ще изложим основната идея на модел 3 на IBM. Това е генеративен модел, описващ процеса на трансформация на английско изречение в българско като поредица от четири стъпки, с всяка от които е асоциирана съответна вероятност. Процесът е илюстриран на Фигура 5, където е показано преобразуването на „*Batman did not fight any cat woman*." в „*Батман не се е бил с никаква жена котка*.". На първата стъпка се решава с колко думи ще се превежда всяка една английска дума: 0,1,2..., което се контролира от съответна вероятност $n(k|e_i)$, моделираща „плодовитостта" на съответната английска дума $e_i$. В нашия случай, спомагателният глагол *did* изчезва, глаголът *fight* се превежда с три български думи, а останалите английски думи се превеждат с по една дума. На втората стъпка на някои позиции се вмъква празната дума: това става с еднаква вероятност P-NULL, независеща от думите и позицията. В нашия случай се вмъква празна дума, която по-късно ще съответства на предлога *с*, за който няма съответна дума в английското изречение. На третата стъпка всяка английска дума $e_i$ се превежда с единствена съответна българска дума $b_j$ съгласно вероятността за превод $t(b_j|e_i)$. Накрая, някои от българските думи се разместват съгласно вероятност $d(j|k)$: в нашия случай се разместват *жена* и *котка*.

[9] През 2003 г. Ох и Ней предлагат модификация на Модел 5 на IBM, която наричат Модел 6 [*Och&Ney, 2003*]. Модели 5 и 6 обаче представляват по-скоро теоретичен интерес и почти не се използват в реални системи.

Четирите вероятности лесно могат да се научат, ако за всяка двойка изречения в двуезичния корпус са дадени съответни подравнявания на ниво дума както на Фигура 6. В общия случай обаче такива подравнявания не са налични и се налага автоматичното им получаване с помощта на алгоритъм за максимизиране на очакването, описан в [*Brown&al.,1993*].




***Фигура 5. Модел 3 на IBM:*** *Това е генеративен модел, описващ процеса на трансформация на английско изречение в българско като поредица от четири стъпки. Първо се решава с колко думи ще се превежда всяка една английска дума: 0,1,2... Второ, на някои позиции се вмъква празната дума. Трето, всяка английска дума се превежда със съответна българска. Накрая, някои от българските думи се разместват. С всяка от стъпките е асоциирана съответна вероятност (показана вдясно).*


***Фигура 6. Подравняване на ниво думи.***

Таблици 2 и 3 показват примерни преводи, научени с помощта на модел 3 на IBM за английските думи *main* и *farmers* заедно със съответните вероятности за превод в проценти. В Таблица 2 се вижда, че *main* се превежда с форми на прилагателните *основен* и *главен*: в единствено/множествено число, в мъжки/женски/среден род, с/без членуване. Таблица 3 е още по-интересна и показва, че моделът е успял да научи, че *farmers* може да се преведе не само като *фермери*, но и като *земеделски/селскостопански/частни стопани/производители*.

*Таблица 2. Български преводи на английската дума* **main***, извлечени с помощта на модел 3 на IBM*.

| Български превод | Вероятност (в %) |
|---|---|
| *основни* | 34,09 |
| *основните* | 22,73 |
| *основната* | 13,64 |
| *основно* | 6,82 |
| *основното* | 6,82 |
| *главната* | 4,55 |
| *важните* | 2,27 |
| *главен* | 2,27 |
| *главните* | 2,27 |
| *главният* | 2,27 |
| *основният* | 2,27 |

*Таблица 3. Български преводи на английската дума* **farmers***, извлечени с помощта на модел 3 на IBM*.

| Български превод | Вероятност (в %) |
|---|---|
| *фермери* | 82,76 |
| *стопани* | 6,37 |
| *селски* | 2,67 |
| *селскостопанските* | 2,36 |
| *земеделските* | 1,88 |
| *земеделски* | 1,80 |
| *частни* | 1,13 |
| *производители* | 1,02 |

Важно ограничение на моделите на IBM е, че могат да учат само отношения от тип 1:M (едно към много), но не и M:1 (много към едно) или M:M (много към много), т.е. една английска дума може да поражда 0, 1 или повече български думи, но една българска дума може да се поражда само от една английска дума. Това означава, че ако на Фигура 5 вместо *any cat woman* имахме *the woman*, нямаше да можем да генерираме *жената* едновременно от *the* и от *woman* (както би било правилно), а щеше да трябва да приемем, че *woman* генерира *жената*, а *the* изчезва. Очевидно това създава трудности с определителния член при превод от английски на български с модел 3 на IBM, тъй като се губи връзката между членната форма в българското и в английското изречение.

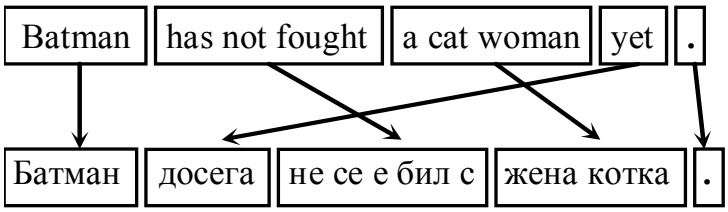
Друг проблем е невъзможността за отчитане на контекста: напр. *interest rate* трябва да се преведе като *лихвен процент*, докато *interest in* – като *интерес към*. Неотчитането на контекста силно затруднява и съгласуването по род, число, падеж, членуването и др. и може да доведе до генериране на фрази като „*главната прокурорите*”, които невинаги могат да се коригират успешно от езиковия модел.

### *Превод с цели фрази*

*Преводът с цели фрази* [*Koehn&al.,2003*] дава решение на повечето от проблемите, свързани с моделите на IBM. Отново става дума за генеративен модел, описващ процеса на трансформация на английско изречение в българско като поредица от стъпки, но този път на ниво цели фрази. Първо, английското изречение се разделя на “фрази”. След това, всяка английска фраза се превежда със съответна българска. Накрая, някои от българските фрази се разместват. С всяка от стъпките е асоциирана съответна вероятност, която отново се учи от паралелен двуезичен корпус от двойки изречения на български и техните преводи на английски език. На Фигура 7 е даден конкретен пример.

Моделът води до съществено подобрение на качеството на машинния превод, тъй като умее да учи съответствия от тип M:M и дава възможност за по-добро отчитане на локалния контекст. Таблица 4 показва някои автоматично извлечени български фрази и техния английски превод. В първата част на таблицата се вижда, че моделът е научил, че думата *както* може да се превежда по различни начини в зависимост от контекста, например като *both*, *like*, *as*, *as well as*, *in line with*, *and* и др.

Забележете още, че извлечените фрази не представляват непременно лингвистични единици и могат да включват препинателни знаци, което позволява да се научат разлики в правилата за пунктуация между двата езика, например „*както следва :*" → „*as follows ,*". Във втората част на таблицата виждаме някои възможни преводи на *главния*: както като отделна дума (напр. *the base*, *the chief*), така и в комбинация с други думи. Така например, *главни прокурори* (забележете, че в рамките на фразата имаме правилно съгласуване по род и число) се превежда като *chief prosecutors*, докато *главни методисти* – като *senior instructors*, *главни улици* – като *main streets*, а *главно предизвикателство* – като *major challenge*. Накрая, вижда се, че за една българска фраза може да има множество възможни английски алтернативи за превод. Както обяснихме по-горе, подходящият превод в конкретно изречение се избира като се взема предвид не само вероятността за превод на двойката фрази, но и вероятността за превод на цялото изречение според лингвистичния модел.




***Фигура 7. Превод с цели фрази:*** *Това е генеративен модел, описващ процеса на трансформация на английско изречение в българско като поредица от три стъпки. Първо, английското изречение се разделя на фрази. Второ, всяка английска фраза се превежда със съответна българска. Накрая някои от фразите се разместват.*

***Таблица 4. Автоматично извлечени български фрази и техният английски превод.*** *Фразите не са непременно лингвистични единици и могат да включват препинателни знаци.*

| **Българска фраза** | **Английска фраза** |
|---|---|
| както физическа , така и психическа | both physical and psychological |
| както целият регион | like the whole region |
| както те са определени | as defined |
| както и размера | as well as the size |
| както и предишните редовни доклади | in line with previous regular reports |
| както и по други | and in other |
| както и относно | and also about |
| както са | as were |
| както се предоставя на | as provided to |
| както следва : | as follows , |
| главния | the base |
| главния | the chief |
| главния | the main |
| главния | the principal |
| главни прокурори | chief prosecutors |
| главни счетоводители | chief accountants |
| главни архитекти | chief architects |
| главни щабове | main staffs |
| главни улици | main streets |
| главни методисти | senior instructors |
| главно предизвикателство | major challenge |

Забележете, че думите в Таблица 4 не са непременно в основната си форма – могат да бъдат в множествено число, членувани и т.н. Това е следствие от факта, че моделът не знае нищо за морфология, синтаксис, пунктуация и др. За него думата *главния* е толкова различна от *главният* и от *главен*, колкото и от *куче*, *зелени*, *играят*, *да, от* или , (запетая). Напълно възможно е моделът да е научил фраза за правилен превод на *главни прокурори*, но не и за *главен прокурор*. По подобен начин, в някои случаи моделите на IBM може да могат да преведат *главния*, но не и *главен*, например защото никога не са го срещали в паралелния текст, върху който са обучавани. Това не е сериозен проблем при превод между английски и френски език (с които първоначално са експериментирали в IBM), но е сериозен недостатък при превод между български и английски, тъй като българският е по-силно флективен, т.е. има повече различни форми на думите. Така например, в български за *главен* има форми като *главен*, *главния*, *главният*, *главна*, *главната*, *главни*, *главните*, докато на английски за *main* има една-единствена форма, независима от род, число и членуване.

Накрая, при превод с цели фрази най-често не се търси най-доброто произведение

$$\hat{e} = \arg\max_{e} \mathrm{P}(b \mid e)\mathrm{P}(e)$$

а се използва по-обща формула като

$$\hat{e} = \arg\max_{e} \mathrm{P}_{\text{фрази}}(b \mid e)^{\alpha_1}.\mathrm{P}_{\text{думи}}(b \mid e)^{\alpha_2}.\mathrm{P}_{\text{фрази}}(e \mid b)^{\alpha_3}.\mathrm{P}_{\text{думи}}(e \mid b)^{\alpha_4}.\mathrm{P}(e)^{\alpha_5}.\mathrm{length}(e)^{\alpha_6}... \qquad (7)$$

където $\mathrm{P}_{\text{фрази}}(x \mid y)$ и $\mathrm{P}_{\text{думи}}(x \mid y)$ са съответно вероятности за превод с цели фрази и дума-по-дума, length($e$) е броят думи в английския превод (за да се избегне проблемът, че по-късите преводи по принцип имат по-голяма априорна вероятност), а $\alpha_i$ са тегла, които се избират така, че да се постигне максимално *BLEU* върху някакъв допълнителен набор паралелни изречения [*Och,2003*].

### *Превод с йерархични фрази*

Връщайки се към Таблица 4, можем да видим, че някои от фразите неявно кодират сложни синтактични трансформации. Така например двойката фрази

„*както физическа , така и психическа*” → „*both physical and psychological*”

би могла, за подходящи *X* и *Y*, да се обобщи в правилото

*както* X *, така и* Y → *both* X *and* Y

където X и Y са променливи, заместващи една или повече думи.

Разбира се, от таблицата могат да се извлекат много други правила, например:

*както и* X → *as well as* X

*както и* X → *in line with* X

*както и* X → *and also* X

*както и по* X → *and in* X

Тези наблюдения лежат в основата на *модела на Чанг* [*Chiang,2005*], който автоматично учи и използва за превеждане йерархични правила, съдържащи променливи. Правилата на Чанг съдържат един нетерминален символ отляво, и две последователности от нетерминали и терминали отдясно – по една за всеки език. Тези правила могат да се разделят най-общо на следните 4 вида:

- правила за превод на отделни думи:
  - X → *главен* || *main*
- правила за превод на цели фрази:
  - X → *главния* || *the main*
  - X → *главни прокурори* || *chief prosecutors*
- правила, смесващи терминални и нетерминални символи:
  - X → *както* X *, така и* Y || *both* X *and* Y
  - X → *както и* X || *as well as* X
- технически правила:
  - S → X || X
  - S → S X || S X

Вижда се, че йерархичният модел на Чанг е теоретично по-мощен от предходния модел с цели фрази поради правилата от последните две групи. Предимствата на модела се потвърждават и експериментално, а авторът му печели награда за най-добра статия на конференцията по компютърна лингвистика ACL [*Chiang,2005*]. Още примери и по-разширено описание са дадени по-късно в [*Chiang,2007*].

Следва да се отбележи обаче, че ефективността на модела на Чанг силно зависи от двойката езици. Той е много силен при превод между английски и китайски език, които се различават силно в своя словоред, но се представя по-лошо от модела с фрази при превод между двойки европейски езици, при които разликите в словореда са по-малки и предимно локални (напр. размяна на реда на прилагателните и съществителните). Моделът на Чанг не е подходящ и при превод между английски и арабски, където основният проблем не е словоредът, а богатата морфология на арабския език.

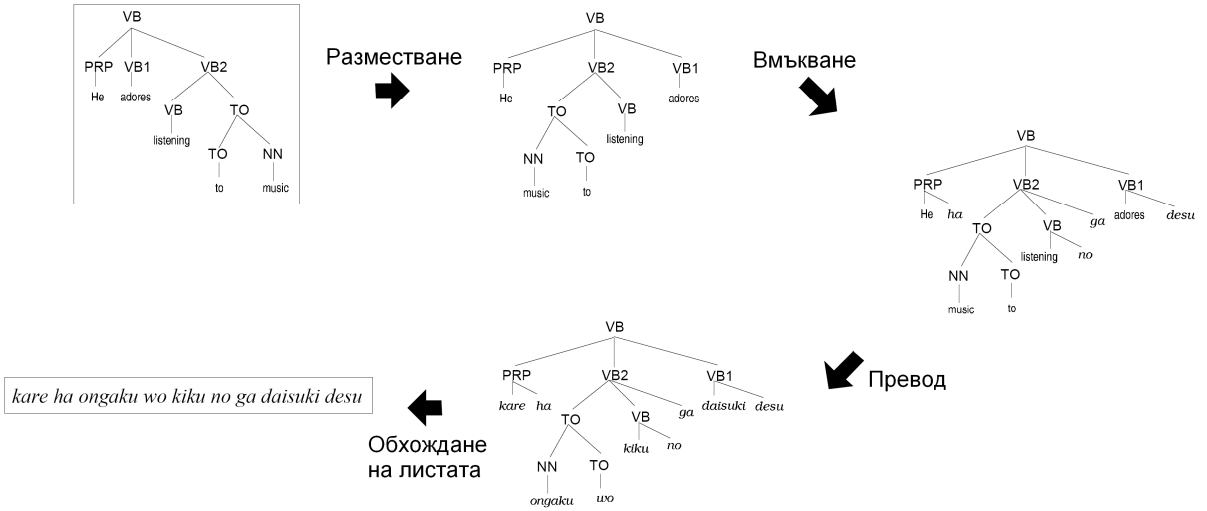
***Превод с използване на синтактична информация***

Съгласуването по род, число и падеж, както и някои особености на словореда създават сериозни проблеми на описаните по-горе модели, непознаващи понятия като *съществително*, *глагол*, *подлог* и др. Всъщност те не знаят дори какво е *дума*[10]: за тях няма разлика между истинска дума и препинателен знак. Затова следващата голяма цел на статистическия машинен превод е прякото моделиране на граматично знание. Задачата не е лесна, отчасти защото автоматичният синтактичен анализ е труден сам по себе си: най-добрите синтактични анализатори за английски език работят с 91% точност за качествен вестникарски текст, какъвто са обучени да анализират [*Charniak&Johnson,2005*], но качеството им пада значително при преминаване към друг домейн, например био-медицински. Въпреки това, най-добрите съвременни системи за синтактичен превод вече са изравнени по качество със системите за превод с цели/йерархични фрази, като в някои случаи са дори по-добри.

Един теоретично изчистен ранен пример за синтактичен превод е моделът на Ямада и Найт [*Yamada&Knight,2001*], който описва процеса на трансформация на английско синтактично дърво в японско на три стъпки: (1) разместване на листа и поддървета, (2) вмъкване на допълнителни възли, и (3) превод на английските листа на японски език дума по дума. Накрая японското изречение се прочита от листата на дървото. Конкретен пример е показан на Фигура 8. Моделът е изключително подходящ за превод от английски на японски език, при което се налага вмъкване на множество допълнителни думи: това става на строго определени синтактични позиции, което прави синтактичния анализ абсолютно необходим.

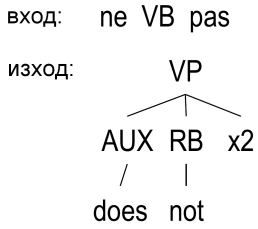
[10] При някои езици, напр. китайски, определянето на границите между думите е трудно поради липсата на писмени разделители и силната многозначност на различните комбинации от последователни йероглифи. Впрочем, понятието *дума* в китайски не е добре дефинирано и от лингвистична гледна точка. При някои европейски езици пък има нужда от допълнително сегментиране на сложни думи. Например, в италиански местоименията често се изписват „залепени" в края на глагола, напр. *compramelo = compra+me+lo* (‘*купи ми го*’), а в немски се сливат сложните съществителни като *Stammzell* (‘*стволова клетка*’; за сравнение, в английски език те се пишат разделно, напр. *stem cell*, освен при частична или пълна лексикализация, напр. *healthcare* и *Sunday*.)

Моделът на Ямада и Найт започва със синтактично дърво на изходния език (английски), което преобразува в изречение на езика, към който се превежда (японски), т.е. моделът преобразува синтактично дърво в последователност от думи.



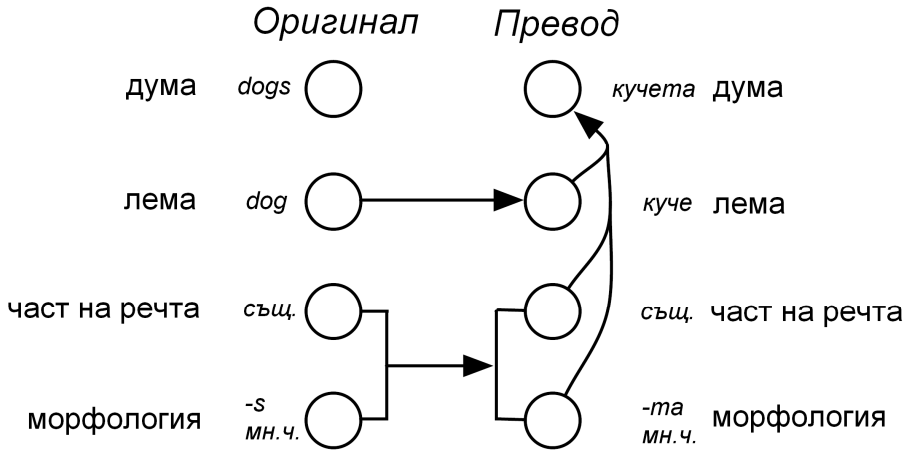

***Фигура 8. Превод от английски на японски език със синтактичния модел на Ямада & Найт.*** *Моделът описва процеса на трансформация на английско синтактично дърво в японско на три стъпки: (1) разместване на листа и поддървета, (2) вмъкване на допълнителни възли, и (3) превод на английските листа на японски дума по дума. Накрая японското изречение се прочита от листата на дървото.* [*Yamada&Knight,2001*]

Точно обратно работи моделът GHKM (по имената на авторите му Гали, Хопкинс, Найт и Марку) [*Galley&al,2004*], който по входна последователност от думи строи синтактични дървета за езика, към който се превежда. За целта GHKM използва правила като показаното на Фигура 9, с помощта на които се строят пълни синтактични английски дървета, които с много голяма вероятност съответстват на граматично правилни изречения. Тъй като за съвременния статистически машинен превод граматично неправилните изречения са много почест проблем от изреченията, при които смисълът е предаден неправилно, GHKM е един от най-успешните съвременни модели за статистически машинен превод.




***Фигура 9. Примерно правило за превод от френски на английски език с GHKM.*** *Правилото превежда стандартното френско отрицание с подходяща английска синтактична структура.*

Накрая, перспективна алтернатива на предходните два синтактични подхода са факторните модели [*Koehn&Hoang,2007*], които позволяват просто моделиране на морфологични и лексикални характеристики на ниво отделна дума. Фигура 10 показва примерен превод на английската дума *dogs* като *кучета*. Процесът включва три стъпки: анализ, транслиране и генериране. Първо, формата *dogs* се анализира като съществително *dog* в множествено число и съответно окончание *-s*. След това, лемата (основната форма), синтактичната и морфологичната информация се транслират поотделно в английските им еквиваленти. Накрая, те заедно генерират правилната българска форма *кучета*.




***Фигура 10. Примерен превод на* dogs → кучета *с използване на факторен модел.***

Основно предимство на факторните модели е, че позволяват отделен лингвистичен модел за всеки фактор, например на ниво морфология (за правилно съгласуване по род и число), на ниво част на речта (за граматично правилна последователност), на ниво лема (за семантично правилна последователност) и на ниво дума (за допълнително изглаждане). Тези лингвистични модели могат да се използват едновременно като отделни множители в обобщената формула (7), в резултат на което се получават граматично много по-правилни преводи. Това обаче е свързано със значително нарастване на пространството на търсене, а оттам и с общо забавяне на системата, което прави факторните модели приложими само при малки обеми учебни данни.

## Примерни преводи

Таблица 5 показва как се държи машинният превод с цели фрази при превод на административен текст от английски на български език. Моделът е обучен върху паралелен корпус от 111 документа: 45 883 двойки изречения, с 1 017 703 и 907 271 думи, съответно на английски и български език. Ще отбележим, че при сериозни индустриални системи обучението обикновено се извършва над 100 пъти по-голям обем паралелни двуезични текстове.

***Таблица 5. Примерни преводи от английски на български език:** с използване на статистически модел с цели фрази, обучен и приложен върху административен текст.*

| Английски оригинал | Човешки превод | Компютърен превод |
|---|---|---|
| European Convention on Mutual Assistance in Criminal Matters | Европейска конвенция за взаимопомощ по наказателно-правни въпроси | Европейска конвенция за взаимопомощ по наказателно-правни въпроси |
| Preamble | Преамбюл | Преамбюл |
| The governments signatory hereto, being members of the Council of Europe, | Правителствата, подписали тази конвенция, в качеството си на членове на Съвета на Европа, | Правителствата, подписали този протокол, членове на Съвета на Европа, |
| considering that the aim of the Council of Europe is to achieve greater unity among its members; | считайки, че целта на Съвета на Европа е да се постигне по-голямо единство между неговите членове, | считайки, че целта на Съвета на Европа е постигането на по-голямо единство между своите членове, |
| believing that the adoption of common rules in the field of mutual assistance in criminal matters will contribute to the attainment of this aim; | убедени, че приемането на общи правила в областта на правната помощ по наказателни дела ще допринесе за постигането на тази цел, | убедени, че приемане на общи правила в областта на правна помощ по наказателни дела ще допринесе за постигането на тази цел, |
| considering that such mutual assistance is related to the question of extradition, which has already formed the subject of a convention signed on 13th december 1957, | считайки, че правната помощ е свързана с въпроса за екстрадицията, която вече бе предмет на конвенцията, подписана на 13 декември 1957 година, | считайки, че тази взаимна помощ е свързана с въпроса за екстрадиция, който вече е образувано предмет на конвенция, подписана в 13th декември 1957 година, |
| have agreed as follows: | се споразумяха за следното: | се споразумяха за следното: |

Таблица 6 показва резултата от прилагането на същия модел за *адаптация*[11] от „македонски" към стандартен български език. Моделът е обучен върху паралелен корпус от вестникарски новини, съдържащ около 15 000 двойки изречения, с около 350 000 думи на български език. Това е около 300 пъти по-малко от нормалния обем текст, използван при сериозно обучение на индустриални системи.

***Таблица 6. Примерна адаптация от „македонски" към стандартен български език:*** *с използване на статистически модел с цели фрази, обучен и приложен върху вестникарски новини.*

| „Македонски" оригинал | Човешки превод | Компютърен превод |
|---|---|---|
| СКОПЈЕ, Македонија -- Според дипломатски извори, првата мировна мисија на ЕУ, која ќе биде распоредена во Македонија на 31-ви март, ќе го носи името Конкордија. | СКОПИЕ, Македония -- Първата мироопазваща мисия на ЕС, която ще бъде разположена в Македония на 31 март, ще се нарича "Конкордия", съобщиха дипломатически източници. | СКОПИЕ, Македония - Според дипломатически източници, първата мироопазваща мисия на ЕС, която ще бъде разположена в Македония на 31 март, ще носи името на Конкордия. |
| Околу 27 земји навестија подготвеност да учествуваат во шестмесечната мисија. | Тя ще продължи шест месеца. Около 27 страни са заявили желание да участват нея. | Около 27 страни намекнаха готовност да участват в шестмесечната мисия. |
| Се очекува македонскиот Парламент в среда (26-ти март) да го одобри нацрт договорот за статусот на силите на ЕУ, со кој на трупите на Унијата им се гарантира дипломатски статус и имунитет. | Очаква се македонският парламент да одобри проекта на Споразумението за статута на силите в сряда (26 март), гарантирайки по този начин дипломатически статут и имунитет на съюзническия контингент. | Очаква се македонският парламент в сряда (26 март) да одобри проект на споразумението за статута на силите на ЕС, с който на войниците на съюза да се гарантира дипломатически статут и имунитет. |
| Во вторникот, во меѓувреме, во Скопје започна меѓународна конференција за децентрализација, при што заменик-премиерот Муса Џафери го одржа воведниот говор. | Междувременно, във вторник в Скопие започна международна конференция по въпросите на децентрализацията. Тя бе открита с реч на вицепремиера Муса Джафери. | Междувременно, във вторник в Скопие започна международна конференция, посветена на децентрализацията и вицепремиерът Муса Джафери изнесена изказването. |
| Во другите вести, претседателот Борис Трајковски за Утрински весник потврди дека експертски тим на министерствата за внатрешни работи и за одбрана ја завршил својата работа на првата национална стратегија за безбедност и одбрана. | В други новини, президентът Борис Трайковски потвърди за Утрински весник, че екипът от експерти към министерствата на вътрешните работи и отбраната е завършил работата си по проекта за националната стратегия за сигурност и отбрана. | Други новини, президентът Борис Трайковски за Утрински вестник, че експертен екип на министерствата на вътрешните работи и отбраната е завършил работата си на първата национална стратегия за сигурност и отбрана. |
| Нацртот е поднесен до претседателскиот кабинет за ревизија. | Проектът е изпратен в президетството и кабинета за одобрение. | Нацртот бе представен на председателстващия кабинет за преразглеждане. |

## Използваем ли е машинният превод?

Макар и все още твърде далеч от качеството на професионалния човешки превод, днес машинният превод е вече използваем в ситуации, в които е достатъчно предаването на най-общия смисъл на текста, например при разглеждане на страница на чужд език в Интернет. Неслучайно функцията за автоматичен превод на *Google* е най-използваната сред предлаганите от компанията.

Автоматичният превод има какво да предложи и на професионалния преводач, например почти двукратно ускоряване на процеса на превод чрез подходящи подсказки в така наречените *преводачески памети* (у нас е най-популярна е *Trados*). За някои хора-преводачи редактирането на компютърен превод се оказва още по-добра алтернатива, тъй като спестява нуждата от ръчно набиране на по-голямата част от текста. В САЩ дори има компании, специализирали се в областта на редактирането на компютърен превод, които предлагат висока скорост и качество на конкурентна цена. Накрая, но не на последно място, машиният превод може да даде много добри резултати, ако оригиналният изказ се опрости. Така например, *Xerox* от години превежда цялата си техническа документация напълно автоматично, без човешка намеса, благодарение на опростения език (*Caterpillar English*), който използва при съставянето на английския оригинал на техническата си документация.

[11] Р. България официално счита *македонският јазик* за регионална писмена форма на българския. В тази връзка, Министерството на външните работи третира процеса на "превод" към стандартен български като *адаптация*.

## Заключение

С появата на статистическия подход през 1991 г. настъпва революция в областта на машинния превод, последвана от нова малка революция през 2003 г., когато е предложен ефективен модел за превеждане с цели фрази. Оставаме в очакване на следващата революция, която да постави синтаксиса, морфологията, и граматичното знание изобщо, на полагащото им се централно място в процеса на машинен превод.

## Литература


[*Bond,2005*] Francis Bond. *Translating the Untranslatable: A Solution to the Problem of Generating English Determiners*. CSLI Series in Computational Linguistics. 2005.

[*Brown&al.,1993*] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert. L. Mercer. *The mathematics of statistical machine translation: parameter estimation.* Computational Linguistics, 19(2), pp. 263-311, 1993.

[*Charniak&Johnson,2005*] Eugene Charniak, Mark Johnson. *Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking.* In Proceedings of ACL. pp. 173-180, 2005.

[*Chen&Goodman,1999*] Stanley F. Chen, Joshua Goodman, *An empirical study of smoothing techniques for language modeling*, Computer, Speech and Language, vol. 13, pp. 359-394, 1999.

[*Chiang,2005*] David Chiang. *A hierarchical phrase-based model for statistical machine translation*. In Proceedings of ACL, pp. 263–270, 2005.

[*Chiang,2007*] David Chiang. *Hierarchical phrase-based translation.* Computational Linguistics vol. 33, number 2, pp. 201–228, 2007.

[*Galley&al,2004*] Michel Galley, Mark Hopkins, Kevin Knight, Daniel Marcu. *What's in a translation rule?* In Proceedings of HLT-NAACL'2004. pp. 273-280, 2004.

[*Hutchins,2003*] John Hutchins. *ALPAC: the (in)famous report.* Readings in machine translation, ed. S.Nirenburg, H.Somers and Y.Wilks. pp. 131-135., The MIT Press, Cambridge, MA, 2003.

[*Koehn&al.,2003*] Philipp Koehn, Franz-Josef Och, Daniel Marcu. *Statistical phrase-based translation.* In Proceedings of HLT/NAACL, Edmonton, Canada, 2003.

[*Koehn&Hoang,2007*] Philipp Koehn and Hieu Hoang. *Factored Translation Models*. In Proceedings of EMNLP-CoNLL'2007, pp. 868-876, 2007.

[*Nagao,1984*] Makato Nagao. *A framework of a mechanical translation between Japanese and English by analogy principle.* In: A.Elithorn and R.Banerji (eds.) Artificial and human intelligence (Amsterdam: North-Holland), pp. 173-180, 1984.

[*Och,2003*] Franz-Josef Och. *Minimum error rate training in statistical machine translation*. In Proceedings of ACL, pp. 160-167, 2003.

[*Och&Ney,2003*] Franz-Josef Och, Hermann Ney. *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, vol. 29, number 1, pp. 19-51, 2003.

[*Papineni&al.,2002*] Kishore Papineni and Salim Roukos and Todd Ward and Wei-jing Zhu. *Bleu: a Method for Automatic Evaluation of Machine Translation*. In Proceedings of ACL, pp. 311–318, 2002.

[*Weaver,1955*] Warren Weaver. *Translation (1949).* In: Machine Translation of Languages, MIT Press, Cambridge, MA, 1955.

[*Yamada&Knight,2001*] Kenji Yamada, Kevin Knight. *A Syntax-based Statistical Translation Model*. In Proceedings of ACL, pp. 523-530, 2001.



*д-р Преслав Иванов Наков*
*Институт по паралелна обработка на информацията, БАН,*
*Секция по лингвистично моделиране*
*(02) 979-6607*
*nakov@lml.bas.bg*