صفدر رشید*

انگریزی سے اردو مبنی بر کارپس مشینی ترجمے کے مراحل: ایک توضیحی مطالعہ

Abstract:

Processes of Corpus-Based Machine Translation from English to Urdu: A Descriptive Study

This study aims to explain the processes of corpus-based Machine-Assisted Translation (MAT) from English to Urdu. It asserts that Translation Tools are of great significance for making the process of translation fast and systematic. It is a fact that the use of technology in the phenomenon of translation is of great importance in this global situation where the demand of translated texts has increased a lot. It is also a reality that human hands seem inadequate to fulfill the present-day requirement of this IT-based global world where inter-human communication is playing a pivotal role for various purposes. The study also focuses on the importance of translation as theoretically and conceptually driven phenomenon. This paper helps in understanding the processes involved in MAT especially those dealing with big translation projects where human translators have time constraints and accuracy issues.

Keywords: corpus-based, machine-assisted, translation, Urdu, English.

علمی اشتراک اور انسانوں کے مابین ابلاغ کے لیے عصرِ حاضر میں ترجمے کے کردار میں بہ تدریج اضافہ ہوا ہے۔ اس عالمی تناظر میں اس مقصد کے لیے مختلف شعبوں میں برق رفتاری سے سامنے آنے والے مواد کا ترجمہ کیے جانے کی ضرورت پہلے سے کہیں زیادہ ہے، لیکن اس مشکل کام کو بروقت سرانجام دینا انسانوں کے لیے بہت کٹھن ہے۔ ترجمے کی مسلسل بڑھتی ضروریات کو پورا کرنے کے لیے مشینوں کو اس قابل بنایا جاسکتا ہے اور استعمال میں لایا جاسکتا ہے۔ مشین اور مشینوں میں استعمال ہونے والے سافٹ ویئرز ماخذی زبان سے ہدفی زبان میں خود کار ترجمے کا عمل سرانجام دے سکتے ہیں۔ واضح

رہے کہ دنیا میں ترجمے کا زیادہ تر کام ایسے متنوں کا نہیں ہے جو ادبی یا ثقافتی حیثیت رکھتے ہوں۔ پیشہ ور مترجمین کی اکثریت سے سائنسی اور تکنیکی دستاویزات، تجارتی اور کاروباری سرگرمیوں، انتظامی نوعیت کی دستاویزات، قانونی دستاویزات، ہدایات ناموں، زرعی اور طبی نصابی کتب، صنعتی اختیارناموں، دستی اشتہارات، اخباری رپورٹوں وغیرہ کے تراجم کے لیے خدمات حاصل کی جاتی ہیں۔ بسا اوقات یہ کام جتنا دقت طلب ہوتا ہے اتنا ہی فوری اہمیت کا حامل ہوتا ہے۔علاوہ ازیں یہ کام بے کیف اور تکراری ہوتا ہے اور اس کے ساتھ ساتھ اس میں درستی اور توافق درکار ہوتا ہے۔ اس طرح کے ترجمے کی مانگ ترجمے کے پیشے کی روایتی گنجائش سے کہیں زیادہ بڑھ رہی ہے۔ اس ضمن میں کمپیوٹر کی مدد واضح اور فوری طور پر درکار ہے۔ کسی مشین ٹرانسلیشن سسٹم (Machine Translation System) کے عملی فائدے کا تعین آخرکار اس کے ماحصل (output) کے معیار سے کیا جاتا ہے۔ لیکن انسان یا مشین کا کیا گیا کون سا ترجمہ 'اچھا' ترجمہ سمجھا جاتا ہے اسے مختصراً بیان کرنا قدرے مشکل بات ہے۔ کیوں کہ اس کا زیادہ تر انحصار اُن مخصوص حالات پر ہے جن میں یہ کیا گیا ہے اور اُن قارئین/سامعین پر ہے جن کے لیے یہ کیا گیا ہے۔ اصل سے مماثلت، درستی، مفہومیت، مناسب اسلوب اور مخصوص لسانی شکل (register) وہ تمام معیارات ہیں جنھیں مدِنظر رکھا جاسکتا ہے، بایں ہمہ وہ ذاتی آرا کے زمرے میں ہی رہتے ہیں۔ ہمیں پاکستان کی ملکی لسانی ضروریات کو پورا کرنے کے لیے کمپیوٹرائی لسانیات، لسانی انجینئرنگ، مشینی ذہانت، نیچرل لینگوئج پراسسنگ، متوازی اور تقابلی کارپس (Parallel and Comparable Corpora) کے شعبوں میں اپنی صلاحیت میں اضافہ کرنے کی ضرورت ہے۔

بلاشبہ اکیسویں صدی ٹیکنالوجی کی صدی ہے اور تمام شعبہ ہائے زندگی میں انسانی ترقیاں ٹیکنالوجی کی ترقیوں پر منحصر ہیں۔ حالیہ ماضی میں ماہرینِ لسانیات نے بھی انفارمیشن ٹیکنالوجی سے بہت استفادہ کیا ہے۔ ترقی یافتہ اقوام پہلے ہی زبان اور ٹیکنالوجی، ترجمے اور ٹیکنالوجی، متوازی کارپس وغیرہ پر کام کا آغاز کر چکی ہیں۔ اب وقت کا تقاضا ہے کہ پاکستان اس سمت میں قدم بڑھائے اور اس مقصد کے لیے اعلیٰ قسم کے سسٹم تخلیق کرنے کے لیے ہماری قابلیت کو بہتر بنائے۔ ہمیں ملکی لسانی وسائل کے سامان کے لیے لسانی انجینئرنگ، مشینی ترجمے (machine translation) اور مشینی اعانتی ترجمے (Machine Assisted Translation) کے سافٹ ویئرز پر کام کو تیز کرنے کی ضرورت ہے۔ یہ ایک حقیقت ہے کہ مشینیں دستاویزات کا ترجمہ کرتے ہوئے سو فیصد درست ترجمہ مہیا نہیں کرسکتیں، خاص طور پر ہماری قومی زبانوں کی صورت میں، لیکن ہمیں اس مقصد کے لیے انفارمیشن ٹیکنالوجی کو استعمال کرنا چاہیے اور جہاں تک ہو سکے اس سے فائدہ اٹھانا چاہیے۔ ہم اعلیٰ قسم کی مشینیں استعمال کر کے فرہنگیں، لغات تیار کر سکتے ہیں ٹرانسلیشن میموریز (TM) بنا سکتے ہیں۔ ٹرانسلیشن میموریز سے مراد ایسا ڈیٹا بیس ہے جس

میں ترجمہ کرنے کے ساتھ ساتھ اضافہ ہوتا رہتا ہے۔ اس طرح اس میں پہلے سے ہی ماخذی زبان اور ہدفی زبان کے ترجمہ شدہ الفاظ، جملے، جملوں کے حصے بلکہ پیرا گراف یا ان کے حصے موجود ہوتے ہیں۔ اس طرح انسانی مترجم کوالفاظ کے متبادل ملنے میں خاصی آسانی ہو جاتی ہے۔

مشینی ترجمے سے مراد کمپیوٹر ٹیکنالوجی کو استعمال کرتے ہوئے انسانی مدد سے یا اس کے بغیر ایک زبان (ماخذی زبان) سے دوسری زبان (ہدفی زبان) میں ترجمہ کرنے کا عمل ہے۔ مختصراً مشینی ترجمہ تحقیق کا کوئی الگ شعبہ نہیں ہے۔ یہ لسانیات، کمپیوٹر سائنس، مشینی ذہانت کے شعبوں، نیز ترجمے کے نئے نظریات کی مدد سے بہتر مشینی سسٹم کو تیار کرنا ہے۔ لہٰذا یہ ایک اطلاقی تحقیق کا شعبہ ہے لیکن اس شعبے نے بہر حال کافی حد تک ایسی تکنیکیں اور تصورات پیش کیے ہیں جنھیں کمپیوٹر پر مبنی لینگوئج پراسسنگ جیسے دیگر شعبوں میں استعمال کیا جاسکتا ہے۔ ۱۹۵۰ء کے عشرے کے اوائل میں محققین نے خودکار ترجمے پر کام کا آغاز کر دیا تھا۔ مشینی ترجمے کو پہلی کامیابی اس وقت ملی جب ۱۹۵۴ء میں چند روسی جملوں کو انگریزی میں ترجمہ کر لیا گیا۔ اُس وقت یہ توقع ظاہر کی گئی کہ چند ہی برسوں کی بات ہے جب مکمل خودکار مشینی ترجمہ ایک حقیقت بن جائے گا۔ لہٰذا امریکہ میں اس پراجیکٹ پر کثیر رقم خرچ کی جانے لگی، مگر کوئی پیش رفت نہ ہوسکی۔ ۱۹۸۰ء میں خودکار مشینی ترجمے سے توقعات ختم کر کے مشینی ترجمے کے شماریاتی ماڈل پر توجہ دی جانے لگی اور اب تک یہی طریقہ کار مستعمل ہے، یعنی مشین ٹرانسلیشن کی بجائے CAT ۔

پاکستان میں سپریم کورٹ کے اردو کے نفاذ کے فیصلے[۱] کے بعد اب ترجمے کی اہمیت میں اضافہ ہوا ہے۔ کہا جاسکتا ہے کہ ترجمے کے بغیر قومی زبان کے نفاذ میں درپیش مسائل کا حل ممکن نہیں کیوں کہ ترجمہ ہی وہ واحد عمل ہے جس سے لسانی وسائل کو بڑھایا جا سکتا ہے اور دوسری زبانوں سے ہر شعبۂ زندگی کے علم کو قومی زبان میں لا کر اردو زبان کو جدید تقاضوں کے مطابق بنایا جا سکتا ہے۔ اس مقصد کے لیے ٹیکنالوجی سے مدد لینا ایک ناگزیر عمل ہے۔

ماہرین لسانیات کے مطابق اب انھی زبانوں کو بقا ملے گی جو ٹیکنالوجی کی زبانیں بن جائیں گی۔ مشینی ترجمہ کمپیوٹرائی لسانیات (computational linguistics) کا ذیلی مضمون ہے جس کے تحت تحقیق کی جا رہی ہے کہ کسی سافٹ وئیر کی مدد سے متن یا آواز کا ایک زبان سے دوسری زبان میں ترجمہ کیا جا سکے۔ مشینی ترجمے کا انتہائی بنیادی استعمال یہ ہے کہ ایک زبان کے لفظوں کو دوسری زبان کے لفظوں میں تبدیل کیا جائے، مگر اس طرح خود کار ترجمے کا مقصد پورا نہیں ہوسکتا۔ اس کا سبب یہ ہے کہ لفظ جزو جملہ(phrases) کے لحاظ سے اور جزو جملہ کسی جملے میں اپنے سیاق و سباق کے مطابق معنی پیدا کرتے ہیں اور یہ محض اسی ضمن میں نہیں بلکہ اسے جبر سے تعبیر کیا جا سکتا ہے کہ لفظ صرف اپنے سیاق میں ہی معنی خیز ہوتا ہے۔

کمپیوٹر کی مدد سے ترجمے کے دو طرح کے طریقہ کار موجود ہیں: ایک مشین ٹرانسلیشن (MT) اور دوسرا کمپیوٹر اسسٹڈ ٹرانسلیشن (CAT)۔ CAT سے مراد ہے کہ کمپیوٹر سافٹ ویئر کی مدد سے پہلے سے محفوظ شدہ ذولسانی ڈیٹا کو استعمال کرتے ہوئے مذکورہ سافٹ وئیر ترجمہ نگار کو مدد فراہم کرے، جب کہ MT اپنے اندر موجود ذولسانی ڈیٹا اور اس میں موجود گرامر کے قوانین کی مدد سے خود کارانہ ترجمے میں مدد دیتا ہے۔ کسی بھی CAT سافٹ ویئر (مثلاً Trados، Wordfast یا MemoQ) کا استعمال کنندہ اس کے مختلف پیرا میٹر طے کر سکتا ہے، مثلاً نیا جملہ پہلے سے موجود جملوں سے کس قدر فی صدی مطابقت (concordance) رکھے گا تو یہ سافٹ وئیر اسے خود کارانہ انداز میں ترجمہ کر کے پیش کر دے گا؛ یعنی ۴۰، ۵۰ یا ۶۰ فی صد؟ مثلاً پہلے سے ایک جملہ موجود ہے: ''Soon Ahsan will stop playing.'' اور دوسرا جملہ ''Soon baby will stop playing.'' ہے۔ دونوں جملوں میں مماثلت کے باعث ٹرانسلیشن میموری پہلے جملے کا ترجمہ سامنے لے آئے گی۔ اگر میموری میں پہلے جملے کا ترجمہ ''جلد ہی احسن کھیلنا بند کر دے گا'' موجود ہے تو دوسرا جملہ لکھنے پر پہلے جملے کا ترجمہ از خود سکرین پر ظاہر ہو جائے گا۔ جیسا کہ اس مثال میں آٹھ میں سے چھ الفاظ دونوں جملوں میں مشترک ہیں۔

یہاں یہ بات بھی ذہن میں رہے کہ CAT سافٹ ویئر وقت کے ساتھ ساتھ ترجمہ نگار کے مزاج سے ہم آہنگی پیدا کرتا ہے، جس کی وجہ سے ترجمے کا ایک معیاری اسلوب بن رہا ہوتا ہے۔ ایک انسان ہمیشہ ایک کام کو ہر مرتبہ کم و بیش مختلف انداز میں کرتا ہے جس کی وجہ سے یکسانیت نہیں آ سکتی۔ CAT میں ترجمے میں یکسانیت یا معیار بندی یا ایک ہی اسلوب کو قائم کیا جا سکتا ہے۔

مشینی ترجمے کی بہت سی اقسام ہیں، مگر عملاً یہ تمام دو بنیادی اقسام سے نکلی ہیں یعنی Rule Based MT یا مبنی بر قواعد مشینی ترجمہ، اور Corpus Based MT یا Example Based MT۔ RBMT ماخذ اور ہدفی زبان کے تجزیے پر مبنی ہوتی ہے۔ دوسری طرف CBMT کارپس میں ذخیرہ شدہ ذولسانی متن کے تجزیے پر مبنی ہوتی ہے۔ اسی طرح کارپس پر مبنی MT کو دو بڑی اقسام میں تقسیم کیا جا سکتا ہے:

> … Memory Based MT, i.e. memory heavy, linguistic light and Example-Based MT i.e. memory light and linguistic heavy. While the former systems implement an austere theory of meaning, the latter make use of rich representations[2].

جوں جوں ٹرانسلیشن میموری میں اضافہ ہوتا جاتا ہے CBMT وقت کے ساتھ ساتھ خود بخود بہتر ہوتی جاتی ہے۔

ٹرانسلیشن میموری کی فیصدی ٹیگنگ (concordance) کی جا سکتی ہے کہ جب ہدفی زبان ، اردو، کے کسی فقرے یا جملے کے ترجمے کا خاص تناسب، مثلاً ۳۰ فیصد، مماثل ہو جائے تو وہ خود کار طریقے سے مماثلت ظاہر کر دے۔ اس طرح مترجم کے پاس اختیار ہو گا کہ وہ جس قدر چاہے حصہ لے لے۔ اس سے وقت کی بھی بچت ہے۔ دراصل یہ اس مفروضے پر قائم کیا گیا ہے کہ ایک زبان کے ہر لفظ کا متبادل لفظ دوسری زبان میں ممکنہ طور پر موجود ہے۔ اسی سے ملتی جلتی Example Based Machine Translation ہے جو متوازی متن (parallel text) کے ساتھ ذولسانی کارپس استعمال کرتی ہے جو پہلے سے موجود مثالوں کی بنیاد پر کام کرتی ہے۔

مشینی ترجمہ کے ان تمام طریقہ ہائے کار کے متوازی جس تکنیک کی طرف ماہرین کی توجہ مرکوز ہے اُسے deep learning کہا جا رہا ہے۔ اس تکنیک میں بجائے اس کے کہ کمپیوٹر کو بہت سا ترجمہ شدہ مواد فراہم کیا جائے اور پھر concordance کی بنیاد پر مترجم ترجمے کے رد، قبول یا اس میں تدوین کا فیصلہ کرے، کمپیوٹر کو ماخذ اور ہدفی زبان کے قواعد (جس میں صرف ونحو کے ساتھ ساتھ روزمرہ بھی شامل ہوتا ہے) فراہم کر کے، جملے کی ساخت کے اصول بھی فراہم کر دیے جائیں اور مشین کو جو کہ درحقیقت انتہائی تیز رفتار اور اچھی صلاحیت کا حامل کمپیوٹر ہوتا ہے، ان بتائے ہوئے اصول وقوانین کو سمجھنے کے لیے کہا جاتا ہے۔ اس تفہیم میں اگرچہ بہت سا وقت لگ سکتا ہے جس کا انحصار کمپیوٹر کی صلاحیت پر ہوتا ہے لیکن اس کے بعد اخذ شدہ نتائج (ڈیٹا) کو کسی بھی کم تر درجے کے کمپیوٹر کو دے کر بہت بہتر اور تیزی سے نتائج حاصل کیے جا سکتے ہیں۔ اگر ''احسن کیسا ہے؟'' اور ''کیا وہ بیمار ہے؟'' جیسے جملے جو قواعدی اعتبار سے استفہامیہ ہونے کے ساتھ ساتھ باہم جڑے ہوئے ہیں، کسی ایسے سافٹ وئیر کو دیے جائیں جس کی ترجمہ کاری کی تکنیک مبنی بر ٹرانسلیشن میموری ہو تو نتائج کچھ یوں ہو سکتے ہیں'' How is Ahsan?'' لیکن دوسرے جملے کا ترجمہ کرتے وقت اس تکنیک میں مسائل سامنے آ سکتے ہیں کہ ''Is he ill?'' یا '' Is she ill?'' دونوں میں سے کون سا ترجمہ درست ہے۔ لہٰذا ہر بار مترجم کو مداخلت کر کے درستی کرنی پڑے گی۔ لیکن اس کے برعکس deep learning میں مشین کو اس بات کی تفہیم کرا دی جاتی ہے کہ تذکیر وتانیث کے مسائل سے کس طرح نمٹا جائے۔

فلسفیانہ سطح پر ہم کہہ سکتے ہیں کہ مشین کے ذریعے ترجمے میں سہولت کی بنیاد اس مفروضے پر ہے کہ ایک زبان کے لفظ کا دوسری زبان میں متبادل موجود ہے اور ترجمے کے لیے مشین اپنی ٹرانسلیشن میموری میں دستیاب ہدفی زبان کے پہلے سے موجود تراجم اور ذولسانی لغات پر کرے گی۔ گویا اس میں یہ مفروضہ بھی شامل ہے کہ پہلے سے دستیاب تراجم درست ہیں۔ تاہم مترجم کے پاس تدوین کا اختیار ضرور ہوتا ہے۔ اس طرح اسے interactive سرگرمی کہا جا سکتا ہے۔

۱۹۶۰ء سے ۱۹۸۰ء تک مشینی ترجمے کے حوالے سے بہت زیادہ تجربات کیے گئے مگر کمپیوٹر کی مدد سے ترجمے کے اصول توقعات کے مطابق طے نہیں ہو سکے۔ یہاں تک کہ کچھ موقعوں پر یہ کہہ دیا گیا کہ مشینی ترجمہ شاید ممکن ہی نہیں ہے۔ بعد ازاں امریکہ، جاپان اور دیگر ممالک کے درمیان صنعتی اور معاشی اشتراک و تعاون کی وجہ سے خود کار ترجمے پر سرمایہ کاری کی گئی، مگر نتیجہ حوصلہ افزا نہیں ملا۔ پھر آئی ٹی کی دنیا کے گوگل اور مائکروسافٹ جیسے بڑے اداروں کے ساتھ ساتھ چھوٹی فرمیں بھی اس میدان میں کود پڑیں۔ اس تمام تر کاوش و کوشش کے نتیجے میں MT کا بہتر سافٹ ویئر شائد ابھی بھی خواب ہے۔ مگر CAT کی صورت میں ایک بہتر اور قابل عمل چیز انسان کے ہاتھ آئی۔ MT میں ٹرانسلیشن میموری اور قواعد دونوں کی موجودگی شرط اوّل ہے۔ جبکہ CAT میں صرف بنیادی قواعد اور لغات کی سطح تک ذولسانی الفاظ کی موجودگی ضروری ہے، جب کہ اس کی ٹرانسلیشن میموری اس کے استعمال کے ساتھ ساتھ بہتر ہوتی جاتی ہے۔ CAT میں سب سے اہم چیز Percentage of Matching کے قواعد ہیں۔ انسانی مترجم کا زیادہ کردار ہونے کے باعث CAT کو Customised Semi Automated Machine Translation کہا جا سکتا ہے۔

جہاں تک کارپس کے ابتدائی مراحل کی تیاری کا تعلق ہے، اس کے لیے ضروری ہے کہ کسی شعبے کے ذخیرے کی نشاندہی کی جائے اور جیسا کہ یہ بات ہم سب کے سامنے ہے کہ اردو زبان کا مواد ڈیجیٹل صورت میں کم دست یاب ہے اسی طرح اردو کے متن کو Optical Character Recognizer یعنی OCR کے ذریعے digitize کرنے میں ابھی بڑی پیش رفت نہیں ہوئی۔ پاکستان میں CRULP اور CLE جیسے ادارے[۳] کچھ عرصے سے تحقیق کر رہے ہیں اور انھیں جزوی کامیابی بھی ملی ہے۔ اردو OCR کا بننا اردو زبان اور مشین ٹرانسلیشن کے لیے ایک انقلاب سے کم نہ ہوگا۔ اس طرح اردو کے ابتدائی مخطوطات اور کتابیں Machine Readable Format میں دستیاب ہو سکیں گی۔

اگر ہمارا کارپس انگریزی اور اردو دو زبانوں پر مشتمل ہے تو اس میں سب سے پہلے انگریزی زبان کے مواد کو digitize کرنے اور اس سلسلے میں OCR سافٹ وئیر اور ٹائپنگ کے طریقہ کار کو اپنایا جائے گا۔ یہاں اس بات کی بھی ضرورت ہے کہ کارپس کے ذخیرے کے لیے بنیادی کلیدی الفاظ اور میموری یونٹ کی بھی تعریف کی ضرورت ہے تا کہ جب مذکورہ انگریزی مواد کا اردو میں ترجمہ ہو تو آسانی کے ساتھ ہو۔ صرف مواد داخل کرنا مسئلہ نہیں کمپیوٹر tools کی مدد سے اسے کسی خاص ترتیب میں لائے بغیر آپ ذولسانی کارپس کے مقاصد کے حصول میں کامیاب نہیں ہو سکتے۔ پہلے مرحلے میں ماخذ زبان مثلاً انگریزی ڈیجیٹل صورت میں منتقل ہو چکی ہو تو ایسے مددگار tools موجود ہیں جس میں ذولسانی کارپس کا مرحلہ طے

کرنے میں آسانی ہوسکتی ہے ۔

دوسرے مرحلے میں TRADOS یا Omega T ترجمے کے عمل کو سہل بناتے ہیں ۔ مگر اسی دوسرے مرحلے میں جانے سے بیشتر متعلقہ لوگوں کی تربیت کا انتظام نہایت ضروری ہے ۔ بظاہر یہ صرف دو مرحلے بیان ہوئے ہیں جس میں انگریزی زبان کے مواد کا digitize کرنا اور بعد ازاں ذخیرہ شدہ مواد کو اردو میں ترجمہ کرنا شامل ہیں ، مگر اس کے لیے زبان کے ماہرین اور کمپیوٹر سائنس کے ماہرین کا اکٹھ اور فزیکل لیب کا قیام وہ ذیلی مراحل ہیں جو انھی مراحل میں شامل ہیں ۔ عملی طور پر یہ سب کچھ کرنے کے لیے systematic طریقہ کار اختیار کرنا پڑتا ہے، تربیت یافتہ ٹیم اور ان کے درمیان ربط پیدا کرنے کے بعد کوئی ادارہ اس قابل ہوسکتا ہے جو یہ کام سر انجام دے سکے ۔ ایک چھوٹی مگر اپنی نوعیت میں بڑی مثال سے وضاحت کی جا سکتی ہے کہ کارپس کا صرف ایک استعمال مشینی ترجمے میں بطور Core Engine Data کا استعمال ہے یعنی کسی مشینی ترجمہ کار کی تربیت کا انحصار کارپس میں موجود ڈیٹا پر ہے ۔ مشہور زمانہ گوگل ٹرانسلیٹر بھی اسی بنیاد پر قائم ہے ۔ اگرچہ عملی طور پر یہ بھی ایک عام اطلاقیہ ہے مگر پیشہ ورانہ انداز میں اور گوگل ادارے کی چھتری تلے ایک بڑے کام کی طرف پیش رفت ضرور ہے۔ یہ ٹرانسلیٹر عمومی استعمال کے لیے تو قدرے ٹھیک ہے مگر کسی مخصوص شعبے کے لیے غیر موزوں ۔

تھیوری سے قطع نظر حقیقت یہ ہے کہ مخصوص شعبوں، خصوصاً سائنسی و تکنیکی شعبوں کے کارپس پر مبنی CAT ہی بہتر نتائج دے رہے ہیں۔ اردو دنیا میں اس حوالے سے کام چوں کہ نہ ہونے کے برابر ہے، لہٰذا آغاز اسی سے ہونا چاہیے۔

> درحقیقت اپنی ضرورت کے مطابق بنائی گئی بنیادی لغت، اور سب سے بڑھ کر انسانی مترجم کی طرف سے تیار کردہ ٹرانسلیشن میموری کے استعمال کی بہ دولت، مخصوص شعبوں (مثلاً قانون، کمپیوٹر سائنس، طب وغیرہ) میں پیغام کا مشینی ترجمہ ہو جائے گا[۴]۔

اس کی ایک بہتر مثال جامعہ گجرات کے مرکزِ السنۂ وعلومِ ترجمہ میں ہونے والے قانون کے ترجمے کی ہے۔ لفظ اور معنی کا رشتہ من مانا یا طے شدہ نہیں بلکہ تفہیم کا تعلق سیاق و سباق کے علاوہ دیگر عناصر سے بھی ہے۔ دوسری طرف مشینی ترجمے میں لفظ اور معنی کا رشتہ طے شدہ ہوگا تو بات بنے گی، اور ایسا سائنسی اور تکنیکی نوعیت کے تراجم میں ممکن ہے، ادبی تراجم میں نہیں۔

علاوہ ازیں ترجمے کے بڑے منصوبوں کو قلیل مدت میں مذکورہ سافٹ ویئر کی مدد کے بغیر مکمل کرنا ممکن ہی نہیں، سافٹ وئیر کے کلیدی الفاظ کی فلٹریشن اور ان کے معنی کے تعین کی وجہ ترجمہ میں یکسانیت کا جزو اہمیت اختیار کر جاتا ہے۔ CAT سافٹ وئیر ترجمہ کرنے والے کی مدد بالکل اسی انداز میں کرتا ہے جیسا کہ Word Processor ، ہجوں، بنیادی گرامر،

رموز اوقاف اور formatting میں ہمارے لیے آسانیاں پیدا کرتا ہے۔ CAT کے سافٹ وئیر میں ترجمہ نگار ہی بہ راہ راست ترجمہ کر رہا ہوتا ہے جس کی وجہ سے اس کا معیار ایک مشینی ترجمہ کار سے بہتر ہوتا ہے ۔

مشینی ترجمے میں گوگل کی آن لائن اور مائکروسافٹ کی بنگ (Bing) ٹرانسلیشن سروسز مفت دستیاب ہیں ۔ یہ عمومی اور سادہ جملوں کے لیے قدرے کارآمد ہیں مگر تخصیصی ترجمے کے لیے بالکل موزوں نہیں۔

یہاں انگریزی کے ایک کارآمد ڈیٹا بیس کا تعارف بھی برمحل ہے۔WordNet انگریزی زبان کا لغوی ڈیٹا بیس ہے۔ یہ ہم معنی الفاظ کو ایک گروپ کی صورت میں اکٹھا کرتا ہے، جنہیں Synsets کہا گیا ہے۔ یہ ان الفاظ کی مختصر تعریف اور مثالیں پیش کرتا ہے۔ اسے لغت اور تھیسارس کا امتزاج کہا جا سکتا ہے۔ چونکہ یہ ویب پر دستیاب ہے اس لیے اسے آٹو میٹک متنی تجزیے کے لیے استعمال میں لایا جا سکتا ہے۔ ڈیٹا بیس اور سافٹ وئیر مفت ڈاون لوڈ بھی کیے جا سکتے ہیں۔ تمام Synsets ایک دوسرے سے معنوی طور پر ربط میں ہیں۔ اپنے سیاق میں معنی بتانے کے لیے اس سے بہتر شاید کوئی اور آن لائن بندوبست نہیں۔ اب WordNet کا ایک پورا سلسلہ وجود میں آ چکا ہے، جس میں یورپ کی کم و بیش تمام بڑی زبانیں موجود ہیں۔ اگر اردو دنیا نے اس ضمن میں خود کوئی پیش رفت نہ کی تو ایک نہ ایک دن اہلِ مغرب ہی اردو کا WordNet تیار کر دیں گے۔

۲۰۰۴ء میں مائیکرو سافٹ اور مقتدرہ قومی زبان کے مابین ایک معاہدہ ہوا[۵] جس کے تحت مائیکروسافٹ آفس اور ونڈوز میں موجود تمام مینیو، ہدایات اور کمپیوٹر سکرین کو اردو میں تبدیل کرنا تھا۔ یہ مواد قریباً نو لاکھ الفاظ پر مشتمل تھا۔ یہاں سب سے بڑا چیلنج ہم معنی الفاظ اور اصطلاحات کے لیے الگ الگ اردو مترادفات لانا تھا۔ ترجمے کا سافٹ وئیر طے شدہ لفظ کے علاوہ دوسرا لفظ قبول ہی نہ کرتا تھا۔ مثلاً choose کے لیے 'انتخاب کریں' اور select کے لیے 'منتخب کریں' طے کیا گیا۔ بعض اوقات کسی اصطلاح کا ترجمہ غیر مانوس معلوم ہوتا ہے، مگر مجبوری وہی کہ ہر اصطلاح کے لیے منفرد لفظ ہو، مثلاً اردو میں history کے لیے 'تاریخ' رائج ہے اور date کے لیے بھی 'تاریخ'۔ عام طور پر تو ہم سیاق و سباق سے اندازہ لگا لیتے ہیں کہ کون سی تاریخ مراد ہے، مگر بطور ایک علیحدہ تصور کے یہاں history کے لیے لفظ 'سابقات' لانا پڑا۔ جملہ معترضہ کے طور پر عرض ہے کہ اصطلاح کے معاملے میں غیر مانوس لفظ سے گھبرانے کی ضرورت نہیں ہوتی۔ جب ایک لفظ ایک تصور کے لیے مختص ہو گیا تو اس پر اتفاق ہو جاتا ہے، نتیجتاً وقت کے ساتھ ساتھ اس سے بے گانگی دور ہو جاتی ہے۔ یہ بات تو ہم سب ہی جانتے ہیں کہ الفاظ اور ان کے معنی ایک من مانے رشتے میں نہیں بندھے ہوتے، معاشرے کا بس ان پر اتفاق ہو جاتا ہے۔

ذیل میں کسی بھی تخصیصی شعبے میں کمپیوٹر سافٹ وئیر کی مدد سے ترجمہ کرنے کے مراحل کا خاکہ پیش کیا جاتا ہے۔ہم

فرض کر رہے ہیں کہ ماخذی زبان انگریزی ہے اور ہدفی زبان اردو۔ دوسرا، ترجمے کے لیے CATاپروچ کے تحت TRADOS سافٹ وئیر کا استعمال کیا جا رہا ہے۔ مراحل حسب ذیل ہیں:

۱۔ سب سے پہلا کام متن کی قسم کا تعین ہے، مثلاً مالیات، کمپیوٹر سائنس، بیالوجی، قانون، زراعت، وغیرہ۔

۲۔ متن کی قسم کے تعین کے بعد، ماخذی زبان، انگریزی، سے ایک بہت بڑا نمائندہ متنی انتخاب، جو کئی ملین الفاظ پر مشتمل ہو۔ یہ انتخاب اسی شعبے کے ماہرین کی کمیٹی کرے گی۔ اس طرح اس مخصوص شعبہ علم کے الفاظ کا بہت بڑا ذخیرہ وجود میں آ جائے گا۔

۳۔ منتخب متن کی ممکنہ تین صورتیں ہوں گی۔ متن کی فیڈنگ یا کمپوزنگ ہوگی۔ دوسری صورت میں متن کو بذریعہ OCR ٹیکسٹ فارمیٹ میں لانا ہوگا۔ تیسری صورت میں متن پہلے ہی کمپوزڈ صورت میں دستیاب ہے۔

۴۔ تمام مواد تکنیکی طور پر صرف سادہ متن (text format) میں ہونا چاہیے۔ اس طرح سارا متن editable format میں موجود ہوگا۔

۵۔ کسی اچھے سافٹ وئیر کو استعمال کرتے ہوئے سارے متن سے کلیدی الفاظ یا keywords نکال لیے جائیں گے، مثلاً دس ملین الفاظ کے ڈیٹا سے دس ہزار کلیدی الفاظ اکٹھے ہو گئے۔

۶۔ تمام کلیدی الفاظ ماہرینِ شعبہ کی کمیٹی کو جائزے کے لیے پیش کیے جائیں گے۔

۷۔ حتمی کلیدی الفاظ اور اصطلاحات کے ممکنہ معنی اس شعبے اور زبان کے ماہرین کی کمیٹی کے حوالے کیے جائیں گے۔

۸۔ غور وخوض کے بعد کلیدی الفاظ اور اصطلاحات کے ترجمے پر اتفاق رائے حاصل کیا جائے گا۔

۹۔ اس طرح وجود میں آنے والی فرہنگ کو کارپس کی ٹرانسلیشن میموری کا حصہ بنا دیا جائے گا۔

۱۰۔ TRADOS میں مترجم کی معاونت کے لیے چند معیاری عمومی ذولسانی لغات بھی رکھی جانی چاہیں۔

۱۱۔ TRADOS جملے کو مختلف فقروں (phrases) میں تقسیم کر کے سامنے لائے گا۔

۱۱۔ ترجمے کے حجم میں اضافے کے ساتھ ساتھ ٹرانسلیشن میموری میں اضافہ ہوتا جائے گا، جس سے نہ صرف معیار بندی ممکن ہوگی بلکہ وقت کی بھی بچت ہوگی۔

۱۲۔ انگریزی اور اردو کے اجزائے کلام (POS) کی ترتیب میں فرق کے باعث ترجمہ شدہ متن بے ربط ہوگا، جس کی اصلاح درکار ہوگی۔

۱۳۔ اس ترجمے کی پڑتال کے لیے کئی سطحوں پر ماہرین کی جائزہ کمیٹیاں تشکیل دی جائیں گی، مثلاً لسانیات، متعلقہ شعبہ علم کے ماہرین، مترجمین وغیرہ جو مختلف پہلوؤں سے ترجمے کا جائزہ لیں گے۔

درج بالا بحث سے یہ نتیجہ اخذ کیا جا سکتا ہے کہ پاکستانی سیاق میں مشینی ترجمے کی اہمیت میں بہت حد تک اضافہ ہوا ہے۔ اردو زبان کے فروغ میں مشینی ترجمے کی اہمیت سے انکار کرنا اب ممکن نہیں رہا۔ ضرورت اس امر کی ہے کہ اردو کو جدید ٹیکنالوجی سے جوڑا جائے تا کہ آنے والے اوقات میں انھیں مزید فروغ مل سکے۔ اس مقصد کے لیے درج بالا تجربات سے استفادہ کیا جا سکتا ہے۔ اس میں پہلے مرحلے میں اردو کے لیے او سی آر (Optical Character Recognizer) درکار ہے جس پر ابھی کافی کام ہونا باقی ہے۔ اس کے ساتھ ساتھ ہمیں کارپس لسانیات کے علم سے استفادہ کرنا ہوگا۔

حواشی و حوالہ جات

* (پ: ۱۹۷۲ء) اسسٹنٹ پروفیسر، شعبۂ اردو، علامہ اقبال اوپن یونیورسٹی، اسلام آباد۔

۱۔ سپریم کورٹ آف پاکستان، فیصلہ برائے آئینی درخواست نمبر ۲۰۰/۵۶ از کوکب اقبال، ۸ ستمبر ۲۰۱۵ء۔

۲۔ مائیکل کارل (Michael Carl)، A Model of Competence for Corpus-Based Machine Translation، مشمولہ: *COLING '00: Proceedings of the 18th conference on Computational linguistics* (سٹراوڈس برگ: Association for Computational Linguistics Stroudsburg، ۲۰۰۰ء)، ۹۹۷-۱۰۰۱۔

۳۔ CRULP (Center for Research in Urdu Language Processing) کا شمار پاکستان میں زبان اور ٹیکنالوجی کے حوالے سے تحقیق کرنے والے اولین اداروں میں ہوتا ہے۔ یہ ادارہ ڈاکٹر سرمد حسین نے FAST یونیورسٹی کے لاہور کیمپس میں قائم کیا۔ اسی ادارے کی بنیاد پر بعد ازاں ۲۰۱۱ء میں ڈاکٹر سرمد حسین نے ہی یونیورسٹی آف انجینئرنگ اینڈ ٹیکنالوجی، لاہور میں CLE (Center for Language Engineering) قائم کیا۔ یہ ادارہ زبان کے لسانیاتی اور کمپیوٹیشنل پہلوؤں پر تحقیق کا مقصد رکھتا ہے۔

۴۔ میتھیو گاڈئیر (Mathieu Guidère)، Toward Corpus-Based Machine Translation for Standard Arabic، مشمولہ: *Translation Journal* (جلد ۶/۱، ۲۰۰۲ء)۔

> The task can be somewhat simplified by using specialized and regular fields of application as reference corpora. In fact, in specialized fields (legal, computer science, medical etc.) the message will be machine translated essentially by using a customized basic dictionary and, above all, a translation memory created by the human translator.

۵۔ عطش درانی، ''اردو اطلاعیات اور فرہنگ''، مشمولہ: برقیاتی فرہنگ برائے کمپیوٹر (انگریزی، اردو) (اسلام آباد: مقتدرہ قومی زبان، ۲۰۰۵ء)۔

مآخذ:

ایس ڈی ایل ٹریڈوس۔ SDL Trados Studio: Translation Tool۔ https://www.sdltrados.com/۔

اومیگا ٹی۔ *OmegaT: The free translation memory tool.*۔ https://omegat.org/۔

پائم، اے [Pyme, A.]۔ *Exploring Translation Theories*۔ لندن/ نیو یارک: روتلج، ۲۰۱۴ء۔

ٹرُوجِلو، اے [Trujillo, A.]۔ *Translation Engines: Techniques for Machine Translation*۔ نیو یارک: سپرنگر، ۱۹۹۹ء۔

درانی، عطش، ''اردو اطلاعیات اور فرہنگ''، مشمولہ: برقیاتی فرہنگ برائے کمپیوٹر (انگریزی، اردو) (اسلام آباد: مقتدرہ قومی زبان، ۲۰۰۵ء)۔

ڈور، بی جے [Dorr, B. J.]۔ *Machine Translation: A View from the Lexicon*۔ کیمبرج: The MIT Press، ۱۹۹۳ء۔

سپریم کورٹ آف پاکستان، فیصلہ برائے آئینی درخواست نمبر ۵۶/۲۰۰ از کوکب اقبال، ۸ ستمبر ۲۰۱۵ء۔

کارل، مائیکل [Carl, Michael]، A Model of Competence for Corpus-Based Machine Translation، مشمولہ: *COLING '00: Proceedings of the 18th conference on Computational linguistics* (سٹراوُڈس برگ: Association for Computational Linguistics Stroudsburg، ۲۰۰۰ء)، ۹۹۷-۱۰۰۱۔

گاڈئیر، میتھیو [Guidère, Mathieu]، Toward Corpus-Based Machine Translation for Standard Arabic، مشمولہ: *Translation Journal* (جلد ۶/۱، ۲۰۰۲ء)۔

گوٹ، سی [Goutte, C.]/ این کینسیڈا [N. Cancedda]/ ایم ایف ڈامیٹ مین [M. F. Dymetman]۔ *Learning Machine Translation*۔ کیمبرج: The MIT Press، ۲۰۰۹ء۔

ہاسر، آر [Hausser, R.]۔ *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*۔ نیو یارک: سپرنگر، ۲۰۱۴ء۔

ورڈ نیٹ۔ *WordNet: A Lexical Database for English*۔ https://wordnet.princeton.edu/۔

وِلکس، وائے [Wilks, Y.]۔ *Machine Translation: Its scope and limits*۔ نیو یارک: سپرنگر، ۱۹۹۹ء۔