اردو کارپس: تکنیکی تعارف، اہمیت، ضرورت اور دائرہ و لائحۂ عمل

حافظ صفوان محمد چوہان*

Abstract:

This article emphasizes the need of Urdu corpus on the example of the Bank of English and the Corpus of Contemporary American English (COCA) which are serving as the backbone of English language engineering, discourse analysis, corpus & lexicon development and works of the same fiber. This proposed Urdu corpus, namely The Bank of Urdu (TBU), will be a repository of Urdu texts of both written and spoken language gathered in platform-independent & machine-readable Indo-Perso-Arabic script. Since the mentioned English corpora have exactly the same architecture and interface so while comparing the TBU with the structure of English corpora, the name "English Corpus" will refer to both these repositories in this document.

Add to devising its scope, technical and design issues of the architecture & interface of TBU are discussed in this introductory paper. Issues like those of code-mixing, false friends and homonyms in Urdu are addressed. Together, solution is given to standardize the Urdu orthograph for this work. Exemplary web view of the user interface is provided. Available Urdu written texts are mostly literature-oriented, so from the data gathering standpoint the proposed TBU must deviate from standard roadways of the English corpora at many instances. This fact is specially dealt with. A study of word-count and of lexicalizing high-frequency Urdu words in Urdu dictionaries of note is made part of this thesis. Aimed at discourse analysis, language engineering and natural language processing in Urdu, and of course, providing vital base for contemporary Urdu lexicon development, this proposed portal will not only separate Urdu language from Urdu literature but will also cast regional Pakistani languages in stationing their scholarly resources in their own scripts for such researches.

This paper on TBU is a proposal of Dr Hafiz Safwan Muhammad Chohan for giving initial shape to the idea of Urdu Data Bank (UDB) of the Center of Excellence for Urdu Informatics (CEUI), National Language Authority (NLA) Islamabad. Due to homonymy of UDB with the Urdu Data Base, UDB was renamed as TBU at the CEUI in a consensus with the scholars of Urdu, IT professionals and representatives of the GoP from Cabinet Division & Planning Division. In this national workshop viz. "Urdu Informatics- Today & Tomorrow" held on 7-8 June 2008 in the NLA, Dr Chohan also coined the Urdu equivalent of TBU as اردو مثال گھر which was accepted by the participants.

* ٹیلی کمیونیکیشن سٹاف کالج، ہری پور

Acknowledgement & Dedication: Dr Hafiz Safwan Muhammad Chohan has been in contact with Prof John McHardy Sinclair (June 14, 1933 - March 13, 2007), Professor of Modern English Language at Birmingham University, 1965-2000. He pioneered work in corpus linguistics, discourse analysis, lexicography, and language teaching, and was the man behind the machine gun of British National Corpus (BNC) and the Collins COBUILD dictionaries. There is no trend of dedicating research papers to any person but with high regret that this paper (both in Urdu & in English) was not written when he was alive, this effort is being dedicated to him.

کلیدی الفاظ

مشین ریڈایبل اردو، اردو لسانیات، لسانی انجینئرنگ، پلیٹ فارم سے ناوابستگی، استعمالی زبان (Functional Language)، عام بول چال (Informal Spoken Language)، الفاظ شماری، تعددِ استعمال، لغتیاتی تجزیہ، انگریزی کارپس، اردو کارپس۔

مخففات

UDB: Urdu Data Base/ Urdu Data Bank	اردو ڈیٹا بیس/ اردو ڈیٹا بینک:
TBU: The Bank of Urdu	اردو مثال گھر:
COCA: Corpus of Contemporary American English	امریکن انگریزی کارپس:
BNC: British National Corpus	برٹش نیشنل کارپس:
OCR: Optical Character Recognition	بصری حروف شناسی:
LEME: Lexicons of Early Modern English	پرانی انگریزی کے نظائر:
COBUILD: Collins Birmingham University International Language Database	کوبِلڈ:
CCED: Collins COBUILD Advanced Learner's English Dictionary	کولن فوقانی لغت:
CRULP: Centre of Research in Urdu Language Processing	مرکزِ تحقیقاتِ اردو:
CEUI: Centre of Excellence for Urdu Informatics	مرکزِ فضیلت برائے اردو اِطلاعیات:

مقتدرۂ قومی زبان پاکستان: NLA: National Language Authority, Islamabad, Pakistan

اِصطلاحات

اردو کارپس: اردو مثال گھر:	The Bank of Urdu
الفاظ شماری:	Word Count
امریکن انگریزی کارپس:	Corpus of Contemporary American English
اِطلاعیاتی ٹیکنالوجی:	Information Technology
اِطلاعیات:	Informatics
بالائے لغت معنی:	Ultra-dictionary meaning
پرانی انگریزی کے نظائر:	Lexicons of Early Modern English
پلیٹ فارم سے ناوابستہ:	Platform-Independent
تعددِ استعمال:	Word Frequency
دخیل ہم صورت لفظ:	False Friend
ذخیرۂ الفاظ/ متراکمہ:	Wordlist
روزمرہ زبان/ بول چال:	Contemporary [use of] Language
فطری زبان:	Natural Language
کارپس/ مثال گھر/ قاموس الامثال:	Corpus
کولن فوقانی لغت:	Collins COBUILD Advanced Learner's English Dictionary
لغتیاتی تجزیہ:	Lexical Analysis
لغوی اندراج:	Headword
مال خانہ:	Repository
معاصر زبان کے نظائر:	Examples of Contemporary use of Language
مواجہ:	(Internet) Site
(موادکی) جمع آوری:	Data Gathering

☆ تجارتی نشانات: اِس مقالے میں Collins، COBUILD اور The Bank of English کے الفاظ بار بار استعمال کیے گئے ہیں؛ یہ الفاظ ایک برطانوی اشاعتی ادارے Harper Collins Publishers Ltd کے تجارتی نشانات (ٹریڈ مارک) ہیں۔

تعارف

دنیا بھر میں زبانوں پر تحقیق کا کام اِس وقت زوروں پر ہے۔ زبان پر تحقیق سے مراد زبان کی ساخت پرداخت کا مطالعہ بھی ہے اور بین اللسانی تعلقات کا مطالعہ بھی۔ کمپیوٹر کی آمد کے ساتھ ہی زبانوں پر تحقیق کے علم میں نئی اور وسیع تر جہتیں سامنے آنا شروع ہوئیں اور خالص سائنسی انداز میں زبانوں کی ساخت اور اثرات کا جائزہ لیا جانا شروع ہوا۔ اِس مطالعے اور تحقیق کے لیے زبان کے معاصر نظائر کی بنیادی اہمیت ہے۔ چنانچہ مشین ریڈایبل حالت میں دنیا کی کئی زبانوں کے متون اِس مقصد کے لیے کارپس کی صورت میں جمع کیے گئے اور کمپیوٹر/ انٹرنیٹ پر محققینِ زبان ولسانیات کے لیے مہیا کیے گئے ہیں۔[۱] یہ متون وقت گزرنے کے ساتھ کمیت میں بڑھ رہے اور کیفیت میں بہتر ہورہے ہیں۔ لسانی انجینئرنگ (Language Engineering) اور لسانیات (Linguistics) کے سبھی شعبوں میں تحقیق کے لیے ایسے متون کی بنیادی ضرورت ہے۔

''دی بینک آف انگلش'' کے نام سے انگریزی زبان کا ایک بڑا امال خانہ (Repository) جس کے روح ورواں آنجہانی پروفیسر جان میک ہارڈی سنکلیئر (John McHardy Sinclair) تھے، اِس وقت برمنگھم یونیورسٹی برطانیہ میں قائم ہے اور دنیا بھر میں جاری لسانیاتی تحقیقات کے لیے موزوں لسانی متون محققین اور تحقیقی اداروں کو فراہم کررہا ہے۔ اِسی طرح کا ایک مواجہ ''امریکن انگریزی کارپس'' (COCA) ہے، جو ''دی بینک آف انگلش'' سے بھی کئی گنا زیادہ مواد مہیا کیے ہوئے ہے۔ زیرِ نظر مقالے میں اِنھی کارپسوں کے ڈھب پر اردو کارپس بنانے کی ضرورت اور لائحۂ عمل کے بارے میں بتایا گیا ہے اور اِس اردو کارپس کو ''دی بینک آف اردو'' کا نام دیا گیا ہے۔

ڈاکٹر حافظ صفوان محمد چوہان کی جانب سے مرکزِ فضیلت برائے اردو اِطلاعیات، مقتدرۂ قومی زبان اسلام آباد میں ''دی بینک آف اردو'' کے موضوع پر کام شروع کرنے کی تجویز دی گئی ہے۔ ڈاکٹر عطش درانی کی سربراہی میں یہاں پر ''اردو ڈیٹا بینک'' کے نام سے ایک شعبہ پہلے سے قائم ہے، لیکن یہیں پر کام کررہے ایک دوسرے شعبے ''اردو ڈیٹا بیس'' کے مخفف کے ہم آواز اور ہم صورت ہونے (Homonymous) کی وجہ سے اِلتباس (اِن دونوں شعبوں کا انگریزی مخفف UDB ہے) اور ''دی بینک آف انگلش'' کے نام سے بہت دور ہونے کی وجہ سے اِس کا نام بدلنے کی تجویز دی گئی، جسے ''اردو اِطلاعیات: آج اور کل'' کے عنوان سے ۷-۸ جون ۲۰۰۸ء کو ہونے والی ایک قومی ورکشاپ میں ماہرینِ اردو و اِطلاعیات اور حکومتِ پاکستان کے نمائندہ حکامِ بالا نے قبول کیا؛ ڈاکٹر حافظ

صفوان ہی کی تجویز پر اردو کارپس کے لیے ''دی بینک آف اردو'' اور اِس کی متبادل اردو اِصطلاح ''اردو مثال گھر'' بھی قبول کی گئی۔

متذکّرۂ بالا انگریزی کارپس اپنی ساخت اور استعمال (architecture and interface) میں چونکہ بالکل یکساں ہیں اِس لیے اِس مقالے میں جہاں ''دی بینک آف اردو'' سے تقابل کے لیے انگریزی کارپس کی بات کی جائے گی وہاں سہولت کے لیے ''انگریزی کارپس'' کا ایک ہی نام اِن دونوں مواجہات (Sites) کے لیے استعمال کیا جائے گا۔ اِسی طرح اِصطلاحات: دی بینک آف اردو/ اردو مثال گھر/ اردو کارپس بھی باہم مترادف ہیں۔

یہ سفارش بھی کی جاتی ہے کہ لفظِ کارپس کے لیے کوئی متبادل اردو اِصطلاح نہ بنائی جائے کیوں کہ دنیا بھر میں یہ لفظ اپنے مخصوص معنی میں استعمال میں ہے۔ ''کارپس'' کا لفظ بھی اُسی طرح اُردو الیا جائے جیسے مثلاً کمپیوٹر، اِنٹرنیٹ اور ٹیلی فون وغیرہ الفاظ زبانِ اردو کے جسم کا حصہ بن چکے ہیں۔

مجوزہ اردو کارپس بوجوہ انگریزی کارپس سے مختلف ذرائع سے بھی مواد لے گا۔ اِن وجوہ اور اِن ذرائع کے بارے میں تفصیلات بھی اِس ابتدائی/ تعارفی مقالے میں پیش کی گئی ہیں۔

1: کارپس اور دی بینک آف انگلش/ امریکن انگریزی کارپس (COCA): مختصر تعارف

کارپس کیا ہوتے ہیں؟ مختصر جواب یہ ہے کہ یہ مشین ریڈایبل متن ہوتے ہیں جنھیں لسانیاتی تحقیقات کے لیے اِکٹھا اور جمع کیا گیا ہو۔ ملاحظہ کیجیے اوکسفرڈ کی تعریف:

> A corpus is a collection of written material in machine-readable form that has been put together for linguistic research.[2]

دی بینک آف انگلش تقریباً ۴۵۰ ملین بولے اور لکھے جانے والے الفاظ کا مجموعہ ہے جسے کمپیوٹر میں زبان کے لغتیاتی تجزیے اور تحقیقی استعمال کے لیے اِکٹھا کیا گیا ہے۔ یہ مواد ۱۹۹۰ء سے اب تک کے استعمال ہونے والے، ذخیرہ کردہ متون سے لیا گیا ہے۔ ہر مہینے اِس مواد میں ۱۰ ملین الفاظ کا نیا متن ڈالا جاتا ہے۔ اِس مال خانے میں موجود متون کا بیشتر حصہ برطانوی انگریزی پر مشتمل ہے جب کہ امریکن انگریزی کے متون ۲۵٪ ہیں؛ پانچ فیصد مواد دنیا بھر میں بولی جانے والی انگریزی سے لیا گیا ہے جس میں آسٹریلیائی اور سنگاپور کی انگریزی بھی آتی ہے۔[۳]

امریکن انگریزی کارپس (COCA) میں ۳۸۵ ملین سے زیادہ الفاظ ذخیرہ ہیں اور یہ امریکی انگریزی اور

علاقائی لہجوں کے انگریزی متون کو ہمدست کیے ہوئے ہے۔ ہر سال اِس مقدار میں کم سے کم دو مرتبہ الفاظ کا اضافہ کیا جاتا ہے جو ۲۰ ملین الفاظ پر مشتمل ہوتا ہے۔ اِس مال خانے میں ۱۹۹۰ء سے دورِ حاضر تک کے متون موجود ہیں۔[۴]

انگریزی کارپس میں صرف مشین ریڈ ایبل انگریزی متون جمع ہیں۔ یہ متون پلیٹ فارم سے ناوابستہ (Platform Independent) حالت میں ہیں اور کسی بھی کمپیوٹر پروگرام کے لیے استعمال ہونے کی صلاحیت رکھتے ہیں۔

انگریزی کارپس میں صرف آج کی بولی اور لکھی جانے والی یعنی معاصر، استعمالی انگریزی زبان کے نظائر جمع ہیں؛ زبان کے کلاسیکل استعمال اور یابسات سے اِسے واسطہ نہیں۔ کلاسیکی انگریزی اور مختلف شعبہ ہائے زندگی کی مخصوص زبانوں کے لیے علیحدہ مواجہات مہیا ہیں۔

1.1: انگریزی کارپس کے مآخذ

''دی بینک آف انگلش'' میں لیا گیا مواد اخبارات، میگزین، فکشن اور نان فکشن کتب، پمفلٹ اور خطوط وغیرہ سے لیا گیا ہے۔[۵] اِس کا دوتہائی حصہ میڈیا کی زبان پر مشتمل ہے، یعنی اخبارات اور میگزینوں کے ساتھ ساتھ ریڈیو اور ٹی وی کی زبان۔[۶] اِس مواد کے تقریباً ۲۰ ملین الفاظ عام بول چال پر مشتمل ہیں۔ یہ متن غیر رسمی ملاقاتوں، انٹرویوز اور مباحثوں کو ریکارڈ کر کے لکھا گیا ہے۔[۷] یہ حصہ اِس لیے رکھا گیا ہے کہ انگریزی زبان کے زندہ استعمال اور الفاظ و مرکبات کے مصادیق میں در آنے والے تازہ ترین رجحانات سے باخبر رہا جا سکے۔

1.2: انگریزی کارپس میں کہاں سے متن نہیں لیا جا رہا؟

انگریزی کارپس میں ادبی زبان کو شامل نہیں کیا گیا، یعنی ناول اور شاعری کو بینک کا حصہ نہیں بنایا گیا۔ اِسی طرح یہ کارپس سائنسی مقالات اور مختلف شعبوں کی مخصوص زبان (Jargon; slang) کو بھی محفوظ نہیں کرتے۔ وجہ یہ ہے کہ اِن سب جہات کی زبان عام زبان نہیں ہوتی بلکہ اہلِ علم یا مخصوص لوگوں/شعبوں کی زبان ہوتی ہے۔ یہ کارپس اِن پگڈنڈیوں کی نہیں بلکہ شاہراہ کی زبان پر وقت اور صلاحیتیں خرچ کرنے اور کرانے سے متعلق ہیں۔

''دی بینک آف انگلش'' اور بقیہ انگریزی کارپس بھی چلتے ہوئے کاروبار میں اپنا مال لگاتے ہیں نہ کہ محدود یا مخصوص، چھوٹے کاروباروں میں۔

1.3: مواد/متون جمع کرنے کے طریقے

مشین ریڈ ایبل متن مختلف طریقوں سے اِکٹھا کیا جا سکتا ہے۔ مثلاً یہ کہ مختلف ویب سائٹوں سے مطلب کا مواد حاصل کیا جائے۔ اگر مواد اہم ہے لیکن متن (Text) کی شکل میں نہیں ہے تو اِسے متن کی صورت میں ٹائپ کرایا جائے۔ یہ ٹائپنگ معیاری ہونی چاہیے۔ آواز سے متن (Speech-to-text) اور تصویر سے متن (بصری حروف شناسی: OCR) کی جدید مشینوں کو استعمال کر کے بھی یہ مواد جمع کیا جاتا ہے۔

مواد کی جمع آوری (Data Gathering) میں بنیادی بات یہ ہے کہ یہ کم وقت میں اور کم خرچ میں ہو، لیکن اِبہام اور اغلاط سے پاک ہو۔

1.4: انگریزی کارپس کے اِستعمالات

انگریزی کارپس بنانے کی اہم ترین وجہ ایک ایسا مثال گھر/ قاموس الامثال تیار کرنا تھا جو محققین اور لغت نویسوں کو الفاظ کے متعلق زیادہ سے زیادہ اور معیاری معلومات فراہم کر سکے۔ وقت گزرنے کے ساتھ ساتھ اِس کے اِستعمالات کی نئ سے نئ صورتیں سامنے آتی رہی ہیں، اور اِن میں اضافہ روز افزوں ہے۔ چنانچہ انگریزی کارپس میں الفاظ کی سب تصریفی صورتوں (Inflected Orthographic Forms) کا استعمال موجود ہوتا ہے۔ اِن کارپسوں میں نظائر چونکہ موجودہ زبان (یعنی آج کل کی بولی اور لکھی جانے والی زبان) سے لی گئی ہیں لہٰذا زبان کا تازہ ترین، زندہ استعمال مہیا کر کے جاری تحقیقات میں سہولت فراہم کی جاتی ہے۔

2: انگریزی کارپس کیا ہے؟ کیا نہیں؟

انگریزی کارپس ذخیرۂ الفاظ یا مترا کمہ (wordlist) نہیں ہیں بلکہ الفاظ کی جملہ تصریفی صورتوں اور ہر سطحی استعمال کے نظائر (مثالوں) پر مشتمل متون کا مجموعہ ہیں۔ اِن کارپسوں میں میں موجود متون ۱۹۹۰ء اور اِس کے بعد کے ہیں۔ ''دی بینک آف انگلش'' میں ۱۹۹۰ء سے پہلے کے نظائر بینک سے نکال کر سرد خانے میں ڈال دیے گئے ہیں۔[۸] ''امریکن انگریزی کارپس'' میں تو ۱۹۹۰ء سے پہلے کا مواد موجود ہی نہیں ہے۔ چنانچہ یہ کارپس جدید انگریزی زبان کے بینک ہیں نہ کہ جدید انگریزی ادب کے۔

دی بینک آف انگلش، امریکن انگریزی کارپس اور بقیہ تمام انگریزی کارپس[۱] انگریزی لسانیات یا لسانی انجینئرنگ کے مترادف نام نہیں ہیں۔ اِن کی مثال عام بینک کے طور سے لی جا سکتی ہے۔ بینک رقوم جمع کرتا ہے اور

محفوظ رکھتا ہے۔ بینک رقوم کو استعمال نہیں کرتا بلکہ مختلف کاروبار کرنے والوں کو استعمال کے لیے دیتا ہے؛ اُن کے اِس استعمال سے ملنے والے نفع میں شرکت کرتا ہے؛ وغیرہ۔ انگریزی کارپس بھی الفاظ اور اُن کے استعمالات کو صرف محفوظ رکھتے ہیں اور محققینِ زبان ولسانیات اور لغت نویسوں وغیرہ کو اپنے پاس موجود متون پر نفع بخش کام کرنے کی اجازت دیتے ہیں، اور اِن کاموں کے نتائج سے اپنی قدر و قیمت میں اضافہ کرتے رہتے ہیں۔

''کارپس'' اور ''لغت'' میں بحیثیتِ اِصطلاح بھی فرق کرنا ضروری ہے: کارپس کی صورت میں جمع کیے گئے متون پر تحقیقات کر کے لغت بھی شائع ہوسکتا ہے۔ بالفاظِ دیگر، کارپس کے کئی اِستعمالات ہوتے ہیں جن میں سے ایک لغت کی تیاری اور اشاعت بھی (ہوسکتا) ہے، یعنی، لغت کارپس کی مدد سے کی گئی لسانی ولسانیاتی تحقیقات کا ایک نتیجہ ہوسکتا ہے۔ چنانچہ کارپس اور لغت مترادف الفاظ/ اصطلاحات نہیں ہیں۔ کارپس مشین ریڈایبل متون کا مجموعہ ہے؛ یہ الیکٹرانک صورت میں ہوتا ہے۔ جب کہ لغت کتابی صورت میں ہوتا ہے۔ اگرچہ الیکٹرانک لغات بھی موجود ہیں، لیکن ہر الیکٹرانک لغت کارپس کی بنیاد پر بنایا گیا ہو، یہ ہرگز ضروری نہیں۔

3: کلاسیکل انگریزی اور انگریزی کارپس

جیسا کہ اوپر ذکر کیا گیا، انگریزی کارپس انگریزی زبان کے معاصر استعمال کے نظائر پر مشتمل ہیں۔ سوال پیدا ہوتا ہے کہ پھر کلاسیکی انگریزی کہاں گئی؟ جواب یہ ہے کہ کلاسیکل اور ابتدائی اَدوار کی زبان کو الگ سے موجود ڈیٹا بینکوں (مثال گھروں) کی شکل میں محفوظ کیا جاتا ہے۔ اِس کی ایک مثال LEME ہے، جس میں انگریزی کے ایسے نظائر جمع کیے گئے (اور کیے جارہے) ہیں جنھیں ''پرانی'' انگریزی (Early Modern English) کہا جاتا ہے۔ ملاحظہ کیجیے:[۹]

4: اردو کارپس (اردو مثال گھر/ دی بینک آف اردو)

اردو کارپس، دی بینک آف انگلش اور امریکن انگریزی کارپس کے ڈھب پر تجویز کیا گیا اردو متون کا ایسا خزانہ ہے جو الفاظ کے استعمال کے نظائر ذخیرہ کرے گا تاکہ اردو پر بطور فطری زبان (Natural Language) تحقیق ہوسکے۔ اِس میں موجود سارے متون مشین ریڈایبل اردو میں اور پلیٹ فارم سے ناوابستہ حالت میں ہوں گے تاکہ کسی بھی مواجہ پر اردو میں کام کرنے والے محقق یا مشین کو متن فراہم کیا جاسکے۔

اردو کارپس اپنے مال خانے میں بنیادی طور پر سب متون کو اردو کے روایتی رسم الخط

(Indo-Perso-Arabic Script) میں رکھے گا، اگر چہ کسی وقت میں رومن اردو اور دیوناگری کے لیے سہولت (support) بھی فراہم کی جاسکے گی۔ چونکہ فوری طور پر یہ کارپس دائیں سے بائیں لکھے جانے والی اردو کے متون جمع اور فراہم کرے گا اِس لیے اردو کی ہم رشتہ، تمام پاکستانی زبانوں کے متون ذخیرہ کرنے کے لیے اگر کارپس بنائے جائیں (جنھیں ضرور اور فوراً بنانا شروع کرنا چاہیے) تو اِن سب کو اِس اردو کارپس کے ذیلی اِداروں (Subsidiaries) کے طور پر رکھا اور استعمال کیا/کرایا جاسکتا ہے۔

4.1: اردو کارپس کے مآخذ، دائرۂ عمل اور انگریزی کارپس سے اِن کا فرق

شان الحق حقی (۱۹۹۶ء) نے لکھا ہے کہ ’’......اردو پنپنے اور پروان چڑھنے نہ پائی تھی کہ کلاسیکیت کی راہ پر چل نکلی۔ شعر گوئی کے کام کی رہ گئی۔‘‘ اِسی طرح گوپی چند نارنگ (۱۹۶۸ء) نے لکھا ہے کہ:

> ’’......ہماری زبان [اردو] نے شاعری کی آغوش میں آنکھ کھولی تھی۔ جس سماج میں یہ پروان چڑھی تھی وہ داستانیں سنتا اور شعر میں خط لکھتا تھا۔ ہمارے ہاں یہ روایت رہی ہے کہ شاعری ادب ہے اور ادب زبان ہے۔ علمی اردو نثر کی عمر ایک ڈیڑھ صدی سے کم نہیں، لیکن یہ بات دلچسپی سے خالی نہیں کہ ہمارے ہاں افسانے کی زبان، ناول اور ڈرامے کی زبان، اور تو اور ہماری تنقید اور تحقیق کی زبان ہنوز شعر کے اثر سے پوری طرح آزاد نہیں ہوسکی۔ ہمارے نثر نگاروں کی ایک بڑی تعداد آج بھی شعر کے بغیر لقمہ نہیں توڑ سکتی۔......‘‘

بات درست ہے۔ اردو پر بطورِ زبان کوئی تحقیق کرنے، علی الخصوص کمپیوٹر پر کوئی ایسا کام کرنے کا ارادہ کیا جائے تو معلوم ہوتا ہے کہ اردو کا تحریری سرمایہ، زیادہ تر، زبان کے صرف ادَبی اِستعمالات کے نظائر پر مشتمل ہے۔ اِس لیے ابتدا میں اردو کارپس کے منصوبے پر انگریزی کارپس سے خاصے فاصلے پر رہ کر کام کرنا ضروری ہوگیا ہے۔ اردو زبان کی ترقی چونکہ کئی لحاظ سے اردو ادب ہی کی ترقی کی مرادف اور مترادف رہی ہے اِس لیے اردو متون کی جمع آوری میں اردو ادب کو علی الکلیہ دست بسر کرنا یا ’’نہیں‘‘ کہنا ممکن نہیں۔ یاد رہے کہ انگریزی کارپس نے ادَبی متون کو اپنے مال خانے کا حصہ نہیں بنایا۔ بایں وجوہ اِس تجویز میں یہ بات قابلِ ذکر ہے کہ اردو کارپس ابتدائی طور پر کچھ اردو اخبارات، (کسی حد تک) ادب، (کسی حد تک) جدید شاعری، اِنٹرنیٹ اور دیگر شائع کردہ کتابوں سے اپنا رأس

المال (Capital) حاصل کرے گا۔ یہ بہت بنیادی فرق ہے جو انگریزی کارپس اور اردو کارپس میں ہے۔

اردو زبان کے متنوع اور بالخصوص بین الاقوامی استعمال پر نگاہ رکھنے والے اہلِ علم کے مشورے سے آغاز میں ایسا کچھ متن جمع کر کے کام شروع کیا/کرایا جا سکتا ہے۔ اردو زبان کے معاصر اِستعمالات کو پلیٹ فارم سے ناوابستہ، مشین ریڈ ایبل اردو میں جمع کرنے کا کام بھی فوراً شروع کر لیا جائے تاکہ اِس مد میں متون آتے رہیں۔ رفتہ رفتہ کلاسیکل متون اور پرانی اردو کے نظائر کے ذخائر کو LEME کے طرز پر الگ (Denest) کر دیا جائے تاکہ یہ ایک مستقل ماخذ کے طور پر پنپ سکے اور اردو پر کلاسیکی تحقیق میں کام آ سکے۔

فوری طور پر اردو لغت (تاریخی اصول پر)، شائع کردہ اردو لغت بورڈ کراچی سے بھی مدد لی جائے تو انتہائی مناسب اور برمحل ہے کیونکہ اردو کے ذخیرۂ الفاظ کے کلاسیکل استعمال کے سب سے زیادہ نظائر صرف یہیں سے، یکجا مل سکتے ہیں۔ لیکن اِس کی شکل یہ ہوگی کہ اوّلاً اُن مآخذ کی ایک فہرست بنا لی جائے جن پر اِس لغت نے زیادہ انحصار کیا ہے، اور ثانیاً اُن میں سے تازہ ترین مآخذ کو چھانٹ لیا جائے۔ اِن چھٹے ہوئے مآخذ کو مشین ریڈ ایبل اردو میں — اور درست اور معیاری، پلیٹ فارم سے ناوابستہ انداز میں — کمپیوٹر پر محفوظ کرنے کا آغاز کر دیا جائے۔

یہ بات واضح ہے کہ مجوزہ اردو کارپس کے لیے اگر اِسی لغت کے مآخذ پر کامل انحصار کیا جاتا ہے تو لگ بھگ LEME کے طرز پر کلاسیکی اردو کارپس وجود میں آئے گا؛ کلاسیکی اردو کارپس اور اُس کی تکنیک پر گفتگو اِس مقالے کے حیطے سے باہر ہے۔

4.2: اردو لغات: الفاظ شماری اور اِستعمالی زبان کے تناظر میں ایک جائزہ

اردو کے جتنے بھی لغات دستیاب ہیں اُن کے مرتبین و مولفین نے حسبِ استطاعت اور حسبِ ذوق نئے الفاظ اور الفاظ کے تازہ تر معانی کے اندراج کی کوشش ہر دور میں کی ہے، لیکن اِس ضمن میں چلن دار زبان کو باقاعدہ ریکارڈ کر کے کثرتِ استعمال (اِصطلاحاً: تعددِ استعمال) کی بنیاد پر الفاظ کو داخلِ لغت کرنے کی کوشش کہیں بھی نہیں کی گئی۔ اِسی بات کو دوسرے الفاظ میں یوں کہا جا سکتا ہے کہ اردو لغت نویسی کی تاریخ میں زبان ''بولنے'' والوں کی کبھی سنی ہی نہیں گئی اور اُنھیں ''بولنے'' کا موقع نہیں دیا گیا (Users of Urdu have no say in Urdu dictionaries.)۔ لہٰذا درست اِصطلاحی معنی میں اردو میں زبان ''بولنے'' والوں کا کوئی بھی لغت موجود نہیں ہے۔ اردو کے متداول لغات میں ایسے کسی مکمل لغت کا تو کیا ذکر، اگر کسی لغت نویس نے عوام کی زبانوں پر جاری لیکن اردو

کے سکہ بند مصنفین اور اہلِ قلم کا ''اچھوت'' کوئی دھونتال لفظ مارے باندھے اپنے لغت میں شامل کر ہی لیا ہے تو اپنی کراہت بلکہ برأت کے اظہار کے لیے کبھی ''عام''، کبھی ''عوام''، کبھی ''بول چال'' اور کبھی ''عامیانہ'' وغیرہ قبیل کے لیبل کا نظر بٹو بھی اُس اندراج پر ضرور چسپاں کیا ہے۔ اردو لغت نویسی کا عمومی ہنجار اردو ادب ہی کی ضرورتوں کا پورا کرنا رہا ہے؛ صرف ''فرہنگِ تلفظ'' (اشاعتِ اوّل: ۱۹۹۵ء) اور ''لغاتِ روزمرہ'' (اشاعتِ اوّل: ۲۰۰۳ء) کا اِستثنا اِس قانون کو ثابت کر رہا ہے۔

البتہ، ایک لغت، ''علمی اردو لغت'' (اشاعتِ اوّل: ۱۹۷۶ء) کے دیباچے ''سخن ہائے گفتنی'' میں یہ صراحت ملتی ہے کہ اخبارات و رسائل کے مطالعے کے بعد الفاظ کی فہرست تیار کر کے لغت نویسی کا کام کیا گیا۔ اردو کے لغات کی حد تک یہ ایک قابلِ قدر بات ہے۔ لیکن اخبارات وغیرہ میں سے کون کون سے حصوں سے مواد لیا گیا، اور لغت کے لیے الفاظ کی روائی یا ناروائی کس بنیاد پر کی گئی، اِس کے لیے کسی سائنسی اصول کی پیروی کیے جانے کی بابت کچھ نہیں بتایا گیا۔ مناسب معلوم ہوتا ہے کہ مؤلفِ لغت جناب وارث سرہندی کے الفاظ یہاں نقل کر دیے جائیں:

> ''......ہم نے لکیر کا فقیر بننے سے اِحتراز کرتے ہوئے قدیم و جدید کتبِ ادب و لغت، مختلف موضوعات کی کتب اور اخبارات و رسائل کے مطالعہ کے بعد الفاظ کی فہارس مرتّب کیں تا کہ اُن تمام الفاظ کا اِحاطہ کیا جا سکے جن سے قدیم و جدید ادب کے قارئین، طلبہ اور مختلف شعبہ ہائے حیات میں کام کرنے والے عام اردو دانوں کو سابقہ پڑتا ہے۔......''

کسی بھی زبان کے جسم (Body) میں استعمالی زبان (Functional Language) کو کچھ ایسی حیثیت حاصل ہے جیسے بدن میں چہرہ۔ زبان کے مکمل جسم☆ اور استعمالی زبان کا فرق معلوم کرنے کے لیے، یا بلکہ یوں کہیے کہ زبان کے مکمل جسم میں سے استعمالی زبان کے الفاظ کو ممتاز کرنے کے لیے، بہت سا مواد جمع کر کے الفاظ کی فہرست بنائی جاتی ہے اور اِسے حروفِ تہجی کی ترتیب میں رکھنے کی بجائے کثرتِ استعمال کی کلید سے چھانٹ کر ترتیبِ نزولی میں رکھا جاتا ہے۔ انگریزی زبان کی حد تک الفاظ ایسی پہلی فہرست Edward Lee Thorndike (1874–1949) نے ۱۹۲۱ء میں تیار کی جو Teacher's Word Book کے نام سے شائع ہوئی؛ یہ فہرست بیس ہزار الفاظ پر مشتمل تھی۔ اِس وقت دنیا بھر کی اہم زبانوں میں الفاظ کی ایسی فہرستیں آن لائن موجود ہیں؛ اِن فہرستوں کو تھوڑے تھوڑے عرصے کے بعد اَپ ڈیٹ بھی کیا جاتا ہے۔ ملاحظہ کیجیے۔[۱۰] اردو میں

الفاظ شماری کی پہلی با قاعدہ اور بڑے پیمانے پر سائنٹفک کوشش، جو''اردو الفاظ شماری'' کے نام سے کتابی شکل میں شائع بھی ہوئی، ڈاکٹر حسن الدین احمد (۱۹۷۳ء) نے کی۔ ۱۰۹۲۷ الفاظ پر مشتمل یہ فہرست ابھی تک کتابی شکل ہی میں ہے اور کمپیوٹرائز نہیں ہوئی۔ دورِ حاضر میں مرکزِ تحقیقاتِ اردو (CRULP) لاہور نے دسمبر ۲۰۰۷ء میں استعمالی اردو الفاظ کی ایک فہرست تیار کی ہے جو کہ انٹرنیٹ پر موجود ہے۔ یہ فہرست پانچ ہزار الفاظ کی ہے۔ ملاحظہ کیجیے: [۱۱]

الفاظ شماری لغت میں الفاظ کے شمول کے لیے کیا حیثیت رکھتی ہے اور کس طرح کے لغت میں کس نوعیت کے الفاظ کی سمائی ہونی چاہیے، اِس کا علم البتہ ضروری ہے۔ ڈاکٹر حسن الدین احمد ہی کے الفاظ میں:

''......لفظ شماری کا مقصد زبان کی جامع لغت تیار کرنا نہیں ہے۔ لغت کی تدوین میں ہر معلوم لفظ کو شریک کرلیا جاتا ہے۔ لفظ شماری میں وہی الفاظ شامل ہوں گے جو نمائندہ ادب میں موجود ہوں، یعنی الفاظ کی شرکت نمائندہ ادب کی تابع ہوتی ہے۔......''

لیکن اردو کے کن الفاظ کو لغویایا (Lexicalize) جائے اور اِس کے لیے مختلف علوم کے منتہیوں کی کتابوں اور تحریروں ہی کو کسوٹی بنانے کی بجائے بہت سے شعبوں کے بے شمار ''بولنے'' والوں کی گفتگو کو بنیاد بنا کر اور اِن کے زیرِ استعمال الفاظ کو زبان کا چلن مان کر اِس کثرتِ استعمال کو ناپنے کا کمپیوٹرائزڈ نظام بنایا جائے ــــ اِس ضرورت کو اردو لغت نویسوں میں سب سے پہلے شمس الرحمٰن فاروقی (۱۹۸۱ء) نے سمجھا اور ذکر کیا ہے۔ اُنھوں نے لکھا ہے کہ:

''......کون سے الفاظ لغت کا حصہ بن گئے ہیں ...... اُن کا تعین محض گذشتہ لغات، کتابوں، رسالوں، اخباروں کے ذریعہ نہیں ہوسکتا۔ اِن چیزوں کے علاوہ زبان بولنے والوں کی کثیر تعداد کی بول چال ریکارڈ کرنی ہوگی۔ پھر تمام کو کمپیوٹر کے ذریعے سے مرتب کرنا ہوگا۔ اگر کمپیوٹر میسر نہ ہو تو تمام الفاظ کے کارڈ بنا کر اُن کے گریڈ اور کثرتِ استعمال کا تعین کیا جائے۔''

استعمالی زبان کے حدود متعین کرنا بھی ضروری ہے۔ عمومی استعداد کے لوگوں کی عمومی ضروریات اور مختلف شعبوں کے لوگوں کی مخصوص شعبہ جاتی ضروریات کے لیے یکسر مختلف زبان استعمال ہوتی ہے۔ اردو کے ہر لغت نویس نے اِس مضمون میں اپنے ذوق اور اپنے زمانے کی ضرورتوں کے مطابق دادِ چناؤ دی ہے۔ چنانچہ کسی لغت میں الفاظ کی تعداد بہت بڑھ گئی ہے، اور کسی میں ضروری الفاظ بھی بار نہیں پاسکے۔ الفاظ کی سمائی کے اعتبار سے اردو کا اب

تک سب سے بڑا مکمل لغت ''جامع اللغات'' (اشاعتِ اوّل: ۱۹۳۵ء) ہے۔ ذیل میں اِس لغت کے مؤلف خواجہ عبدالمجید کے الفاظ نقل کیے جاتے ہیں تا کہ معلوم ہو سکے کہ اُن کے ہاں استعمالی زبان کی تعریف کیا ہے:

> ''...... اِس لغت میں فارسی، عربی، ہندی، سنسکرت، ترکی، عبرانی وغیرہ کے وہ الفاظ ہیں جو کسی نہ کسی صورت میں اردو زبان میں استعمال ہو رہے ہیں۔ اِن میں وہ الفاظ بھی ہیں جو ہندو اور مسلمان مشکل نویس اپنی سنسکرت یا عربی فارسی کی قابلیت جتانے کی غرض سے جا و بے جا استعمال کرتے رہتے ہیں۔ ہر چند اردو اُن الفاظ کے بغیر بھی فصیح ہو سکتی ہے۔...... ہندوؤں کو تو عربی فارسی زبانوں کے بہت سے مشکل الفاظ کے معنی اردو لغات میں مل بھی جائیں گے لیکن مسلمانوں کو سنسکرت اور ہندی الفاظ کے معانی کسی اردو لغت میں نہیں مل سکتے کیوں کہ اردو لغت نگار اِن کو غیر اردو الفاظ سمجھ کر، بالکل ترک کر دیتے ہیں۔ اِس لیے ہم نے جامع اللغات میں جہاں عربی اور فارسی زبانوں کے نامانوس اور مشکل الفاظ دیے ہیں وہاں سنسکرت اور ہندی زبان کے ادق الفاظ بھی درج کر دیے ہیں۔......''

چنانچہ معلوم ہوا کہ خواجہ عبدالمجید کے ہاں ''استعمالی زبان'' سے Functional Language مراد نہیں ہے جیسا کہ آج ہم مراد لیتے ہیں، اگرچہ اُن کی اِعتنا سے ''جامع اللغات'' کا اردو کے مترا کمہ (wordlist) ہونے کی حیثیت بہر حال قائم ہو گئی ہے۔

4.3: اردو کارپس میں فی الوقت متون کہاں سے نہیں لیے جائیں گے؟

فی الحال اردو کارپس کے مال خانے میں رکھے جانے والے متون بہت ہی مخصوص جگہوں سے لیے جانے کی تجویز ہے۔ یہ کارپس ریڈیو، ٹی وی اور میڈیا کی زبانوں سے (زیادہ تر) مواد نہیں لے گا۔ چنانچہ عام بول چال کی زبان اِس میں عموماً شامل نہیں کی جائے گی؛ لیکن اگر کہیں سے معیاری متن میں محفوظ کردہ بول چال کی زبان مل جائے تو اُسے ضرور لیا جائے گا۔ انگریزی کارپس کی طرح اِس میں سائنسی تکنیک سے متعلق زبان بھی شامل نہیں کی جائے گی، اگرچہ مستقبل میں اِس پر بھی کام کیا جا سکتا ہے۔ فوری طور پر کئی قسم کے انٹرویوز (مصاحبوں) اور ٹاک شوز

(مباحثوں) وغیرہ میں شامل ہونے والے الفاظ بھی اِس کارپس کے مال خانے کا حصہ نہیں بنیں گے۔ وغیرہ وغیرہ۔

اردو کارپس کے مآخذ کو اِتنا محدود کرنے کے وجوہ بہت سے ہیں۔ سرِ دست اِس ضمن میں چند سامنے کی باتیں عرض کی جاتی ہیں:

۱۔ معیاری متن کی جمع آوری کے لیے بہت وقت درکار ہوگا۔ موجودہ اردو متون، زیادہ تر، معیاری نہیں ہیں۔ اِس لیے ادب کے منتخب متون کو بھی معیاری بنا کر کام کرنا ضروری ہے۔ معیاری متون سے مراد اِملاء کے مسائل کا حل بھی ہے۔

۲۔ موجودہ اردو متون، زیادہ تر، مشین ریڈایبل حالت میں بھی نہیں ہیں۔ اِس ضمن میں مرکزِ تحقیقاتِ اردو لاہور کے مدیحہ اعجاز اور ڈاکٹر سرمد حسین (۲۰۰۷ء) نے تفصیل سے بتایا ہے کہ ہمارے ہاں لوگ نہ صرف متن فراہم کرنے میں فراخ دِلی کا ثبوت نہیں دیتے بلکہ اخبارات والے لوگ تو علی العموم متن کو مشین ریڈایبل بنانے کی ضرورت سے بھی آگاہ نہیں ہیں۔ اِس کے بالمقابل انگریزی میں صورتِ حال یکسر مختلف ہے: سبھی اہم اخبارات اور رسالوں نے اپنے متون اِنٹرنیٹ پر مہیا کیے ہوئے ہیں۔ مثال لیجیے کہ ٹائم میگزین والوں نے اپنے سارے متن اور تصاویر، پہلی اِشاعت (۱۹۲۳ء) سے لے کر اب تک کے، مشین ریڈایبل حالت میں رکھے اور مہیا کیے ہوئے ہیں۔[۱۲] کارپس صرف لکھے ہوئے (تحریری) متون کو اپنا حصہ بناتا ہے۔ چنانچہ اردو کارپس میں ذخیرہ کیے جانے والے سارے متون کو پہلے مشین کے لیے قابلِ فہم (مشین ریڈایبل) بنانا ہوگا، جس کے لیے کافی وقت درکار ہے۔ بدیں وجہ فوری طور پر مجوزہ اردو کارپس کا دائرہ کار اِتنا وسیع نہیں کیا گیا۔

۳۔ جیسا کہ اوپر اشارہ کیا گیا، اردو کارپس میں شامل متون کے الفاظ تصدیق شدہ اور معیاری اِملاء میں ہونا ضروری ہیں تا کہ تحقیق کرنے والے لوگ زبان پر تحقیق کریں نہ کہ اِملاء کے مسائل میں اُلجھ کر رہ جائیں۔ دنیا بھر کی تو ایک رہی، خود جزائرِ برطانیہ ہی میں بھانت بھانت کی انگریزی بولی جاتی ہے۔ چنانچہ انگریزی نے مختلف الاملاء اور مختلف الہجا الفاظ کے لیے یہ ترتیب اختیار کی ہے کہ کسی ایک ہجے کو معیاری مان کر لغت میں درج کر کے کام شروع کر دیا ہے جب کہ اختلافی اِملاء کو ساتھ میں رہنے دیا ہے۔ انگریزی کے بلاتخصیص کسی بھی لغت کو دیکھ لیجیے، ہر صفحے پر اِختلافی اِملاء والے کچھ نہ کچھ لفظ ضرور نظر آئیں گے۔ اردو میں ابھی تک ایسا نہیں ہوا۔ اِملاء اور ہجا کے مسائل کبھی مکمل طور پر حل نہیں ہوں گے؛ اور کسی لفظ کا ایک ہی اِملاء

ساری دنیائے اردو قبول کرلے، یہ بھی ممکن نہیں۔ انگریزی میں بھی ایسا نہیں ہو سکا۔ لہٰذا انگریزی کے تتبع میں اختلافِ اِملاء و ہجا کے ساتھ ہی کام کرنا ہوگا۔[۱۳]

4.4: اردو کارپس کے اِستعمالات: آج اور کل

اردو کارپس کا استعمال موجودہ دور کے اردو محققین اور اہلِ معانی کی اوّلین ضرورت ہے۔ اِس کے اِستعمالات کے امکانات علی الکلیہ وہی ہیں جو انگریزی کارپس کے ہیں۔ بلکہ کئی جہات میں یہ امکانات متنوع تر ہیں۔ لسانی اور لسانیاتی تحقیق کے لیے عام استعمال میں آنے والی اردو کی مثالیں بنانا اور جمع کرنا— اور اِس کام کو مستقل طور پر کیے جانا— بے حد ضروری ہے۔ اِس قسم کے ذخیرۂ امثال اور قاموس الامثال کی ضرورت، تیاری اور لائحۂ عمل پر الگ سے بحث موجود ہے (ڈاکٹر حافظ صفوان محمد چوہان: ۲۰۰۷ء) اور اس کے لیے مآخذ اور اردو کے لیے ہمارے ماحول کے مطابق خاص انداز میں کام کرنے کی بابت کئی جہات پر تفصیلی گفتگو کی گئی ہے۔ دورِ حاضر میں اردو میں جاری اور ممکنہ لسانی تحقیقات کے لیے مجوزہ اردو کارپس کی حیثیت، ریڑھ کی ہڈی کی سی ہے۔

اِس ترتیب پر یعنی اردو کارپس کی شکل میں اردو کے متون اگر ذخیرہ کیے جائیں، اور ان کو مستقل طور پر اَپ ڈیٹ بھی کیا جاتا رہے، تو اردو لسانی تحقیق کا نیا منظرنامہ سامنے آتا ہے۔ اور اگر سب پاکستانی زبانوں کے متون (اُن کے روایتی رسوم الخط میں) ذخیرہ کرنے کی بات بھی چل پڑے تو اردو کو وہ مقام ملنے کی امید کی جاسکتی ہے جس کی وہ جائز طور پر حق دار ہے؛ یہ سب کارپس مل کر دنیا میں جاری لسانیاتی تحقیقات کو بہت کام کے بنیادی متون فراہم کر سکتے ہیں۔

طلبہ کی تدریسی ضرورتوں کو پورا کرنے کے ساتھ ساتھ اردو کارپس مختلف علوم و فنون کے اساتذہ اور اساتذۂ زبان وادب کے لیے بھی فائدہ مند ہوگا کیونکہ یہ لوگ اِس کے ذریعے سے اپنے طلبہ کی ضروریات کے مطابق تازہ بتازہ مواد حاصل کریں گے۔ کلاسیکی اردو ادب کے طلبہ و اساتذہ بھی اِس مواجہ کو اپنے لیے ویسا ہی سودمند پائیں گے جیسا کہ یہ جدید لسانیاتی تحقیقات کے لیے ہوگا۔

یہ کارپس اصلاً تو اِنٹرنیٹ/کمپیوٹر پر مہیا ہوگا کہ اِس کا فائدہ اِنھی مواجہات سے سب سے زیادہ ہوگا، لیکن ''دی بینک آف انگلش'' کے تتبع میں چند سال (آسانی کے لیے سمجھ لیجیے کہ ہر پانچ سال) کے بعد جمع شدہ مواد سے حاصل کی گئی نظائر کی بنیاد پر اِس سے ایک لغت بھی تیار کر کے شائع کیا جاسکے گا۔ یہ بات ظاہر و باہر ہے کہ یہ لغت

تازہ ترین، معاصر اردو زبان کا آئینہ دار ہوگا۔ کولن فوقانی لغت (Collins COBUILD Advanced Learner's English Dictionary) کے تازہ یعنی چوتھے ایڈیشن کو اِس لغت کی مثال میں ملاحظہ کیا جا سکتا ہے۔ اپنی علمی اِفادیت کے ساتھ ساتھ ایسا لغت شائع کرنا اِس لیے بھی ضروری ہے کہ یہ ایک دیر تک قابلِ استعمال، نظر آنے والی چیز ہوتا ہے۔

4.4.1: اردو کارپس کے اِستعمالات: لغتیاتی تجزیے اور اردو لسانیات

اِس نگاہ سے دیکھیں تو بنیادی طور پر یہ کارپس لغت نویسوں اور زبان دانوں کے لیے بڑے کام کی چیز ہے۔ لفظ کے استعمال کی زیادہ سے زیادہ نظائر کی ضرورت اِس لیے ہے کہ بڑے سے بڑا زبان دان اور لغت نویس بھی لفظ کے سب معنوں اور مفاہیم کو ویسے نہیں برت سکتا جیسا کہ عوامی جینیئس کر سکتا ہے۔ لفظ کے مزاج سے متعلق سب حقائق کو جمع کرنا اور اُن کی بنیاد پر نتائج نکالنا اکیلے آدمی کے لیے ممکن ہی نہیں ہے۔

یہ مجوزہ کارپس ہر لفظ اور ہر لغوی اندراج (Headword) کی تہ میں موجود ہوگا۔ لغت نویس ہو یا محقق، وہ جس لفظ کے اندراج پر کام کرنا چاہے اُس کے استعمال کی صدہا صورتیں انگلی کے صرف ایک اشارے سے کمپیوٹر کی سکرین پر آموجود ہوں گی۔ لفظ کا ہر سطحی استعمال اور ہر نوعی معانی مثلاً محاوراتی، اِستعاراتی، فرضی، مرادی، حقیقی، مجازی، اصطلاحاتی، موضوعی، وقتی، سلینگی، وغیرہ، سامنے پا کر ہر طرح کی تحقیق میں سہولت مل سکے گی۔ یہاں تک کہ الفاظ کے بالائے لغت معنی بھی سامنے ہوں گے۔ لفظ مجرد شکل میں لیکن مختلف معنی و مفہوم میں بھی سامنے ہوگا اور اپنی جملہ تصریفی شکلوں، تعلیقیوں (Affixes) اور مرکبات کی صورت میں بھی۔

یکرو اِملائی صورتوں والے بہت سے الفاظ ایک سے زیادہ قواعدی حیثیت کے حامل ہوتے ہیں۔ یہ صورتِ حال دنیا کی ہر زبان میں پائی جاتی ہے، کسی میں کم اور کسی میں زیادہ۔ مثلاً اِس مثالی جملے:

"بچہ آیا کے ساتھ آیا۔"

میں آیا کے ایک اِملاء/ لفظ کی دو یکسر مختلف قواعدی حیثیتیں ہیں جب کہ صوت وصورت بالکل ایک ہی ہے۔ لسانی انجینئرنگ میں جاری ترقیات کے سبب سے مجوزہ کارپس کے مواجہ کا کمپیوٹر پروگرام اِتنا "سمجھ دار" (Artificially Intelligent) ہوگا کہ یہ ہم شکل اور ہم آواز لیکن معنی میں مختلف تجنیس کے حامل الفاظ (Homonyms) کے درمیان فرق کر سکے گا؛ چنانچہ یہ پروگرام ہر دو آیا کی اصل (root) کے مطابق نتائج فراہم

کرے گا۔ اسی طرح یہ پروگرام اس مثالی جملے:

[مکرر] ''سالوں کی محنت سے اشرف آج اس مقام پر ہے۔''

میں جہاں سالوں کے لفظ کو نری اِملائی حالت (Orthograph) میں پیش کر سکے گا وہیں اس جملے کے سیاق و سباق سے یہ فیصلہ بھی کر سکے گا کہ یہاں یہ لفظ سال سے مشتق ہے یا سالا سے۔ بالکل یہی اَحوال ذیل کے دونوں جملوں میں لفظ کمروں کے لیے ہوگا، کہ آیا یہ لفظ کمر سے مشتق ہے یا کمرہ سے:

[مکرر] الف: دونوں کمروں میں سفیدی ہوگئی۔

ب: دونوں کی کمروں کا ناپ ایک ہی ہے۔

یہی صورتِ حال Code-mixing کے لیے ہوگی۔ چنانچہ اِس پروگرام کو یہ طے کرنے کے لیے تیار (Train) کیا جاسکے گا کہ مندرجۂ ذیل مثالی جملے:

[دیگر] ''میں دفتر سے لیٹ آیا اور آتے ہی بستر پر لیٹ گیا۔''

میں پہلا لیٹ اصلاً انگریزی لفظ Late ہے جب کہ دوسرا لیٹ اردو مصدر لیٹنا کی ایک تصریفی شکل ہے۔ مجوزہ اردو کارپس کے اندر اردو-انگریزی Code-mixing کی یہ بالکل سادہ شکل ہے؛ ذیل میں اِس کی ایک نسبتاً پیچیدہ شکل کی مثالیں پیش کی جاتی ہیں:

ایسا لفظ جو دخیل ہو، اور اُسی کا ہم صورت اور ہم صوت اپنی زبان میں بھی موجود ہو، اور یہ دخیل لفظ اپنی زبان کے قاعدوں سے تصریفی شکلیں اختیار کرلے تو اِسے اپنی زبان کے لفظ کا False Friend کہتے ہیں۔ اردو میں صرف یہی نہیں ہوتا کہ انگریزی کے دخیل الفاظ اپنی اصلی حالت میں رہیں بلکہ دوسری بہت سی زبانوں کے دخیل الفاظ کی طرح انگریزی کے دخیل الفاظ بھی اردو کے قاعدوں سے جمع اور دیگر تصریفی شکلوں میں بدل جاتے ہیں۔ ذیل میں دو دو جملوں پر مشتمل دو سیٹ ملاحظہ کیجیے جن میں انگریزی کے دو الفاظ cream اور scene اردو کے ایک عام قاعدے سے جمع بنے ہیں:

[دیگر] الف: آئس کریموں سے بھرا ڈیپ فریزر خراب ہوگیا۔

ب: اللہ سخیوں کا سخی اور کریموں کا کریم ہے۔

[مکرر] الف: تھیٹر میں لڑائی والے سینوں پر لوگ جذباتی ہوجاتے تھے۔

ب: صدر نے جیتنے والوں کے سینوں پر تمغے آویزاں کیے۔

مندرجۂ بالا جملوں میں پہلے سیٹ کے پہلے جملے میں لفظ کریموں کی لغت [cream + وں لاحقۂ جمع] ہے جب کہ دوسرے جملے میں موجود اِسی صوت وصورت کے لفظ یعنی کریموں کی لغت [کریم + وں لاحقۂ جمع] ہے۔ اگر کریم کو اردو کا لفظ مانا جائے اور cream کو دخیل، تو اِس صورت میں cream + وں سے بننے والے کریموں کو False Friend کہا جائے گا۔ یہی صورت جملوں کے دوسرے سیٹ میں موجود ہے: اِس میں پہلے جملے میں سینوں کی لغت [scene + وں لاحقۂ جمع] ہے جب کہ دوسرے جملے میں موجود سینوں کی لغت [سینہ + وں لاحقۂ جمع] ہے۔ سینہ چونکہ اردو کا لفظ ہے اِس لیے scene + وں سے بننے والا سینوں اردو والے سینوں کا False Friend ہوا۔ علیٰ ہٰذا۔ مجوزہ اردو کارپس کو اِس قسم کے مسائل کے حل کے لیے تیار کیا جا سکے گا۔

لیکن لسانی انجینئرنگ جتنی بھی ترقی کر لے، کسی بھی فطری زبان پر تحقیق کے لیے انسانی عنصر (Human Element) کی بنیادی ضرورت ہے کیوں کہ یہ زبان انسان بولتے ہیں، مشینیں نہیں۔ مثلاً یہ شعر دیکھیے: ؎

[دیگر] کچھ لفظ درختوں کے تنوں پر بھی کھدے ہیں

جنگل کی گواہی تجھے شہروں میں نہ آلے [۱۴]

اِس شعر کے پہلے مصرع میں لفظ تنوں پر غور کیجیے۔ یہ لفظ بیک وقت تن سے بھی مشتق ہوسکتا ہے اور تنا سے بھی۔ اِسی طرح دوسرے مصرع میں لفظِ آلے پر غور کیجیے۔ نری اِملائی حالت کے نقطۂ نظر سے دیکھیں تو بیک وقت یہ لفظ [آ + لینا (مصدر)] کی ایک تصریفی صورت بھی ہوسکتا ہے اور اور لفظِ آلہ کی اِمالی صورت (Induction) بھی، اور لفظِ آلہ کی جمع بھی۔ پہلے مصرع میں دو میں سے کون سا، اور دوسرے مصرع میں تینوں میں سے کون سا ایک لفظ یہاں پر مراد ہے؟ اِن سوالات کے جوابات کے لیے اردو کے اہلِ علم کے پاس جائے بغیر بات نہیں بنے گی۔ چنانچہ اِس قسم کے مسائل کے حل کے لیے اردو اِطلاعیات کے شعبے کے لوگوں کو علمائے زبانِ اردو کے شانہ بشانہ چلنا ہوگا۔ درست تر الفاظ میں یوں کہا جائے گا کہ اِن لوگوں کو ساتھ لے کر چلنا ہوگا۔ اردو پر بطورِ فطری زبان تحقیق کرنے والوں کو اردو کی رمزیں جاننے والے علما کی ضرورت ہمیشہ باقی رہے گی۔

اردو زبان پر یہ اور اِس جیسے کام کارپس کے بغیر بھی ہوتے رہے ہیں، اور اُن سب کی اپنی اہمیت اور حیثیت بھی تسلیم ہے۔ تاہم اِس بات کے تسلیم کر لینے میں کوئی امر مانع نہیں کہ کارپس کی موجودگی میں یہ کام بہت بہتر اور بہت متنوع — اور بہت وسیع تناظر میں — ہوسکتے ہیں۔ یہ بات بھی واضح ہے کہ یہ کارپس مہیا متون کی مقدار کے پیمانے پر جتنا بڑا ہوگا، لغت نویسوں اور اردو پر بحیثیتِ زبان تحقیق کرنے کے جویاؤں کی اُسی قدر یاوری کر سکے گا۔

مشین ریڈ ایبل اردو میں مہیا ایسا اردو کارپس دنیا بھر کے لسانیاتی محققین کا ایک مشترک خواب ہے جو اُن کے لیے چیلنج، اردو زبان، پر تحقیقات کے لیے سب سے وقیع سروسامان ہوگا۔[۱۵] ڈاکٹر اینڈریو ہارڈی (۲۰۰۳ء) نے بھی اردو کے بارے میں یہی کہا ہے۔

4.4.2: اردو کارپس کے تحقیقی اور لغتیاتی استعمال کی تصویری/ ویب شکل

ذیل میں اردو کے دو مصادر: اترنا/ اتارنا اور چڑھنا کی چند تصریفی صورتوں کو لے کر انتہائی مختصر سے دو جدول نموتنًا دیے جارہے ہیں تاکہ اندازہ کیا جاسکے کہ مجوزہ اردو کارپس کا کمپیوٹر/ اِنٹرنیٹ پر استعمال (Interface) کس طرح سے ہوگا۔ یہ بات ظاہر ہے کہ یہاں پر پیش کیے گئے مثالی جملے، سارے کے سارے، خود سے بنائے گئے ہیں نہ کہ کسی مہیا متن سے لیے گئے ہیں۔ اِن جداوِل کے سرسری جائزے ہی سے معلوم ہوسکتا ہے کہ ایک لفظ کی بدلتی تصریفی شکلوں، مرکبات، سابقوں اور لاحقوں، مقولوں، ضرب الامثال اور مختلف استعمالات میں کیسا تنوع ملتا ہے؛ یہی تنوع اور اِس کا مطالعہ — اور اِس ذخیرۂ امثال اور قاموس الامثال کا ہمہ وقت اور ہر ایک کے لیے موجود اور مہیا ہونا — ہی اِس مجوزہ اردو کارپس کے بارے میں سو باتوں کی ایک بات ہے۔

پہلا مصدر: اترنا/ اتارنا

بچوں نے مل جل کر سارا سامان	اتار	لیا۔
اُس کے سر پر جو بھوت چڑھا ہے اُس کا	اتار	میرے پاس ہے۔
اِس طرح سات سروں کے	اتار	چڑھاؤ سے سولہ سُر قائم کیے گئے۔
ندیوں کے	اتار	چڑھاؤ کی وجہ سے مصنوعی ذرائع آبپاشی بنائے گئے۔
بار بار کی	اترائی	چڑھائی سے ہمارا سانس پھول گیا۔
یہاں سے ایک خطرناک	اترائی	شروع ہوجاتی ہے۔
خانہ بدوشوں کو لباس کے لیے بھیڑ کی	اترن	اون درکار ہوتی ہے۔
لنڈا بازار میں گوروں کی	اترن	بکتی ہے۔
اِس سٹیشن پر کوئی مسافر نہیں	اترا	
قرآن اِس لیے	اترا	ہے کہ اچھی عادتیں اور خصائل سکھائے۔
دس دن سے اُس کے پیٹ میں کوئی دانہ نہیں	اترا	
خواجہ صاحب گھبرا کے گاڑی سے	اترے	اور ایک جانب جا کھڑے ہوئے۔
بالآخر صدر صاحب کرسی سے	اترے	اور قوم نے سکھ کا سانس لیا۔
بچیاں گاڑی سے	اتریں	اور سکول کے اندر چلی گئیں۔
اللہ کی طرف سے برکتیں	اتریں	اور وہ خوش حال ہوگئے۔

بچوں نے ضد کی کہ وہ جھولوں سے نیچے نہیں	اتریں	گے جب تک چھلیاں ختم نہ کرلیں۔
مسافر ریل گاڑی سے	اتر	گئے۔
حجاج کے دل میں یہ بات	اتر	گئی۔
سامنے بیٹھے شخص کا نام میرے ذہن سے	اتر	گیا ہے۔
جس کی اپنی	اتر	گئی اُسے دوسرے کی آبرو سے کیا [واسطہ]۔
اُس کا وہ کولھا پھر سے	اتر	گیا ہے جس کا آپریشن ہوا تھا۔
سارا سامان خشک گودی پر	اتر	گیا۔
روٹیاں تنور سے	اتر	گئیں۔
مریض کی آنکھ میں موتیا	اتر	آیا۔
سکول کے ورائٹی پروگراموں میں اہم لوگوں کی نقلیں	اتاری	جاتی ہیں۔
قائدِ اعظم نے شیروانی	اتاری	اور ہینگر میں لٹکا دی۔
بیٹیوں کے رشتے ہوئے تو اُن کو سر سے بوجھ	اترتا	محسوس ہوا۔
بچہ سیڑھیوں پر چڑھتا اور	اترتا	رہا۔
سامان کو فہرست سے ملا کر قلی کو	اتروائی	دے دیجیے۔
بچوں نے مل کر بھاری میز نیچے	اتروائی	اور پھر کھیل کود میں لگ گئے۔
:		
:		

دوسرا مصدر: چڑھنا

ماموں نے تپائی کھسکائی اور اُسی پر	چڑھ	بیٹھے۔
بچے کی زبان پر اللہ اللہ	چڑھ	گیا۔
استانیوں کی چار چار ماہ کی تنخواہ	چڑھ	گئی تھی۔
بیماری میں نادیہ کی آنکھیں کسی قدر	چڑھ	جایا کرتی تھیں۔
نادیہ پانچویں پاس کر کے چھٹی میں	چڑھ	گئی تو کسی قدر سکون ہوا۔
ماما بہت	چڑھ	چڑھ کر بولتی تھی۔
لڑکے مالٹوں کے باغ میں درختوں پر	چڑھ	دوڑے۔
:		
:		

اِن جداوِل میں صرف مصادر ہی نہیں بلکہ اسما، افعال اور اَعلام وغیرہ کی بنیاد پر بھی مواد سامنے لایا جا سکے گا۔ یہاں تک کہ نرے حروفِ جار کے مطالعے کے لیے بھی مواد سامنے لایا جا سکے گا۔ نیز جداوِل کی ہر سطر (Record)

کے ساتھ یہ معلومات بھی پیش کی جاسکے گی کہ زیرِ بحث لفظ (word under question) گرامر کے اعتبار سے کس حیثیت کا حامل ہے، یعنی، یہ لفظ مصدر ہے، فعل ہے، یا علَم ہے، وغیرہ۔

4.5: اردو کارپس اور انگریزی کارپس میں فرق

جیسا کہ پہلے ذکر کیا گیا، مواد کی جمع آوری کے نقطۂ نظر سے مجوزہ اردو کارپس انگریزی کارپس سے بہت لحاظ سے مختلف ہے۔ اختلاف کے کچھ وجوہ بھی اوپر گزر چکے ذیلی عنوان 4.1 میں ضمناً بیان کردیے گئے ہیں۔ ذیل میں ایک جدول کی شکل میں یہ اِختلافات الفبائی ترتیب میں واضح کیے جارہے ہیں۔

ایک اہم فرق جو انگریزی اور اردو کارپس میں ہے، یہ ہے کہ انگریزی کارپس ہر متن کو اپنے پاس، یعنی اپنے مال خانے میں رکھتے ہیں اور جو بھی کام کراتے ہیں، اِسی متن کی بنیاد پر کراتے ہیں؛ جب کہ مجوزہ اردو کارپس فوری طور پر اردو کے ہر اُس متن کو بھی قبول کرے گا جو پلیٹ فارم سے ناوابستہ، مشین ریڈایبل حالت میں، کہیں سے بھی (Non-repository) مہیا ہوسکے۔

ماخذ	انگریزی کارپس	اردو کارپس
اخبارات	+	+
ادب (نثر)	–	+ (چند منتخب متون)
اِنٹرنیٹ	+	+
انٹرویوز (مصاحبے)	+	-☆
تقاریر (منتخب تقاریر)	+	-☆
ٹیلی وژن (خبریں+ڈرامہ+ٹاک شوز)	+	-☆
ریڈیو (خبریں+ڈرامہ+ٹاک شوز)	+	-☆
شاعری	–	+ (چند منتخب متون)
عام بول چال	+	-☆

(☆ والے نشانات کا مطلب یہ ہے کہ اگر اِن مآخذ سے مواد مشین ریڈایبل اردو میں مل جاتا ہے تو ضرور لیا جائے گا۔)

5: خاتمہ

زبان کی ساخت اور سائنسی اصولوں پر لسانیاتی اور لغتیاتی تحقیق کے لیے وسیع پیمانے پر جمع کیے گئے معاصر استعمالی نظائر کی بنیادی اہمیت ہے۔ اِس وقت مشین ریڈایبل اور پلیٹ فارم سے ناوابستہ حالت میں دنیا کی کئی زبانوں کے متون اِس مقصد کے لیے جمع کیے گئے اور کمپیوٹر/ اِنٹرنیٹ پر محققین کو مہیا کیے گئے ہیں۔ دنیا کی بڑی زبانوں میں جدید لغات بھی اِنھی متون پر کی گئی تحقیقات کی بنیاد پر بنتے ہیں۔ اِس مقالے میں اردو کے روایتی رسم الخط میں اردو کارپس (دی بینک آف اردو/ اردو مثال گھر) کی صورت میں ایسے متون جمع کرنے اور مہیا کرنے کی ضرورت اور اہمیت ذکر کی گئی ہے تاکہ اردو پر بحیثیتِ فطری زبان تحقیق ہو سکے، اور اِس اہم کام کے لیے لائحۂ عمل اردو اِطلاعیات کی زبان میں بتایا گیا ہے۔

تحریر: ۱۳/ جولائی ۲۰۰۸ء، مطابق ۱۰/ رجب المرجب ۱۴۲۹ھ

نظرِ ثانی اور چند اضافے: ۱۱/ نومبر ۲۰۰۸ء

پس نوشت (Postscript):

اِس مقالے کی تیاری میں Collins COBUILD لغت کے پہلے ایڈیشن (1987)، دوسرے ایڈیشن (1995) اور کولن فوقانی لغت یعنی Collins COBUILD English Dictionary for Advanced Learners, Major New Edition (2003) سے آزادانہ استفادہ کیا گیا ہے۔ حواشی میں دیے گئے صفحات کے حوالے اِسی آخرالذکر ماخذ کے ہیں۔ اِس لغت کا چوتھا ایڈیشن بھی شائع ہو چکا ہے۔ مزید معلومات کے لیے رک: www.cobuild.collins.co.uk

مزید مطالعہ:

1. en.wikipedia.org/wiki/COBUILD
2. en.wikipedia.org/wiki/Bank_of_English
3. en.wikipedia.org/wiki/BYU_Corpus_of_American_English
4. corpus.byu.edu/
5. www.americancorpus.org/
6. Dr Sarmad Hussain & Madiha Ijaz: Coupus Based Urdu Lexicon Development, present at: crulp.org/Publication/papers/2007/corpus_based_urdu_lexicon_development.pdf
7. Kashif Riaz: Empirical Stop Word Identification in Urdu Corpora,

present at: irsg.bcs.org/FDIA/2007/fdia2007.php
8. Dara Becker & Kashif Riaz: A Study in Urdu Corpus Construction, present at: acl.ldc.upenn.edu/W/W02/W02-1201.pdf

حوالہ جات

۱۔ احمد، ڈاکٹر حسن الدین (۱۹۷۳ء) ''طریقِ کار اور اصول'' دیباچہ، مشمولہ ''اردو الفاظ شماری'' ، وَلا اکیڈمی، عزیز باغ، سلطان پورہ، حیدرآباد دکن، انڈیا۔ ص ۳۱
۲۔ حقی، شان الحق (۱۹۹۶ء) ''اردو الفاظ میں چھوت چھات'' مقالہ، مشمولہ ''لسانی مسائل و لطائف''، مقتدرہ قومی زبان، اسلام آباد۔ ص ۲۷
۳۔ سرمد حسین، ڈاکٹر و مدیحہ اعجاز (۲۰۰۷ء) ''Corpus Based Urdu Lexicon Development'' مقالہ، مشمولہ CLT07، شعبۂ کمپیوٹر سائنس، پشاور یونیورسٹی۔ ص ۸۷
۴۔ صفوان محمد چوہان، ڈاکٹر حافظ (۲۰۰۷ء) ''اردو لغت (تاریخی اصول پر): بدلتے لسانی تناظر میں چند تجاویز''، مقالہ، مشمولہ ''جرنل آف ریسرچ''، بہاء الدین زکریا یونیورسٹی ملتان، شمارہ-۱۲۔ ص ۲۸۲
۵۔ عبدالمجید، خواجہ (۱۹۳۵ء) ''جامع اللغات کا خاکہ'' اور ''غیر ضروری الفاظ'' مشمولہ ''مقدمہ'' از مؤلف ''جامع اللغات:'' دوسرا اردو سائنس بورڈ ایڈیشن، اردو سائنس بورڈ، لاہور: ۲۰۰۳ء۔ ص ۲۲، ۲۳
۶۔ فاروقی، ڈاکٹر شمس الرحمٰن (۱۹۸۱ء) ''اردو لغات اور لغت نگاری'' مقالہ، مشمولہ ''تنقیدی افکار''، الہ آباد اردو رائٹرز گلڈ، انڈیا۔ پہلا ایڈیشن: ۱۹۸۳ء۔ ص ۱۹۲
۷۔ نارنگ، ڈاکٹر گوپی چند (۱۹۶۸ء) ''اردو زبان کے مطالعے میں لسانیات کی اہمیت'' مقالہ، مشمولہ ''اردو زبان و لسانیات''، رامپور رضا لائبریری، رامپور، انڈیا۔ پہلا ایڈیشن: ۲۰۰۶ء۔ ص ۱-۲۹۰
8. Hardie, Dr Andrew (2003): Unpublished PhD Thesis "The Computational Analysis of Morphosyntactic Categories in Urdu," Lancaster University, UK. P-43

حواشی:

۱۔ اِنٹرنیٹ پر دنیا بھر میں موجود بڑے کارپس ملاحظہ کیجیے: http://corpus.byu.edu/

۲۔ ملاحظہ کیجیے: http://www.worldwidewords.org/topicalwords/tw-cor1.htm

۳۔ ملاحظہ کیجیے: http://www.americancorpus.org/

۴۔ ملاحظہ کیجیے: http://www.cobuild.collins.co.uk

5. Written texts come from newspapers, magazines, fiction and non-fiction books, brochures, leaflets, reports, and letters. (p-xiv)
6. Two-thirds of the corpus is made up of media language: newspapers, magazines, radio and TV. Ibid.
7. Informal spoken language is represented by recordings of everyday casual conversation, meetings, interviews and discussions. Ibid.

۸۔ ''دی بینک آف انگلش'' کی بنیاد پر بننے والا پہلا لغت Collins COBUILD English Dictionary ہے جس کا پہلا ایڈیشن ۱۹۸۷ء میں سامنے آیا۔ سرد خانے میں ڈالنے کا یہ مطلب نہیں ہے کہ اِسے اب استعمال نہیں کیا جارہا، بلکہ مراد یہ ہے کہ اِسے عام طور سے استعمال میں نہیں لایا جارہا اور مخصوص مواقع کے لیے یا عندالطلب رکھا گیا ہے۔

۹۔ ملاحظہ کیجیے: http://leme.library.utoronto.ca/

۱۰۔ ملاحظہ کیجیے: http://www.bckelk.ukfsn.org/menu.html

۱۱۔ ملاحظہ کیجیے:

crulp.org/Downloads/ling_resources/wordlists/UrduHighFreqWords(5000).pdf

۱۲۔ ملاحظہ کیجیے: http://corpus.byu.edu/time/

۱۳۔ یہ بات ڈاکٹر خواجہ محمد زکریا نے مجلسِ ترقیِ ادب لاہور میں منعقدہ ''اردو املاء کے مسائل'' کے عنوان سے ہونے والے ایک اجلاس میں کہی۔ (۲۰۰۷ء)

۱۴۔ اردو کے کسی خالص لفظ کی اردو ہی کے قواعد کے مطابق بنائی گئی جمع جس سے وہ لفظ ایک سے زیادہ مصادر (Infinitives) کی طرف بیک وقت مشیر ہو جائے، کے مثالی جملے کی تلاش میں یہ شعر پروفیسر عابد صدیق کے مجموعۂ کلام ''پانی میں ماہتاب'' سے ملا۔ ملاحظہ کیجیے: ص ۱۵۵۔

۱۵۔ انگریزی ترکیب: Knowledge-base کے لیے مجھے تا حال ''سروسامان'' سے بہتر اردو مترادف نہیں ملا۔

تشکر (Acknowledgement):

[ا] ڈاکٹر حافظ صفوان محمد چوہان برٹش نیشنل کارپس (BNC) اور Collins COBUILD لغات کے روح و رواں، برمنگھم یونیورسٹی برطانیہ کے شعبۂ جدید لسانیات کے پروفیسر John McHardy Sinclair (1933–2007) سے طالب علمانہ رابطے میں رہے ہیں۔ اُنھیں بہت افسوس ہے کہ یہ مقالہ (اردو اور انگریزی دونوں زبانوں میں) اُن کی وفات سے پہلے نہیں لکھا جاسکا۔ تحقیقی مقالے کو کسی کے نام معنوَن کرنے (Dedication) کی روایت نہیں ہے، پھر بھی یہ مقالہ اعترافِ کمال کے طور پر آنجہانی پروفیسر جان میک ہارڈی

سنکلیئر کے نام معنون کیا جاتا ہے۔

[۲] محترمہ قرۃ العین، اسسٹنٹ انفارمیٹکس آفیسر، مرکزِ فضیلت برائے اردو اطلاعیات، مقتدرۂ قومی زبان اسلام آباد نے اِس مقالے کے کچھ ابتدائی نکات ۸/ جون ۲۰۰۸ء کو مقتدرہ میں ایک مشاورتی میٹنگ کے دوران میں نوٹ کیے اور اُنھیں اِن پیج میں لکھوا کر ای میل کیا۔ راقمان الحروف اُن کے شکر گزار ہیں۔ محترمہ قرۃ العین نے ''Scope of Urdu Data House'' کے نام سے خود بھی ایک مقالہ دوسری شیخ ایاز عالمی کانفرنس برائے زبان و ادب (SAICLL) منعقدہ سندھ یونیورسٹی جامشورو (پاکستان) میں مورخہ ۸/ نومبر ۲۰۰۸ء کو پیش کیا۔ اِس مقالے کے مطالعے کی سفارش کی جاتی ہے۔

[۳] قارئین سے گزارش کی جاتی ہے کہ وہ محترمہ مدیحہ اعجاز اور ڈاکٹر سرمد حسین کا انگریزی مقالہ: Corpus Based Urdu Lexicon Development مشمولہ CLT07، شعبۂ کمپیوٹر سائنس، پشاور یونیورسٹی۔ ص-۸۷ ضرور ملاحظہ فرمائیں۔ اردو کارپس کے موضوع پر یہ ایک اہم مقالہ ہے۔ اِسی طرح دارا بیکر (Dara Becker) اور کاشف ریاض کے مقالے: A Study in Urdu Corpus Construction کے مطالعے کی بھی سفارش کی جاتی ہے۔ یہ مقالہ acl.ldc.upenn.edu/W/W02/W02-1201.pdf پر موجود ہے۔ راقمان الحروف اِن دونوں مقالات کے مصنفین کے شکر گزار ہیں۔

مآخذ:

الف: کتابیات

۱۔ احمد، ڈاکٹر حسن الدین، ''اردو الفاظ شماری،'' وَلا اکیڈمی، عزیز باغ، سلطان پورہ، حیدرآباد دکن، انڈیا۔ ۱۹۷۳ء

۲۔ حقی، شان الحق، ''لسانی مسائل و لطائف،'' مقتدرۂ قومی زبان، اسلام آباد۔ ۱۹۹۶ء

۳۔ عابد صدیق، ''پانی میں ماہتاب،'' دوسرا ایڈیشن، الحمد پبلی کیشنز، لاہور۔ ۲۰۰۶ء

۴۔ عبدالمجید، خواجہ، ''جامع اللغات،'' دوسرا اردو سائنس بورڈ ایڈیشن، اردو سائنس بورڈ، لاہور۔ ۲۰۰۳ء

۵۔ عبدالرحیم، ڈاکٹر، ''پردہ اُٹھادوں اگر چہرۂ الفاظ سے،'' دوسرا ایڈیشن، بیت الحکمت، لاہور۔ ۲۰۰۵ء

۶۔ فاروقی، ڈاکٹر شمس الرحمٰن، ''تنقیدی افکار،'' پہلا ایڈیشن، الہ آباد اردو رائٹرز گلڈ، انڈیا۔ ۱۹۸۳ء

۷۔ نارنگ، ڈاکٹر گوپی چند، ''اردو زبان اور لسانیات،'' رامپور رضا لائبریری، رامپور، انڈیا۔ ۲۰۰۶ء

۸۔ وارث سرہندی، ''علمی اردو لغت''، پچیسواں ایڈیشن، علمی کتب خانہ، اردو بازار، لاہور۔۲۰۰۸ء

ب: رسائل اور تحقیقی جرائد

ا۔ جرنل آف ریسرچ، بہاءالدین زکریا یونیورسٹی ملتان۔ شمارہ-۱۲(۲۰۰۷ء)

2. Proceedings of the Conference on Language & Technology (CLT07) at Bara Gali Summer Campus، شعبۂ کمپیوٹر سائنس، پشاور یونیورسٹی (۲۰۰۷ء)

ج: انٹرنیٹ سائٹس (چند منتخب سائٹس)

1. http://leme.library.utoronto.ca/
2. http://www.titania.bham.ac.uk/docs/svenguide.html

د: تکنیکی مشاورت

ا۔ ڈاکٹر خواجہ محمد زکریا، سابق پرنسپل، اورینٹل کالج، جامعۂ پنجاب، لاہور [اردو و انگریزی]

۲۔ ڈاکٹر شمس الرحمن فاروقی، 29/C, Hastings Road, Allahabad-211001، انڈیا [لغت نویسی]

۳۔ ڈاکٹر گوپی چند نارنگ، D-252, Sarvodaya Enclave, New Delhi 110017، انڈیا [اردو لسانیات]

۴۔ خواجہ غلام ربانی مجال، ۲۸-گلستان کالونی، لین نمبر-۲ نیشنل پارک روڈ۔ راول پنڈی [اردو]

۵۔ حافظ محمد اختر ندیم، لیکچرر شعبۂ انگریزی، گورنمنٹ ڈگری کالج، میاں چنوں [انگریزی]

۶۔ راؤ صفدر رشید، ڈیٹا بینک سپروائزر، مرکزِ فضیلت برائے اردو اطلاعیات، مقتدرۂ قومی زبان، اسلام آباد [اردو]

۷۔ وصی اللہ کھوکھر، ایم جی ایچ سالوشنز، کامونکے [اردو لغت نویسی و اردو اِطلاعیات]