ក្រសួងអប់រំ យុវជន និងកីឡា
នាយកដ្ឋានស្រាវជ្រាវនិងនវានុវត្តន៍

OPEN ACCESS

ទស្សនាវដ្តីស្រាវជ្រាវកម្ពុជាសម្រាប់ការអប់រំ និងស្ថេម
Cambodian Journal of Education and STEM

អត្ថបទស្រាវជ្រាវ (Original Article)

ការកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់ក្នុងអត្ថបទជាភាសាខ្មែរ៖ ការវិភាគតាមកម្មវិធី Corpus Linguistics

Identifying High-Frequency Words in Khmer Texts: A Corpus Linguistics Analysis

ខូយ ប៊ុនឡូត* និង អាន ពិសិដ្ឋ
ក្រុមហ៊ុន សង្គប្ប អនុវត្តន៍ រាជធានីភ្នំពេញ ប្រទេសកម្ពុជា
*អ្នកនិពន្ធទទួលបន្ទុកឆ្លើយឆ្លង៖ bunlot89@gmail.com

Bunlot Khoy* and Piseth An
Sangapac Anuwat Company, Phnom Penh, Cambodia
*Corresponding author: bunlot89@gmail.com

https://doi.org/10.62219/cjes.2024214

ទទួលបានអត្ថបទ៖ ២ តុលា ២០២៣	កែសម្រួល៖ ៣១ មករា ២០២៤	យល់ព្រមឱ្យបោះពុម្ព៖ ១២ កុម្ភៈ២០២៤
Received: 2 October 2023	Revised: 31 January 2024	Accepted: 12 February 2024

មូលន័យសង្ខេប

បំណិនអាន ស្តាប់ និយាយ និងសរសេរ ដើរតួនាទីយ៉ាងសំខាន់ក្នុងការប្រាស្រ័យទាក់ទងគ្នារបស់មនុស្ស។ ប្រសិនបើអ្នកសិក្សា ឬគ្រូបង្រៀនមិនចាប់អារម្មណ៍អំពីពាក្យប្រើប្រាស់ញឹកញាប់ដែលជាមូលដ្ឋាន ដើម្បីបំពេញជំនាញទាំងបួននោះទេ ពួកគាត់ទំនងជាចំណាយពេលវេលាច្រើន និងមិនមានភាពងាយស្រួលក្នុងការទទួលបានលទ្ធផលល្អក្នុងការសិក្សា ឬការបង្រៀននោះទេ។ ហេតុនេះ ពាក្យប្រើប្រាស់ញឹកញាប់ដើរតួនាទីសំខាន់ក្នុងការជួយអ្នកសិក្សាឱ្យសម្រេចបានតាមគោលដៅ និងជួយសម្រួលដល់អ្នករៀបចំកម្មវិធីសិក្សា ឬអ្នកបង្កើតកម្មវិធីផ្សេងៗ ដើម្បីបង្កើតកម្មវិធីប្រើប្រាស់បានយ៉ាងងាយស្រួលដល់សាធារណជន។ ការសិក្សានេះមានគោលបំណងកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់ក្នុងអត្ថបទជាភាសាខ្មែរ តាមនិយាមស្តង់ដា ៤០ដង ក្នុង១លានពាក្យ ដែលត្រូវបានវិភាគចេញពីប្រភពទិន្នន័យការសែតកោះសន្តិភាព គេហទំព័រគោះយើងអាន! (បង្កើតដោយអង្គការមូលនិធិអាស៊ី) សៀវភៅប្រជុំរឿងព្រេង សៀវភៅគតិលោក និងសៀវភៅអក្សរសិល្ប៍ខ្មែរកម្រិតថ្នាក់វិទ្យាល័យ។ ដើម្បីវិភាគទិន្នន័យ ការសិក្សានេះប្រើកម្មវិធី AntConc ដែលជាកម្មវិធី Corpus Linguistics ដែលត្រូវបានប្រើប្រាស់ក្នុងការវិភាគកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់។ វចនានុក្រមភាសាខ្មែរ ឆ្នាំ២០២២ ក៏ត្រូវបានប្រើប្រាស់ដើម្បីកំណត់ថ្នាក់ពាក្យនៃពាក្យប្រើប្រាស់ញឹកញាប់ដែលទាញចេញពីនិយាមស្តង់ដា ៤០ដង ក្នុង១លាន

ពាក្យ។ ជាលទ្ធផល ការស្រាវជ្រាវនេះអាចកំណត់បានបញ្ជីពាក្យប្រើប្រាស់ញឹកញាប់ដែលមានចំនួន ១ ៩៧៤ពាក្យ និងថ្នាក់ពាក្យជានាមសព្ទដែលប្រើប្រាស់ច្រើនជាងគេមានចំនួន ១ ០០៨ពាក្យ។ លទ្ធផលស្រាវជ្រាវនេះអាចជួយសម្រួលដល់គ្រូបង្រៀន អ្នករៀបចំកម្មវិធីសិក្សា អង្គការក្រៅរដ្ឋាភិបាល ឬដៃគូពាក់ព័ន្ធ ក្នុងការពិចារណាទៅលើពាក្យប្រើប្រាស់ញឹកញាប់ និងពាក្យនាមសព្ទក្នុងការរៀបចំអត្ថបទអំណាននៅក្នុងកម្រិតមូលដ្ឋានដំបូងៗ។

ពាក្យគន្លឹះ: ពាក្យប្រើប្រាស់ញឹកញាប់ អត្ថបទអំណាន ភាសាខ្មែរ Corpus Linguistics

Abstract

Reading, listening, speaking, and writing skills are essential in human communication. Learners or teachers who do not understand the high-frequency words which are the foundation for understanding the four skills are more likely to spend much time and less likely to get good results. Therefore, high-frequency words play an essential role in helping learners achieve their goals and in helping curriculum designers or developers create applications that are easily accessible to the public. This study aims at identifying the high-frequency words in the standard of NormFreq in Khmer text 40 times per 1 million words analyzed from texts from Koh Santepheap newspapers, Let's Read! (developed by the Asia Foundation), books on a collection of Khmer/Cambodian folktales, a set of Khmer wisdom books, and books on Khmer literature at the high school level. AntConc is a corpus linguistics program used to analyze high-frequency words, and the Khmer dictionary 2022 is used to classify the parts of speech of the most frequent words, drawn from the standard of NormFreq 40 times per 1 million words. As a result, this study identified a list of 1,974 high-frequency words, with nouns being the most commonly part of speech, comprising 1,008 words. These research findings may assist teachers, curriculum developers, NGOs, or relevant partners in considering high-frequency words and nouns when preparing reading texts or materials for basic or elementary levels.

Keywords: High-frequency words; reading texts; Khmer language; corpus linguistics

សេចក្តីផ្តើម

Corpus គឺជាកម្មវិធីវិភាគភាសាដែលត្រូវបានចងក្រង និងប្រមូលអត្ថបទសរសេរ និងនិយាយជាប្រភពនៃកស្តុតាងសម្រាប់ការពិពណ៌នាអំពីលំអាននៃធម្មជាតិ រចនាសម្ព័ន្ធ និងការប្រើប្រាស់ភាសា (Biber et al., 1998)។ ការពិពណ៌នាភាសាដែលផ្អែកលើ Corpus ជាធម្មតា បង្ហាញមិនត្រឹមតែបញ្ជីពាក្យនោះទេ វាថែមទាំងបង្ហាញអំពីលក្ខណៈលម្អិតនៃការប្រើប្រាស់ពាក្យប្រាកដនិយមតាមលក្ខណៈវិទ្យាសាស្ត្រផងដែរ (Kennedy, 2001)។ Francie & Kucera (1979) ក៏បានបញ្ជាក់យ៉ាងច្បាស់ថា Corpus គឺជាកម្មវិធីវិភាគភាសាមានតាំងពីទសវត្សរ៍ ១៩៦០ ដែលមានចំនួន ១លានពាក្យប្រមូលពីប្រភពអត្ថបទចម្រុះ សមាមាត្រគ្នានៅឯសាកលវិទ្យាល័យប្រោន (Brown University)។ រហូតដល់ពាក់កណ្តាលសតវត្សរ៍ទី ២០ អ្នកឯកទេសអប់រំ និងអ្នកភាសាវិទូនៅសហរដ្ឋអាម៉េរិកក៏បានប្រើប្រាស់ Corpus ដែលប្រមូលពាក្យបានចំនួន ១៨លាន

ពាក្យ ដើម្បីរកពាក្យប្រើប្រាស់ញឹកញាប់ជាងគេនៅក្នុងភាសាអង់គ្លេស និងដើម្បីបង្កើតកម្មវិធីសិក្សាឱ្យកាន់តែប្រសើរឡើង សម្រាប់ការកែលម្អការអប់រំអក្ខរកម្ម (Francis & Kucera, 1979)។ រយៈពេល ៣៥ឆ្នាំក្រោយមកទៀត Corpus នេះមានភាពល្បីសុសសាយដល់សហភាពអឺរ៉ុប (Kennedy, 2001)។

ការអនុវត្ត Corpus នៅក្នុងវិស័យភាសាវិទ្យាបានរីកចម្រើនយ៉ាងឆាប់រហ័សក្នុងរយៈពេលជាច្រើនទសវត្សរ៍មកនេះ។ Tognini-Bonelli (2001) បានអះអាងថា Corpus Linguistics ត្រូវបានអ្នកស្រាវជ្រាវជាច្រើនឯកភាពថា ជាវិធីសាស្ត្រស្រាវជ្រាវបែបវិទ្យាសាស្ត្រសុទ្ធសាធ ពីព្រោះភាសាវិទូបានបញ្ចូលនូវសកម្មភាពនៃការប្រមូលទិន្នន័យ និងប្រែក្លាយទិន្នន័យពីគុណវិស័យទៅជាបរិមាណវិស័យ ដើម្បីកំណត់ពាក្យឱ្យអ្នកសិក្សាបានច្បាស់លាស់។ លើសពីនេះ Weisser (2016) បានបង្ហាញថា Corpus Linguistics គឺជាវិធីសាស្ត្រដែលផ្តល់នូវការអភិវឌ្ឍនៃការយល់ដឹងអំពីភាសាដែលមានប្រភពទិន្នន័យពិតប្រាកដ ដើម្បីឱ្យអ្នកសិក្សារៀនអំពីរបៀបស្រាវជ្រាវដោយមិនផ្អែកលើមូលដ្ឋានទ្រឹស្តីតែមួយគត់នោះទេ។ Wiegand & Mahlberg (2019) ក៏បានពន្យល់ថា Corpus Linguistics មិនត្រឹមតែជាវិធីសាស្ត្រសម្រាប់ពិពណ៌នាលក្ខណៈភាសាផ្សេងៗ និងគ្រប់បរិបទទាំងអស់នោះទេ ប៉ុន្តែវាថែមទាំងមានលក្ខណៈវិទ្យាសាស្ត្រផងដែរ។ លើសពីនេះទៀត Reppen (2009) បានកំណត់អំពីអត្ថប្រយោជន៍នៃការប្រើប្រាស់ Corpus Linguistics ដែលមានប្រភពទិន្នន័យគ្រប់បរិបទ និងងាយស្រួលទាញយកមករៀបចំអត្ថបទជាភាសាអង់គ្លេសឱ្យត្រូវនឹងគោលបំណងរបស់គ្រូបង្រៀន និងកំណត់បានច្បាស់ទាំងថ្នាក់ពាក្យ ល្បះឃ្លា និងពាក្យ។

ដោយសារអត្ថប្រយោជន៍ និងគុណតម្លៃនៃ Corpus Linguistics ត្រូវបានគាំទ្រ និងផ្តល់ទំនុកចិត្តខ្ពស់ដោយភាសាវិទូជាច្រើន (Reppen, 2009; Tognini-Bonelli, 2001; Weisser, 2016; Wiegand & Mahlberg, 2019) កម្មវិធីនេះអាចយកមកវិភាគអត្ថបទជាភាសាខ្មែរបាន ដើម្បីកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់។ លើសពីនេះ ពាក្យប្រើប្រាស់ញឹកញាប់ក្នុងភាសាខ្មែរហាក់បីដូចពុំមានអ្នកស្រាវជ្រាវចាប់អារម្មណ៍ក្នុងការសិក្សាលើអត្ថបទជាភាសាខ្មែរ និងពុំមានកម្មវិធី Corpus Linguistics ប្រើប្រាស់ ដើម្បីបង្កើតបញ្ជីពាក្យប្រើប្រាស់ញឹកញាប់នៅឡើយទេ។ ទោះយ៉ាងណាក៏ដោយ Khoy (2021) បានសិក្សាអំពីពាក្យប្រើប្រាស់នៅកម្រិតថ្នាក់មូលដ្ឋានដោយប្រមូលពាក្យត្រួតគ្នាប្រហែល ៧០០ ០០០ពាក្យក្នុងការសិក្សា និងបានបង្ហាញលទ្ធផលថាកម្រិតថ្នាក់មូលដ្ឋានគួរប្រើប្រាស់ពាក្យចំនួន ៩ ០០០ពាក្យមិនត្រួតគ្នា និងមួយឆ្នាំសិក្សាត្រូវរៀនពាក្យថ្មីបន្ថែមចំនួន ១ ០០០ពាក្យ រាប់បញ្ចូលតាំងពីថ្នាក់ទី១ ដល់ថ្នាក់ទី៩។ បើពិនិត្យឱ្យកាន់តែច្បាស់បន្ថែមទៀត Khoy (2021) បានផ្តល់ជាទុន ឬតម្រុយសម្រាប់អ្នកសិក្សាស្រាវជ្រាវក្រោយៗពិចារណាលើពាក្យប្រើប្រាស់កម្រិតមូលដ្ឋានក្នុងការសរសេរអត្ថបទ ប៉ុន្តែពុំមានបង្ហាញចំនួននិយាមស្តង់ដាច្បាស់លាស់ក្នុង១លានពាក្យ ដើម្បីគាំទ្រការស្រាវជ្រាវនោះទេ។ ដោយភាពខ្វះខាតក្នុងការស្រាវជ្រាវបែបនេះ ការស្រាវជ្រាវនេះមានគោលបំណងប្រើប្រាស់កម្មវិធី Corpus Linguistics ដើម្បីកំណត់ពាក្យប្រើញឹកញាប់នៅក្នុងភាសាខ្មែរ។ ដូច្នេះ ការស្រាវជ្រាវនេះប្រើប្រាស់កម្មវិធី AntConc Corpus Linguistics (Anthony, 2022) ក្នុងការកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់ដែលទាញចេញពីប្រភពទិន្នន័យអត្ថបទសរសេររបស់កាសែត និងសៀវភៅជាភាសាខ្មែរ។

ក៏ប៉ុន្តែ ពាក្យប្រើប្រាស់ញឹកញាប់ជាងគេត្រូវបានកំណត់ទៅតាមទម្រង់ផ្សេងៗគ្នា។ Lynn (1973) បានសម្គាល់ពាក្យប្រើប្រាស់ញឹកញាប់ជាងគេ ជាពាក្យដែលមានចំនួនដងលើសពី ៣៩ដងនៅក្នុងចំណោមពាក្យដែលមាននៅក្នុងទិន្នន័យរបស់ Corpus Linguistics។ Dang et al. (2017) បានកំណត់ថាពាក្យប្រើប្រាស់ញឹកញាប់ជាងគេ គឺជាពាក្យដែលមានចំនួនច្រើនជាង ៧០% គ្របដណ្តប់លើអត្ថបទអំណាន។ Nation (2012) បានបញ្ជាក់ថាពាក្យចំនួន ២ ០០០ពាក្យនៅ

ក្នុងភាសាអង់គ្លេសត្រូវបានចាត់ទុកជាពាក្យប្រើប្រាស់ញឹកញាប់ជាងគេ។ Van Zeeland & Schmitt (2013) បានបង្ហាញថាពាក្យប្រើប្រាស់ញឹកញាប់ជាងគេមានប្រហែល ២ ០០០ ដល់ ៣ ០០០ពាក្យ ដែលមានលក្ខណៈគ្រប់គ្រាន់ ដើម្បីយល់ពីការនិយាយគ្នាប្រចាំថ្ងៃ និងអានអត្ថបទរឿងនិទានជាដើម។ Laufer & Ravenhorst-Kalovski (2010) ក៏បានអះអាងថាចំនួនពាក្យប្រហែលពី ៣ ០០០ ទៅ ៥ ០០០ពាក្យប្រើប្រាស់ញឹកញាប់ជាងគេ គឺគ្រប់គ្រាន់សម្រាប់អានអត្ថបទ និងមើលភាពយន្តជាភាសាអង់គ្លេស។ ការលើកអំណះអំណាងរបស់អ្នកស្រាវជ្រាវខាងលើ បានផ្តល់ជាភស្តុតាងជាក់លាក់មួយអំពីពាក្យប្រើប្រាស់ញឹកញាប់ ក៏ប៉ុន្តែពាក្យប្រើប្រាស់ញឹកញាប់ហាក់បីដូចមិនទាន់ជាទម្លឹកឧណ៌មូលដ្ឋានគ្រប់គ្រាន់សម្រាប់ការស្រាវជ្រាវនេះទេ។

ដើម្បីមានមូលដ្ឋានគ្រប់គ្រាន់ក្នុងការកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់ ការស្រាវជ្រាវនេះពឹងផ្អែកទៅលើនិយាមស្តង់ដា (Normed Frequency ឬ NormFreq) ជាចំនួនពាក្យនៅក្នុងចំនួន ១លានពាក្យ។ Bestgen (2020) បានសិក្សាអំពីការប្រៀបធៀបពាក្យដោយការប្រើប្រាស់ Corpus Linguistics ផ្សេងៗគ្នា ដើម្បីកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា ២០ដងក្នុង ១លានពាក្យ។ Francis et al. (1982) បានបញ្ជាក់ថា Corpus Linguistics ដែលមានពាក្យត្រួតគ្នាចំនួនប្រហែល ៣,៥លានពាក្យ ត្រូវកំណត់យកពាក្យប្រើប្រាស់ញឹកញាប់ចំនួន ១០០ដង និងមាននិយាមស្តង់ដាយ៉ាងហោចណាស់ក៏ចំនួន ២៥ដងក្នុង ១លានពាក្យផងដែរ។ យោងតាម Biber & Barbieri (2007) ពាក្យដែលមាននៅក្នុង Corpus Linguistics និងស្ថិតនៅក្នុងនិយាមស្តង់ដា ៤០ដងឡើងទៅ ត្រូវបានកំណត់ថាជាពាក្យប្រើប្រាស់ញឹកញាប់ ទោះបីពាក្យទាំងនោះ មានប៉ុន្មានក៏ដោយ។ Chen & Baker (2016) បានគូសបញ្ជាក់ថាពាក្យប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា ៤៥ដងក្នុង ១លានពាក្យទើបអាចយកជាផ្លូវការក្នុងការកំណត់ពាក្យនោះបាន។ Gardener & Davies (2013) បានពន្យល់ថា ដើម្បីឱ្យមានលក្ខខណ្ឌនៃការប្រើប្រាស់ពាក្យញឹកញាប់ លុះត្រាពាក្យនោះកើតឡើងនៅក្នុងនិយាមស្តង់ដា យ៉ាងតិចបំផុតក៏ ៥០ដងក្នុង ១លានពាក្យដែរ។ ការលើកឡើងរបស់អ្នកស្រាវជ្រាវមុនៗ បានបង្ហាញពីលក្ខណៈវិនិច្ឆ័យលើពាក្យប្រើប្រាស់ញឹកញាប់ ដើម្បីធ្វើការសន្និដ្ឋាន និងបញ្ជ្រាបពាក្យទាំងនោះទៅក្នុងអត្ថបទអំណានក្នុងកម្រិតមូលដ្ឋាន។ អ្នកស្រាវជ្រាវទាំងនោះ បានផ្តល់អំណះអំណាង និងភស្តុតាងជាក់លាក់មួយ ដែលអាចគាំទ្រការសិក្សាដែលប្រើប្រាស់លក្ខខណ្ឌសមស្របទៅនឹងបរិបទ និងពាក្យក្នុងកម្មវិធី Corpus Linguistics ជាភាសាខ្មែរ។ ដូច្នេះ គោលបំណងនៃការស្រាវជ្រាវនេះ គឺសាកល្បងកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ ដែលទាញចេញពីកម្មវិធី AntConc Corpus Linguistics។

ដូចបានរៀបរាប់ខាងលើ អ្នកស្រាវជ្រាវជាច្រើន (Reppen, 2009; Tognini-Bonelli, 2001; Weisser, 2016; Wiegand & Mahlberg, 2019) បានអះអាងថា Corpus Linguistics ជាឧបករណ៍វិភាគភាសាបែបវិទ្យាសាស្ត្រ និងមានសុក្រឹតភាពខ្ពស់ ដើម្បីកំណត់រកពាក្យ ឃ្លា ល្បះ ទៅតាមបរិបទ និងគោលបំណងរបស់ការស្រាវជ្រាវ។ ដូច្នេះ ការស្រាវជ្រាវនេះឃើញពីផលប្រយោជន៍នៃកម្មវិធី AntConc Corpus Linguistics ដែលអាចវិភាគអត្ថបទជាភាសាខ្មែរបានដូចភាសាអង់គ្លេសដែរ។ ម្យ៉ាងវិញទៀត ភាសាខ្មែរមានអាយុកាលរាប់ពាន់ឆ្នាំមកហើយ ក៏ប៉ុន្តែភាសានេះ ពុំទាន់មានពាក្យប្រើប្រាស់ញឹកញាប់ដែលទាញចេញពី Corpus Linguistics និងកំណត់ចំនួនតាមនិយាមស្តង់ដាក្នុង ១លានពាក្យនៅឡើយ។ ដើម្បីធ្វើការសន្និដ្ឋានតាមក្បួនវិទ្យាសាស្ត្រ ការស្រាវជ្រាវនេះបានបំផុសសំណួរចំនួនពីរ ដើម្បីធ្វើការសិក្សា៖

១. តើពាក្យប្រើប្រាស់ញឹកញាប់ក្នុងភាសាខ្មែរមានចំនួនប៉ុន្មាន បើយោងតាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យដែលទាញចេញពីកម្មវិធី AntConc Corpus Linguistics?

២. តើថ្នាក់ពាក្យមួយណាដែលត្រូវបានប្រើប្រាស់ច្រើនជាងគេ បើយោងតាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ?

ទិធីសាស្ត្រស្រាវជ្រាវ

ប្រភពទិន្នន័យនៃការស្រាវជ្រាវ

ប្រភពទិន្នន័យសម្រាប់ការស្រាវជ្រាវនេះ បានដកស្រង់ចេញពីកាសែតកោះសន្តិភាព (www.kohsantepheapdaily.com.kh) អត្ថបទនៅក្នុងគេហទំព័រ តោះយើងអាន! (www.letsreadasia.org) សៀវភៅប្រជុំរឿងព្រេង សៀវភៅគតិលោក និង សៀវភៅអក្សរសិល្ប៍ខ្មែរកម្រិតថ្នាក់វិទ្យាល័យ (ថ្នាក់ទី១០ ដល់ទី១២)។ អត្ថបទអំណានត្រូវបានប្រមូលចេញពីគេហទំព័រ កាសែតកោះសន្តិភាពចំនួន ១ខែ (ខែកុម្ភៈ ឆ្នាំ២០២៣) ដើម្បីយកមកសិក្សានៅក្នុងការស្រាវជ្រាវនេះ ពីព្រោះកាសែតនេះជា កាសែតដែលមានភាពវ័យចំណាស់មួយនៅក្នុងប្រទេសកម្ពុជា និងមានប្រជាប្រិយសម្រាប់អ្នកអាន (Khoy et al., 2021)។ បន្ថែមពីនេះទៀត អត្ថបទនៅក្នុងគេហទំព័រ តោះយើងអាន! (Let's Read, 2016) ដែលមានចំនួនអត្ថបទពីកម្រិត0 ដល់ កម្រិត៥ ដែលបានផ្សព្វផ្សាយដល់អ្នកអានគ្រប់កម្រិត និងឥតគិតថ្លៃ ក៏ត្រូវបានដកស្រង់ជាទិន្នន័យសម្រាប់ការស្រាវស្រាវ នេះផងដែរ។ បន្ថែមពីនេះទៅទៀត ការស្រាវជ្រាវនេះ ក៏បានពឹងផ្អែកលើប្រភពទិន្នន័យយកចេញពីសៀវភៅប្រជុំរឿងព្រេង ខ្មែរដែលមានចំនួន ៩ភាគ សៀវភៅគតិលោកចំនួន ១០ភាគ ដែលជាសៀវភៅមានប្រជាប្រិយភាពសម្រាប់ប្រជាពលរដ្ឋខ្មែរ តាំងពីចុងសតវត្សរ៍ទី២០ រហូតដល់បច្ចុប្បន្ន ដោយផ្នែកខ្លះនៃសៀវភៅទាំងពីរប្រភេទនេះ ក៏ត្រូវបានក្រសួងអប់រំ យុវជន និង កីឡាដកស្រង់អត្ថបទខ្លះដាក់នៅក្នុងកម្មវិធីសិក្សាភាសាខ្មែរ។ ប្រភេទទិន្នន័យចុងក្រោយគឺអត្ថបទអក្សរសិល្ប៍ខ្មែរ (រឿង ប្រលោមលោក) កម្រិតថ្នាក់វិទ្យាល័យចំនួន ៤ក្បាល ដែលមានថ្នាក់ទី១០ ចំនួន១ក្បាល ថ្នាក់ទី១១ ចំនួន១ក្បាល និងថ្នាក់ ទី១២ ចំនួន២ក្បាល។ រឿងប្រលោមលោកទាំងនោះ មានចំណងជើងដូចជា៖ (ក) រឿងកុលាបប៉ៃលិន (ខ) រឿងចៅកែចិត្ត ចោរ (គ) រឿងផ្កាស្រពោន និង (ឃ) រឿងមហាចោរទល់ដែន។

នីតិវិធីក្នុងការប្រមូលទិន្នន័យ

ចំពោះការប្រមូលទិន្នន័យពីប្រភពកាសែតកោះសន្តិភាពអនឡាញ គឺមានរយៈពេល ១ខែ ដែលមានចំណងជើងរងេ ដូចជា នយោបាយ ជំនឿនិងសាសនា ជីវិតនិងសង្គម កម្សាន្ត បច្ចេកវិទ្យា កីឡា សុខភាព និងព័ត៌មានសំខាន់ៗ។ ក្នុងរយៈ ពេល ១ខែ អត្ថបទនៃកាសែតកោះសន្តិភាពដែលត្រូវបានដកស្រង់ចេញមកមានចំនួន ៩០ ០០០ចំណងជើង និងចំនួន ២៣ ២២៥ ៧៧៩ពាក្យ។ ចំពោះសៀវភៅប្រជុំរឿងព្រេង មានចំនួន ៩ភាគដែលមានចំនួន ២៤៨ចំណងជើង និងមានចំនួន ២៦៦ ៧៦១ពាក្យ។ សៀវភៅគតិលោកមានចំនួន ១០ភាគ ដែលមានចំនួន ២៤៨ចំណងជើង និងមានចំនួន ៨៦ ៧៥២ ពាក្យ។ អត្ថបទនៅក្នុងគេហទំព័រ តោះយើងអាន! មានចំនួន ៥២៤ចំណងជើង និងមានចំនួន ១៧៣ ៥០៣ពាក្យ និង សៀវភៅអក្សរសិល្ប៍ខ្មែរកម្រិតថ្នាក់វិទ្យាល័យមានចំនួន ៤រឿងប្រលោមលោកផ្សេងៗគ្នា និងមានចំនួន ៥៥ ៤០៥ពាក្យ។ អត្ថបទដែលបានដកស្រង់ចេញពីប្រភពទាំងនោះ ត្រូវបានបម្លែងជាទម្រង់ File Notepad ++ រួចបំប្លែង (Encode) ជា UTF-8-BOM ដើម្បីយកមកញែកពាក្យនៅក្នុងកម្មវិធី PAN Khmer Line Breaking ។ បន្ទាប់មកទៀត អ្នកស្រាវជ្រាវបាន ដកស្រង់អត្ថបទអំណានចេញពីគេហទំព័រ តោះយើងអាន ! (www.letsreadasia.org) ជាទម្រង់ File EPUB រួចបម្លែងវា ទៅជាទម្រង់ File Notepad ++ ដែលមានចំនួន ៥២៤អត្ថបទអំណាន។ ក្រោយពីប្រមូលអត្ថបទទាំងអស់បានរួចរាល់ហើយ អ្នកស្រាវជ្រាវបានយកអត្ថបទទាំងអស់នោះ ទៅញែកជាពាក្យដាច់ដោយឡែកតាមមេពាក្យ (ឧទាហរណ៍៖ មាន) និងអនុ ពាក្យ (ឧទាហរណ៍៖ មានមុខ មានកម្ម មានកូន) របស់វចនានុក្រមសម្តេចសង្ឃរាជ ជួន ណាត វចនានុក្រមខ្មែរ ២០២២ និង

ឈ្មោះភូមិ ឃុំ/សង្កាត់ ស្រុក/ខណ្ឌ និងខេត្ត/ក្រុង ក្នុងព្រះរាជាណាចក្រកម្ពុជា ដែលមានមេពាក្យសរុបប្រហែល ៧០ ០០០ ពាក្យ ដើម្បីធានានូវសេរណាភាព និងសុក្រឹតភាពនៃគោលពាក្យទាំងអស់នោះ។

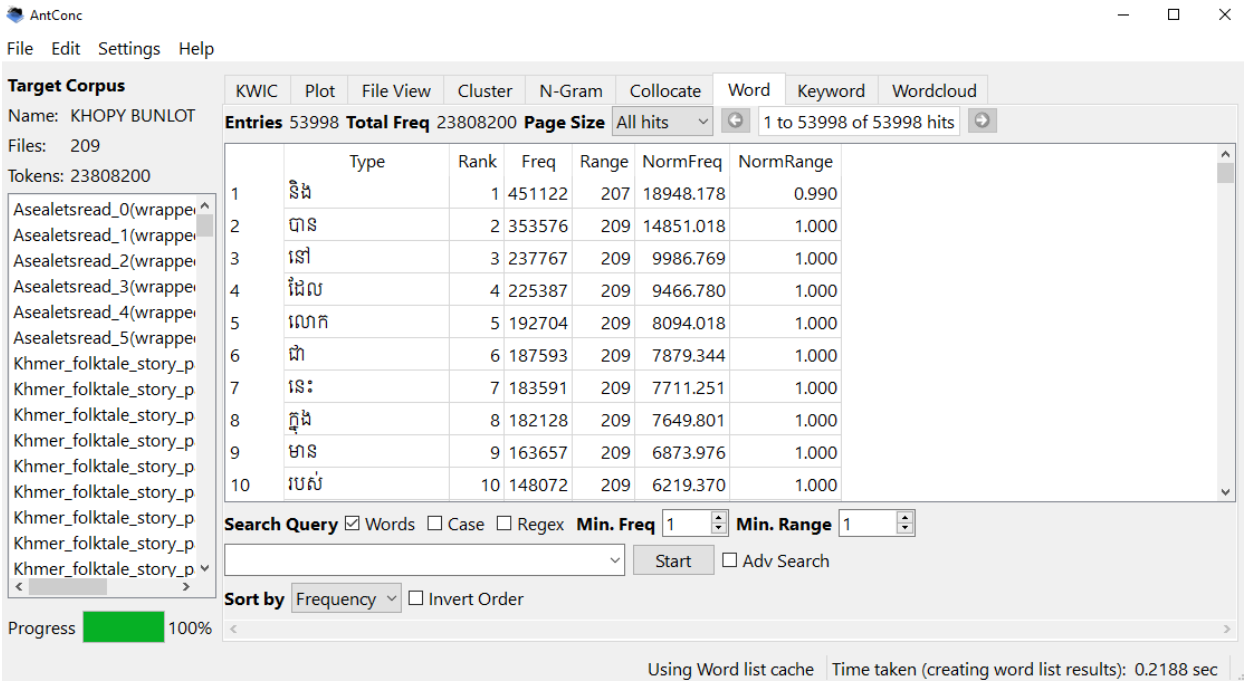
ឧបករណ៍ស្រាវជ្រាវ និងការវិភាគទិន្នន័យ

បន្ទាប់ពីញែកពាក្យចេញពីប្រភពអត្ថបទទាំងអស់មក អ្នកស្រាវជ្រាវបានប្រើប្រាស់ឧបករណ៍វិភាគទិន្នន័យដែលមានឈ្មោះថា AntConc (Version 4.2.0) ដើម្បីបង្កើត Corpus Databases ជាភាសាខ្មែរ (Anthony, 2022)។ បន្ទាប់មក អ្នកស្រាវជ្រាវបានធ្វើងផ្ទាត់ថ្នាក់ពាក្យដែលទាញចេញពីវចនានុក្រមភាសាខ្មែរ ដើម្បីកំណត់ថ្នាក់ពាក្យមួយណាដែលត្រូវបានប្រើប្រាស់ច្រើនជាងគេ តាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ (Biber & Barbieri, 2007)។ ដើម្បីឆ្លើយតបទៅនឹងសំណួរស្រាវជ្រាវទី១ អ្នកស្រាវជ្រាវបានយកទិន្នន័យជាអត្ថបទទាំងអស់ដាក់ចូលនៅក្នុងកម្មវិធី AntConc រួចបង្កើតជា Corpus Databases ជាភាសាខ្មែរ ដើម្បីវិភាគរកពាក្យប្រើប្រាស់ញឹកញាប់ចេញពីអត្ថបទដែលប្រមូលបានពីប្រភពទិន្នន័យទាំងអស់នោះដោយកំណត់ទៅតាមលក្ខខណ្ឌ និងនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ។ ប្រសិនបើពាក្យទាំងអស់នោះមិនស្ថិតនៅតាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យទេ អ្នកស្រាវជ្រាវត្រូវលុបពាក្យទាំងនោះចេញពីការវិភាគ។ បន្ទាប់មក អ្នកស្រាវជ្រាវបានជ្រើសរើសពាក្យដែលមាននៅក្នុងកម្មវិធី Corpus Databases ៤០ដងក្នុង ១លានពាក្យ យកមកសិក្សា ដើម្បីកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់ ដែលទាញចេញពីប្រភពទិន្នន័យអត្ថបទអំណាន ដូចដែលបានរៀបរាប់ខាងលើទាំងអស់។ ចំពោះសំណួរស្រាវជ្រាវទី២ អ្នកស្រាវជ្រាវបានធ្វើងផ្ទាត់ពាក្យប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ ជាមួយវចនានុក្រមភាសាខ្មែរដែលមានថ្នាក់ពាក្យបានកំណត់រួចរាល់។ ប៉ុន្តែដោយពាក្យខ្លះមានថ្នាក់ពាក្យច្រើនជាងមួយ អ្នកស្រាវជ្រាវបានកំណត់ថ្នាក់ពាក្យនោះតាមថ្នាក់ពាក្យដែលត្រូវមានច្រើនជាងគេក្នុង Corpus Databases ជាភាសាខ្មែរ។ ជាឧទាហរណ៍ ពាក្យ «ជា» មានថ្នាក់ពាក្យ ជានាមសព្ទ គុណនាម និងកិរិយាសព្ទ។ ការកំណត់ថ្នាក់ពាក្យនៃពាក្យ «ជា» ក្នុង Corpus Databases ជាកិរិយាសព្ទ ដោយផ្អែកលើនិយមន័យរបស់ពាក្យ «ជា» នៅក្នុងវចនានុក្រមខ្មែរ ២០២២ ដែលជាវចនានុក្រមអេឡិចត្រូនិក ដែលត្រូវបានដាក់ឱ្យប្រើប្រាស់បើកចំហជំនាន់ទី ២.០ កាលពីចុងឆ្នាំ២០២៣។ នេះជាសម្រង់សេចក្តីផ្តាល់របស់ពាក្យ «ជា» ដែលត្រូវបានពន្យល់នៅក្នុងវចនានុក្រមខ្មែរ ២០២២៖ «ពាក្យសម្រាប់និយាយចង្អុលនាមសព្ទឱ្យដាច់សេចក្តីដោយឡែក (ឧទាហរណ៍៖ វិជ្ជាជាទ្រព្យដ៏ប្រសើរក្នុងលោក, ឈ្មោះនេះជាមនុស្សស្លូតត្រង់)» (Royal Academy of Cambodia, 2023)។

លទ្ធផលស្រាវជ្រាវ និងការពិភាក្សា

ពាក្យប្រើប្រាស់ញឹកញាប់បំផុត

បន្ទាប់ពីប្រមូលទិន្នន័យដែលបានពីប្រភពអត្ថបទទាំង៥រួច អ្នកស្រាវជ្រាវបានយកទិន្នន័យទាំងនោះមកវិភាគរកពាក្យប្រើប្រាស់ញឹកញាប់ជាភាសាខ្មែរដោយកំណត់តាមលក្ខខណ្ឌ និងនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ។ មុននឹងបង្ហាញលទ្ធផលស្រាវជ្រាវ អ្នកស្រាវជ្រាវសូមបង្ហាញទិន្នន័យមូលដ្ឋាន អំពីប្រភពអត្ថបទដែលបានដាក់បញ្ចូលនៅក្នុងកម្មវិធី AntConc Corpus Databases ដើម្បីវិភាគរកពាក្យប្រើប្រាស់ញឹកញាប់ជាភាសាខ្មែរ។



	Type	Rank	Freq	Range	NormFreq	NormRange
1	និង	1	451122	207	18948.178	0.990
2	បាន	2	353576	209	14851.018	1.000
3	នៅ	3	237767	209	9986.769	1.000
4	ដែល	4	225387	209	9466.780	1.000
5	លោក	5	192704	209	8094.018	1.000
6	ជា	6	187593	209	7879.344	1.000
7	នេះ	7	183591	209	7711.251	1.000
8	ក្នុង	8	182128	209	7649.801	1.000
9	មាន	9	163657	209	6873.976	1.000
10	របស់	10	148072	209	6219.370	1.000

រូបភាពទី១៖ ប្រភពទិន្នន័យនៅក្នុងកម្មវិធី AntConc Corpus Databases ដែលផ្តោតលើពាក្យប្រើប្រាស់ញឹកញាប់

រូបភាពទី១ បង្ហាញថាមានប្រភពទិន្នន័យជាឯកសារចំនួន ២០៩ឯកសារ (Files) ដែលមានពាក្យត្រួតគ្នា (Tokens) ចំនួន ២៣ ៨០៨ ២០០ពាក្យ និងពាក្យមិនត្រួតគ្នា (Entries) ចំនួន ៥៣ ៩៩៨ពាក្យ។ ចំពោះពាក្យប្រើញឹកញាប់នៅក្នុងប្រភពទិន្នន័យទាំងអស់ដែលប្រមូលបានពីអត្ថបទទាំង៥ប្រភព មានពាក្យ «និង» ឈរលេខរៀង (Rank) ទី១ ដែលមានពាក្យប្រើប្រាស់ញឹកញាប់ (Freq) ចំនួន ៤៥១ ១២២ដង ក្នុងចំណោមពាក្យជិត ២៤លានពាក្យ និងមាននិយាមស្តង់ដា (NormFreq) ចំនួន ១៨ ៩៤៨ដង ក្នុងចំនួន ១លានពាក្យ។ បន្ទាប់មក មានពាក្យ «បាន» «នៅ» «ដែល» «លោក» «ជា» «នេះ» «ក្នុង» «មាន» និង «របស់» ដែលមានលំដាប់រៀងគ្នាតាមលំដាប់ និងមាននិយាមស្តង់ដា ចន្លោះពី ៦ ៨៧២ដង ដល់ ១៤ ៨៥១ដង ក្នុង ១លានពាក្យ។ រីឯ ពាក្យចុងក្រោយ គឺពាក្យ «របស់» ដែលត្រូវបានប្រើប្រាស់ញឹកញាប់រហូតដល់ ១៤៨ ០៧២ដង ក្នុងចំណោមជិត ២៤លានពាក្យ និងនិយាមស្តង់ដាមានចំនួន ៦ ២១៩ដង ក្នុង ១លានពាក្យ។ សរុបមក កម្មវិធី AnCont Corpus Databases បានបង្ហាញអំពីចំនួនពាក្យប្រើប្រាស់ញឹកញាប់ចំនួន ១០ពាក្យ ក្នុងចំណោមពាក្យជិត ២៤លានពាក្យ និងមាននិយាមស្តង់ដាលើសពី ៦ ០០០ដង ក្នុង ១លានពាក្យ។ ដូច្នេះ ពាក្យទាំងនេះឆ្លុះបញ្ចាំងអំពីពាក្យប្រើប្រាស់ញឹកញាប់ជាភាសាខ្មែរក្នុងចំណោមពាក្យជិត ២៤លានពាក្យ ដែលអាចកំណត់អត្តសញ្ញាណនៃពាក្យខ្មែរ និងបង្ហាញអំពីពាក្យប្រាកដប្រជាដើម្បីឱ្យអ្នកអាន អ្នកសិក្សា អ្នកបង្កើតកម្មវិធីផ្សេងៗ ស្វែងយល់បន្ថែមអំពីគោលពាក្យចម្បងទាំងនេះ ក្នុងការរៀបចំកម្មវិធីសិក្សាណាមួយដែលពួកគាត់មានបំណងចង់អភិវឌ្ឍ។

នាមសព្ទ កិរិយាសព្ទ និងគុណនាមដែលប្រើប្រាស់ញឹកញាប់បំផុត

លទ្ធផលសម្រាប់សំណួរស្រាវជ្រាវទី១ ដែលបានទទួលពីការប្រើប្រាស់ឧបករណ៍ AntConc Corpus Databases ក្នុងការកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ ត្រូវបានបង្ហាញលម្អិតនៅក្នុងតារាងទី១ និងរូបភាពទី២ ខាងក្រោមនេះ៖

តារាងទី១៖ *នាមសព្ទ កិរិយាសព្ទ និងគុណនាម ចំនួន១០ ដែលប្រើប្រាស់ញឹកញាប់បំផុតតាមនិយាមស្តង់ដា*

ល.រ	នាមសព្ទ			កិរិយាសព្ទ			គុណនាម		
	ពាក្យ	ចំនួន	%	ពាក្យ	ចំនួន	%	ពាក្យ	ចំនួន	%
១	លោក	១៩២ ៧០៤	០,៨១	បាន	៣៥៣ ៥៧៦	១,៤៩	នាក់	១២៥ ៤១៩	០,៥៣
២	របស់	១៤៨ ០៧២	០,៦២	នៅ	២៣៧ ៧៦៧	១,០០	មួយ	៨៣ ២៨០	០,៣៥
៣	ឆ្នាំ	១៤២ ៧២៦	០,៦០	មាន	១៦៣ ៦៥៧	០,៦៩	ម្នាក់	៥៧ ៦៤៤	០,២៤
៤	ឈ្មោះ	៨០ ៨៨៣	០,៣៤	គឺ	១៤៨ ០៦៦	០,៦២	ចុងក្រោយ	៥៦ ៩៣២	០,២៤
៥	ទី	៨០ ៥៦៦	០,៣៤	ឱ្យ	១២៦ ៨២២	០,៥៣	ស្មើ	៣៨ ៤៧៤	០,១៦
៦	រថយន្ត	៧៤ ៤២៦	០,៣១	ប្រកួត	១១៦ ៦៩៥	០,៤៩	សម្រាប់	៣៧ ៨៥៥	០,១៦
៧	ផ្ទះ	៧៤ ២៧៨	០,៣១	ដោយ	១១៥ ៥៤៤	០,៤៩	ថ្មី	២២១ ២៧	០,០៩
៨	ចំនួន	៧៤ ០៦៧	០,៣១	ទៅ	៨៦ ៣៥៧	០,៣៦	លាន	១៨ ៥២៩	០,០៨
៩	ប្រទេស	៧២ ៥៣៣	០,៣០	ឈ្នះ	៧៨ ៧៣២	០,៣៣	ធំ	១៦ ៣០៩	០,០៧
១០	អាយុ	៦២ ៣៣២	០,២៦	តាម	៧៦ ១៧៥	០,៣២	ល្អ	១៤ ៥៥៩	០,០៦

សម្គាល់៖ នាមសព្ទចំនួន ១០០៨ពាក្យ កិរិយាសព្ទចំនួន ៦០៥ពាក្យ និងគុណនាមចំនួន ១៦៤ពាក្យ ត្រូវបានចងក្រងដោយអ្នកស្រាវជ្រាវ និងអាចអាន ឬទាញយកបានដោយឥតគិតថ្លៃ តាមតំណ៖ https://shorturl.at/cfj18

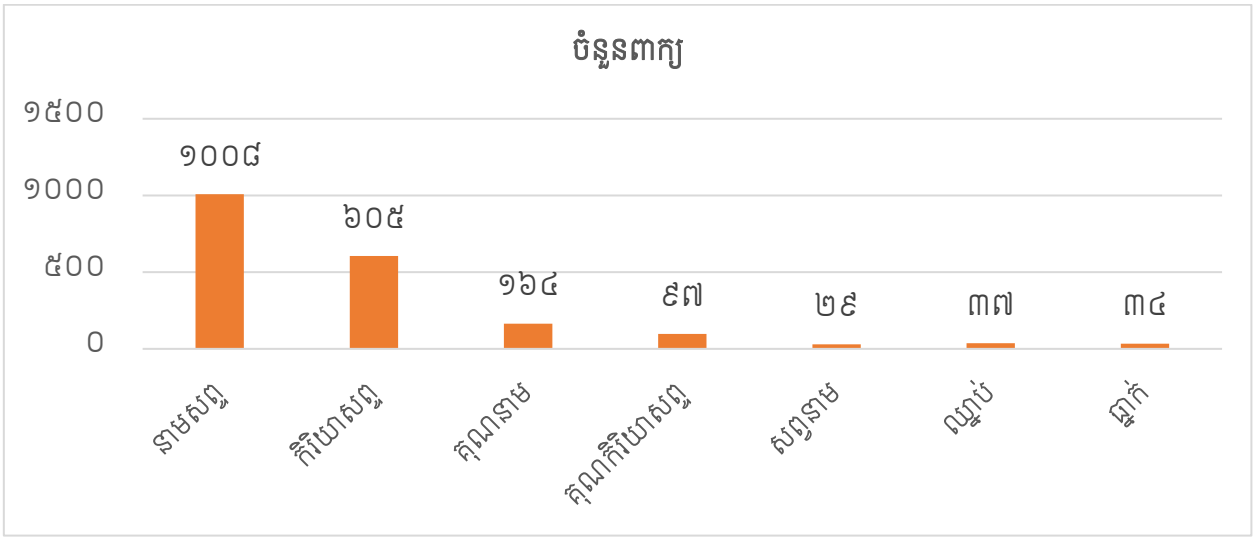
តារាងទី១ បង្ហាញឱ្យឃើញថា នាមសព្ទ កិរិយាសព្ទ និងគុណនាម ចំនួន១០ពាក្យ ត្រូវបានប្រើប្រាស់ញឹកញាប់បំផុតតាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ។ ក្នុងចំណោមថ្នាក់ពាក្យទាំងបីប្រភេទ គេសង្កេតឃើញថាភាគរយច្រើនជាង ០,៥ ដែលស្ថិតនៅក្រុមនាមសព្ទមានចំនួន ៣ពាក្យ គឺ «លោក» «របស់» និង «ឆ្នាំ»។ ពាក្យដែលស្ថិតនៅក្រុមកិរិយាសព្ទមានចំនួន ៥ពាក្យ គឺ «បាន» «នៅ» «មាន» «គឺ» និង «ឱ្យ» និងពាក្យដែលស្ថិតនៅក្រុមគុណនាមមានចំនួន ១ពាក្យ គឺ «នាក់»។ នាមសព្ទដែលត្រូវបានគេប្រើប្រាស់ច្រើនជាងគេ គឺពាក្យ «លោក» ដែលមានចំនួន ១៩២ ៧០៤ពាក្យ ស្មើនឹង ០,៨១% ខណៈនាមសព្ទដែលត្រូវបានគេប្រើប្រាស់ច្រើនជាងគេក្នុងកម្រិតទី១០ គឺពាក្យ «អាយុ» ដែលមានចំនួន ៦២ ៣៣២ពាក្យ ស្មើនឹង ០,២៦%។ ចំពោះកិរិយាសព្ទ ពាក្យ «បាន» ត្រូវបានប្រើប្រាស់ញឹកញាប់ជាងគេ គឺមានចំនួន ៣៥៣ ៥៧៦ពាក្យ ស្មើនឹង ១,៤៩% ខណៈពាក្យ «តាម» ត្រូវបានប្រើប្រាស់ច្រើនជាងគេក្នុងកម្រិតទី១០ ដែលមានចំនួន ៧៦ ១៧៥ ស្មើនឹង ០,៣២%។ គុណនាមដែលត្រូវបានប្រើប្រាស់ជាអតិបរមា គឺពាក្យ «នាក់» ដែលមានចំនួន ១២៥ ៤១៩ពាក្យ ស្មើនឹង

0,៥៣% ខណៈគុណនាមដែលគេប្រើប្រាស់ច្រើនជាងគេក្នុងកម្រិតទី១០ គឺពាក្យ «ល្អ» មានចំនួន ១៤ ៥៥៩ពាក្យ ស្មើនឹង 0,0៦%។ ដូច្នេះ ក្រុមពាក្យកិរិយាសព្ទដែលមានចំនួនពាក្យ និងភាគរយច្រើនជាងគេ ត្រូវបានគេប្រើប្រាស់នៅក្នុងអត្ថបទទាំង ២០៩អត្ថបទ។ សរុបមក យោងតាមប្រភពទិន្នន័យនៃអត្ថបទអំណាន ពាក្យដែលត្រូវបានប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ មានចំនួន ១ ៩៧៤ពាក្យ ដែលមានបង្ហាញក្នុងរូបភាពទី២។ ពាក្យទាំងនេះទំនងជាពាក្យស្រួលជាងគេ ដែលអ្នកនិពន្ធនៅក្នុងស្ថាប័នបោះពុម្ពទាំងនោះរំលេចចេញ ដើម្បីឱ្យអ្នកអានគ្រប់ស្រទាប់យល់ខ្លឹមសារនៃអត្ថបទអំណាននោះបាន។

លទ្ធផលនៃការសិក្សានេះដូចនឹងការស្រាវជ្រាវរបស់ Nation (2012) និង Van Zeeland & Schmitt (2013) ដែលបានកំណត់ថាពាក្យប្រើប្រាស់ញឹកញាប់មានចំនួន ២ ០០០ពាក្យ និងការស្រាវជ្រាវរបស់ Biber & Barbieri (2007) ដែលបានកំណត់និយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ។ ប៉ុន្តែ លទ្ធផលដែលបានរកឃើញក្នុងសិក្សានេះ គឺខុសពីការសិក្សារបស់ Laufer & Ravenhorst-Kalovski (2010) ដែលបានអះអាងថាពាក្យប្រើប្រាស់ញឹកញាប់មានចំនួនពី ៣ ០០០ ទៅ ៥ ០០០ពាក្យ និងកំណត់និយាមស្តង់ដា ២០ដង (Lynn, 1973) ២៥ដង (Bestgen, 2020) ៤៥ដង (Francis et al., 1982) និង ៥០ដង (Gardner & Davies, 2013) ទើបអាចកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់បាន។ លទ្ធផលស្រាវជ្រាវនេះបានបង្ហាញថាពាក្យចំនួន ១ ៩៧៤ពាក្យបានបំពេញតាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ ដែលពាក្យទាំងនេះមានសារៈសំខាន់សម្រាប់ការយល់ដឹងអំពីខ្លឹមសារអត្ថបទនៅគ្រប់កម្រិតនៃអត្ថបទអំណានផ្សេងៗក្នុងភាសាខ្មែរ។

ថ្នាក់ពាក្យដែលត្រូវបានប្រើប្រាស់ច្រើនជាងគេ

ដើម្បីឆ្លើយតបនឹងសំណួរស្រាវជ្រាវទី២ អ្នកស្រាវជ្រាវបានយកពាក្យដែលត្រូវបានប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យមកផ្ទៀងផ្ទាត់ជាមួយវចនានុក្រមខ្មែរ ឆ្នាំ២០២២។ ពាក្យដែលប្រើប្រាស់ញឹកញាប់ ចំនួន ១ ៩៧៤ពាក្យ ត្រូវបានបែងចែកជាថ្នាក់ពាក្យតាមលំនាំភាសាខ្មែរ ដែលមានដូចជា នាមសព្ទ កិរិយាសព្ទ គុណនាម គុណកិរិយាសព្ទ សព្វនាម ឈ្នាប់ និងឆ្នាក់។ លទ្ធផលថ្នាក់ពាក្យដែលត្រូវបានប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា មានបង្ហាញក្នុងរូបភាពទី២។



រូបភាពទី២៖ ថ្នាក់ពាក្យដែលត្រូវបានប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា

រូបភាពទី២ បង្ហាញអំពីថ្នាក់ពាក្យនៃពាក្យដែលត្រូវបានប្រើប្រាស់ញឹកញាប់តាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ។ ថ្នាក់ពាក្យទាំងនេះ រួមមាន នាមសព្ទ កិរិយាសព្ទ គុណនាម គុណកិរិយាសព្ទ សព្វនាម ឈ្នាប់ និង ធ្នាក់។ គួរកត់សម្គាល់ថា ក្នុងចំណោមពាក្យចំនួន ១ ៩៧៤ពាក្យ ដែលត្រូវបានប្រើប្រាស់ញឹកញាប់បំផុត មានតែថ្នាក់ពាក្យរបស់ឧទានសព្ទតែប៉ុណ្ណោះ ដែលពុំមាននៅក្នុងពាក្យប្រើប្រាស់ញឹកញាប់បំផុត។ ដោយឡែក នាមសព្ទគឺជាថ្នាក់ពាក្យដែលត្រូវបានប្រើប្រាស់ច្រើនជាងគេនៅក្នុងចំណោមថ្នាក់ពាក្យទាំងអស់ដែលមានចំនួន ១ ០០៨ពាក្យ ស្មើនឹង ៥០,៨១% នៃចំនួនពាក្យជិត ២ ០០០ពាក្យ។ លទ្ធផលនេះមានន័យថា បើក្នុងពាក្យប្រើប្រាស់ញឹកញាប់ត្រូវបានគេសរសេរ ១ល្អះមាន ១០ពាក្យ គឺមាននាមសព្ទចំនួន ៥ពាក្យនៅក្នុងល្អះនោះរួចទៅហើយ។ លើសពីនេះ កាលណានាមសព្ទត្រូវបានប្រើប្រាស់ច្រើននៅក្នុងភាសានេះជាកស្តុតាងមួយដែលបានបញ្ជាក់ឱ្យឃើញថា នាមសព្ទអាចឱ្យយើងកំណត់អត្តសញ្ញាណដូចជាឈ្មោះ ការពិពណ៌នា និងការពិភាក្សាអំពីអ្វីៗដែលនៅជុំវិញខ្លួនយើងស្តីអំពីនាមរូបី និងនាមអរូបីបានយ៉ាងប្រាកដប្រជា។ លើសពីនេះទៀត នាមសព្ទក៏ដើរតួនាទីយ៉ាងសំខាន់ចំពោះការបង្កើតល្អះ ឬប្រយោគក្នុងការប្រស្រ័យទាក់ទងគ្នាបានយ៉ាងមានប្រសិទ្ធភាពខ្ពស់ (Pescuma et al., 2021; Schmid, 2000)។ លើសពីនេះទៀត នាមសព្ទក៏ដើរតួនាទីយ៉ាងសំខាន់ចំពោះការបង្កើតល្អះ ឬប្រយោគក្នុងការប្រស្រ័យទាក់ទងគ្នាបានយ៉ាងមានប្រសិទ្ធភាពខ្ពស់ (Garcia-Gamez & Macizo, 2019; McGhee-Bidlack, 1991)។ លទ្ធផលសិក្សានេះ ដូចគ្នាទៅនឹងលទ្ធផលដែលបានរកឃើញក្នុងការសិក្សារបស់ Shih et al. (2000) និង Simsek & Gun (2021) ដែលបានសិក្សាអំពីថ្នាក់ពាក្យចេញពី Corpus Linguistics និងបានរកឃើញថានាមសព្ទត្រូវបានគេប្រើប្រាស់ច្រើនជាងគេ គ្រាន់តែភាគរយនៃការប្រើប្រាស់នាមសព្ទខុសគ្នាតែប៉ុណ្ណោះ ដូចជា ២៤% សម្រាប់ Shih et al. (2000) និង ៦៥% សម្រាប់ Simsek & Gun (2021)។

នាមសព្ទចំនួន ១៥ពាក្យដែលត្រូវបានគេប្រើប្រាស់ញឹកញាប់ជាងគេ

រូបភាពទី៣ បង្ហាញពីនាមសព្ទដែលត្រូវបានគេប្រើប្រាស់ច្រើនជាងគេនៅក្នុងចំណោមពាក្យសរុបជិត ២៤លានពាក្យ ឬចំណោមពាក្យដែលត្រូវបានប្រើប្រាស់ញឹកញាប់ជិត ២ ០០០ពាក្យ។ ចំពោះពាក្យដែលសរសេរជិតក្រាស់ជាងគេ ក្នុងពាក្យពពក (Word Cloud) គឺមានន័យថាពាក្យនោះត្រូវបានគេសរសេរ ច្រើនជាងគេ តាមនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ។

រូបភាពទី៣៖ នាមសព្ទចំនួន ១៥ពាក្យត្រូវបានគេប្រើប្រាស់ញឹកញាប់ជាងគេ

រូបភាពទី៣ បង្ហាញអំពីនាមសព្ទ ជាលក្ខណៈពាក្យពពក ដែលត្រូវបានប្រើប្រាស់ច្រើនជាងគេ ក្នុងចំណោមពាក្យជិត ២៤ លានពាក្យ ដែលទាញចេញពីប្រភពទិន្នន័យចំនួន ៥ប្រភេទ ដូចបានរៀបរាប់ខាងលើ។ ក្នុងការបកស្រាយខ្លឹមសារនៃរូបភាពពាក្យពពកនេះ អ្នកស្រាវជ្រាវសូមលើកយកតែពាក្យចំនួន ៣៣ក្យ ដែលប្រើប្រាស់ញឹកញាប់មកអធិប្បាយតែប៉ុណ្ណោះ។ ពាក្យ «លោក» ជានាមសព្ទលំដាប់ទី១ ដែលត្រូវបានប្រើប្រាស់ច្រើនជាងគេ និងពាក្យញឹកញាប់ (Freq) មានចំនួន ១៩២ ៧០៤ពាក្យ និងនិយាមស្តង់ដា (NormFreq) មានចំនួន ៨ ០៩៤ដងក្នុង ១លានពាក្យ។ ពាក្យនេះត្រូវបានគេសរសេរ ដើម្បីកំណត់អត្តសញ្ញាណភេទ និងមុខតំណែងឱ្យបានច្បាស់លាស់ និងត្រូវបានគេនិយមសរសេរភ្ជាប់ពីមុខពាក្យដូចជា «ប្រធានាធិបតី នាយករដ្ឋមន្ត្រី រដ្ឋមន្ត្រី អភិបាល ឧត្តមសេនីយ៍ វរសេនីយ៍ អ្នកនាំពាក្យ ឧកញ៉ា ជំទាវ បណ្ឌិត នាយក ប្រធានសង្ឃ គ្រូ តា យាយ ពូ ក្មួយ និងឈ្មោះបុគ្គលណាម្នាក់ (ឧ. លោក សុខ សម្បត្តិ)»។ នាមសព្ទដែលឈរលំដាប់ទី២ គឺជាពាក្យ «ឆ្នាំ» ដែលមានពាក្យប្រើប្រាស់ញឹកញាប់ចំនួន ១៤២ ៧២៥ពាក្យ និងនិយាមស្តង់ដាចំនួន ៥ ៩៩៥ដង។ អ្នកសរសេរបានប្រើពាក្យនេះ ដើម្បីកំណត់អំពីអាយុ (៤៥ឆ្នាំ ៧៨ឆ្នាំ) ស៊េរីឧបករណ៍ប្រើប្រាស់ (ឡាន ម៉ូតូ ទូរសព្ទ កុំព្យូទ័រ) កាលបរិច្ឆេទ (ថ្ងៃទី១០ ខែសីហា ឆ្នាំ២០១១) រយៈពេល (១០ឆ្នាំ យូរឆ្នាំ ប្រចាំឆ្នាំ) និងឆ្នាំតាមចន្ទគតិ (ជូត ឆ្លូវ ខាល ថោះ)។ បន្ទាប់មក ពាក្យ «ឈ្មោះ» ដែលត្រូវបានប្រើប្រាស់ច្រើនជាងគេលំដាប់ទី៣ មានពាក្យប្រើប្រាស់ញឹកញាប់ចំនួន ៨០ ៨៨៣ ពាក្យ ក្នុងចំណោមពាក្យជិត ២៤លានពាក្យ និងនិយាមស្តង់ដាចំនួន ៣៣៩៧ដងក្នុង ១លានពាក្យ។ ពាក្យនេះត្រូវបានគេប្រើប្រាស់ ដើម្បីកំណត់អត្តសញ្ញាណរបស់មនុស្ស សត្វ ទីកន្លែង និងវត្ថុ ឱ្យកាន់តែច្បាស់ និងបញ្ជាក់ន័យឱ្យអ្នកអានងាយស្រួលក្នុងការឆ្លុះបញ្ចាំង ឬផ្ទៀងផ្ទាត់ជាមួយគំនិតរបស់ពួកគេ។

សេចក្តីសន្និដ្ឋាន

លទ្ធផលនៃការស្រាវជ្រាវនេះបានបង្ហាញពីពាក្យប្រើប្រាស់ញឹកញាប់នៅក្នុងសំណុំទិន្នន័យនៃពាក្យចំនួនជិត ២៤ លានពាក្យ ដែលបានទាញចេញពីអត្ថបទចំនួន ២០៩អត្ថបទ។ ពាក្យដែលប្រើប្រាស់ញឹកញាប់ជាភាសាខ្មែរដែលបានទាញចេញពីកម្មវិធី AntConc Corpus Linguistics មានចំនួនជិត ២ ០០០ពាក្យដែលត្រូវបានប្រើប្រាស់ញឹកញាប់ និងពាក្យនាមសព្ទដែលត្រូវបានប្រើប្រាស់ច្រើនជាងគេក្នុងចំណោមក្រុមពាក្យជាទៃទៀតមានចំនួន ១ ០០៨ពាក្យ។ លទ្ធផលនេះផ្តល់នូវការយល់ដឹងដ៏មានតម្លៃចំពោះរចនាសម្ព័ន្ធនៃការប្រើប្រាស់ពាក្យក្នុងភាសាខ្មែរ។ ពាក្យប្រើប្រាស់ញឹកញាប់មានចំនួនជិត ២ ០០០ពាក្យ ដែលត្រូវបានរកឃើញនៅក្នុងការស្រាវជ្រាវនេះ តាមរយៈនិយាមស្តង់ដា ៤០ដងក្នុង ១លានពាក្យ។ ពាក្យទាំងនេះអាចក្លាយជាបញ្ជីពាក្យកម្រិតមូលដ្ឋានគ្រឹះសម្រាប់ធ្វើការទំនាក់ទំនងប្រកបដោយប្រសិទ្ធភាព និងអាចជួយដល់ការយល់ដឹងអំពីអត្ថបទអំណានកាន់តែប្រសើរ។ មិនថាអ្នកចាប់ផ្តើមដំបូង ឬអ្នកអានកម្រិតខ្ពស់នោះទេ ការចេះពាក្យទាំងនេះបង្កើតបានជាមូលដ្ឋានទូទៅសម្រាប់ការយល់ដឹង និងការចូលរួមចំណែកជាមួយភាសាសរសេរ និងអាចជាភាសានិយាយថែមទៀត។

នៅក្នុងប្រភេទនៃពាក្យប្រើញឹកញាប់នេះ នាមសព្ទបានលេចចេញជាថ្នាក់ពាក្យដែលលេចធ្លោជាងគេ និងមានចំនួន ១ ០០៨ពាក្យ ដែលប្រហែលស្មើនឹង ៥១% នៃចំនួនពាក្យសរុបប្រើប្រាស់ញឹកញាប់។ ភាពលេចធ្លោនេះគូសបញ្ជាក់ពីតួនាទីដ៏សំខាន់របស់នាមសព្ទក្នុងភាសាខ្មែរ។ នាមសព្ទគឺជាអ្នកនាំអត្ថន័យ កំណត់អត្តសញ្ញាណ និងពណ៌នាអំពីពិភពលោក ពីរូបធាតុជាក់ស្តែង រហូតដល់គំនិតអរូបី។ នាមសព្ទក៏ផ្តល់មូលដ្ឋានគ្រឹះ សម្រាប់ល្បះ និងការប្រាស្រ័យទាក់ទងដោយបញ្ជូនគំនិត និងព័ត៌មានប្រកបដោយប្រសិទ្ធភាពខ្ពស់។

លទ្ធផលការស្រាវជ្រាវនេះបានបង្ហាញថាពាក្យដែលត្រូវបានប្រើប្រាស់ញឹកញាប់ច្រើនជាងគេក្នុងចំណោម ២ ០០០ ពាក្យ គឺជានាមសព្ទដែលមានមូលដ្ឋានចម្បងសម្រាប់តភ្ជាប់រវាងការប្រើប្រាស់ភាសា និងការយល់ដឹង។ នាមសព្ទដើរតួជាផ្លូវកាត់ផ្នែកភាសា ដើម្បីធ្វើឱ្យមនុស្សគ្រប់គ្នាទទួលបានបទពិសោធក្នុងការអានអត្ថបទឱ្យមានភាពងាយស្រួល ដោយមិនពឹងផ្អែកលើជំនាញផ្សេងៗដែលពួកគាត់បានសិក្សា។ លទ្ធផលនេះសង្កត់ធ្ងន់លើតម្លៃនៃការបង្រៀន និងរៀនពាក្យស្នូលនៃពាក្យប្រើប្រាស់ញឹកញាប់ជាជំហាន និងជាមូលដ្ឋានគ្រឹះក្នុងការរៀនភាសា។ លើសពីនេះទៅទៀត អ្នកអប់រំ អ្នកបង្កើតកម្មវិធីសិក្សា និងអ្នកស្រាវជ្រាវគួរតែពិចារណាអំពីតួនាទីពាក្យស្នូលនៃនាមសព្ទ និងមុខងាររបស់វា ដោយទទួលស្គាល់ថានាមសព្ទមិនមែនគ្រាន់តែជាផ្នែកនៃភាសាប៉ុណ្ណោះទេ។ នាមសព្ទគឺជាមូលដ្ឋានគ្រឹះក្នុងការបង្កើតទំនាក់ទំនងប្រកបដោយប្រសិទ្ធភាព។

សរុបមក ពាក្យប្រើញឹកញាប់ រួមទាំងពាក្យដែលបានកំណត់អត្តសញ្ញាណចំនួនជិត ២ ០០០ពាក្យទាំងនេះ បម្រើជាមូលដ្ឋានគ្រឹះនៃភាសាខ្មែរសម្រាប់ការយល់ដឹងទៅលើអត្ថបទ និងជួយអ្នកអាន ឬអ្នកសិក្សាឱ្យមានគំនិតយល់ដឹងបានភាគច្រើន។ ជាពិសេស នាមសព្ទលេចចេញជាឆ្អឹងខ្នងនៃភាសាខ្មែរដែលមានជាងពាក់កណ្តាលនៃពាក្យប្រើប្រាស់ញឹកញាប់។ ចំណេះដឹងអំពីនាមសព្ទដែលប្រើប្រាស់ញឹកញាប់ជាងគេក្នុងភាសាខ្មែរ គឺជាធាតុគាំទ្រដ៏មានសារៈសំខាន់ក្នុងការប្រាស្រ័យទាក់ទងប្រកបដោយប្រសិទ្ធភាព និងបង្ហាញពីតួនាទីរបស់នាមសព្ទក្នុងការសម្រួលអត្ថបទសម្រាប់អ្នកអានដែលមានជំនាញផ្សេងៗគ្នា។

ដែនកំណត់ និងអនុសាសន៍នៃការស្រាវជ្រាវ

ការស្រាវជ្រាវនេះមានដែនកំណត់ ដោយសារអត្ថបទភាសាខ្មែរគ្មានដំណកឃ្លា (Zero space) ពីពាក្យមួយ ទៅពាក្យមួយទៀត ដូចពាក្យក្នុងអង់គ្លេស ឬបារាំង។ ភាសាខ្មែរក៏មានពាក្យសរសេរជាមួយជើង «ដ» និងជើង «ត» ដែលមានរូបរាងដូចគ្នា ដែលបង្កភាពស្មុគស្មាញក្នុងការកាត់ពាក្យឱ្យបានត្រឹមត្រូវ ដើម្បីយកពាក្យទាំងនោះមកបញ្ចូលក្នុងកម្មវិធី Corpus Linguistics (AntConc)។ លើសពីនេះទៅទៀត ប្រភពអត្ថបទមិនទាន់សម្បូរបែប និងពាក្យនៅមានចំនួនតិចក្នុងការព្យាករកពាក្យដែលត្រូវបានប្រើប្រាស់ញឹកញាប់ជាងគេ។ ទោះបីជាមានដែនកំណត់ ឬចំណុចខ្វះខាតទាំងនេះក៏ដោយ បើយោងតាមទិន្នន័យដែលបានវិភាគ ដើម្បីកំណត់ពាក្យប្រើប្រាស់ញឹកញាប់ និងថ្នាក់ពាក្យប្រើប្រាស់ច្រើនជាងគេក្នុងភាសាខ្មែរ ការស្រាវជ្រាវនេះអាចផ្តល់អនុសាសន៍មួយចំនួន ដូចខាងក្រោម៖

- អ្នកអប់រំគួរតែយកលទ្ធផលនៃពាក្យប្រើប្រាស់ញឹកញាប់ និងនាមសព្ទជាក់ជាអាទិភាពក្នុងការបង្រៀននិងរៀននៅកម្រិតមូលដ្ឋានដំបូង ដើម្បីជួយឱ្យអ្នកសិក្សាទទួលបានកាន់តែប្រសើរទាំងភាសាសរសេរ និងភាសានិយាយ។
- អ្នករៀបចំកម្មវិធីសិក្សា ឬអ្នកអប់រំគួរតែយកលទ្ធផលដែលបានរកឃើញទាំងនេះមកពិចារណានៅពេលរៀបចំកម្មវិធីសិក្សាភាសាខ្មែរ ដើម្បីធានាថាអ្នកសិក្សាមានការយល់ច្បាស់អំពីនាមសព្ទដោយផ្តោតលើសកម្មភាព និងធ្វើលំហាត់ក្នុងការពង្រឹងការយល់ដឹងរបស់អ្នកសិក្សា។
- អ្នករៀបចំសម្ភារៈសិក្សា និងសកម្មភាពសិក្សាជាភាសាខ្មែរគួរតែរៀបចំតាមកម្រិតជំនាញផ្សេងៗគ្នា ដោយធានាថាអ្នកសិក្សាដែលចាប់ផ្តើមដំបូងត្រូវមានភាពស្ទាត់ជំនាញអំពីនាមសព្ទដែលស្ថិតនៅក្នុងថ្នាក់ពាក្យដែលប្រើប្រាស់ញឹកញាប់បំផុត។
- អ្នកស្រាវជ្រាវក្រោយៗអាចយកលទ្ធផលក្នុងការស្រាវជ្រាវនេះ ទៅជំរុញឱ្យមានការស្រាវជ្រាវបន្ថែមទៅលើគំរូនៃការប្រើប្រាស់ពាក្យនៅក្នុងសំណុំទិន្នន័យឱ្យបានច្រើនជាងនេះ និងរួមបញ្ចូលភាសាផ្សេងៗគ្នា ដើម្បីបន្តស្វែងយល់ពី

ភាពខុសប្លែកគ្នានៃពាក្យប្រើប្រាស់ញឹកញាប់ និងការប្រើប្រាស់ពាក្យនៅក្នុងបរិបទភាសាចម្រុះ។ អ្នកអប់រំភាសា និងអ្នកស្រាវជ្រាវដទៃទៀតគួរទទួលស្គាល់ថា ខណៈពេលដែលនាមសព្ទជាប់ចំណាត់ថ្នាក់ខ្ពស់ជាថ្នាក់ពាក្យដែលមានការប្រើប្រាស់ញឹកញាប់ ការណ៍នេះអាចនឹងប្រែប្រួលដោយផ្អែកលើសំណុំទិន្នន័យ និងភាសាផ្សេងៗក្នុងបរិបទភាសាជុំវិញពិភពលោក។

សរុបមក អ្នកសិក្សាភាសា អ្នកអភិវឌ្ឍន៍កម្មវិធីសិក្សា និងអ្នកស្រាវជ្រាវខាងផ្នែកភាសា គួរគិតគូរ និងចែកចាយពីពាក្យ និងថ្នាក់ពាក្យ ដែលត្រូវបានគេប្រើប្រាស់ញឹកញាប់ក្នុងភាសាខ្មែរឱ្យបានទូលំទូលាយ ដើម្បីជាប្រយោជន៍ដល់អ្នកសិក្សាភាសាខ្មែរឱ្យចំគោលដៅ និងចំណេញពេលវេលា។

សេចក្តីថ្លែងអំណរគុណ

អ្នកនិពន្ធសូមថ្លែងអំណរគុណដល់ក្រុមហ៊ុន សង្គប្ប អនុវត្តន៍។ អ្នកនិពន្ធក៏សូមថ្លែងអំណរគុណដល់និពន្ធនាយក និងអ្នកត្រួតពិនិត្យជំនាញអនាមិករបស់ទស្សនាវដ្តីស្រាវជ្រាវកម្ពុជាសម្រាប់ការអប់រំ និងស្នែម សម្រាប់មតិយោបល់កែលម្អលើអត្ថបទស្រាវជ្រាវនេះ។ ខ្លឹមសារក្នុងអត្ថបទនេះ គឺជាការទទួលខុសត្រូវរបស់អ្នកនិពន្ធ និងមិនឆ្លុះបញ្ចាំងពីទស្សនៈ ឬនិន្នាការនយោបាយរបស់ក្រុមណាមួយឡើយ។

ឯកសារយោង (References)

Anthony, L. (2022). *Laurence Anthony's AntConc* (Version 4.2.0). https://www.laurenceanthony.net/software/antconc/

Bestgen, Y. (2020). Comparing lexical bundles across corpora of different sizes: The Zipfian problem. *Journal of Quantitative Linguistics*, *27*(3), 272–290. https://doi.org/10.1080/09296174.2019.1566975

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Chen, Y. H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*(6), 849–880. https://doi.org/10.1093/applin/amu065

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning*, *67*(5), 959–997. https://doi.org/10.1111/lang.12253

Francis, W. N., & Kucera, H. (1979). Brown corpus manual. *Letters to the Editor*, *5*(2), 9–12.

Francis, W. N., Kučera, H., & Mackie, A. W. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin.

Gardner, D., & Davies, M. (2013). New academic vocabulary list. *Applied Linguistics*, *35*(14), 305–327. https://doi.org/10.1093/applin/amt015

Kennedy, G. (2001). Corpus linguistics. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 2816–2820). Pergamon. https://doi.org/10.1016/B0-08-043076-7/03056-4

Khoy, B. (2021). Measuring Khmer vocabulary size for each grade in basic education. *International Journal of Innovative Science and Research Technology*, *6*(1), 194–202.

Khoy, B., Suon, S., & Khan, B. (2021). A syntactic analysis of Cambodian news discourse on COVID-19 outbreaks: Sentence lengths and structures as predictors of readability. *Ahwaz Journal of Linguistics Studies*, *2*(3), 22–36.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

Let's Read. (2016, October 17). *Let's Read: Children's books: Free to read download translate*. https://www.letsreadasia.org/

Lynn, R. W. (1973). Preparing word-lists: A suggested method. *RELC Journal*, *4*(1), 25–28. https://doi.org/10.1177/003368827300400103

Nation, I. S. P. (2012). *The BNC/COCA word family lists.* http://www.victoria.ac.nz/lals/about/staff/paul-nation

Reppen, R. (2009). English language teaching and corpus linguistics: Lessons from the American National Corpus. In P. Baker (Ed.), *Contemporary approaches to corpus linguistics* (pp. 206–215). Continuum Press.

Royal Academy of Cambodia. (2023). *វចនានុក្រមខ្មែរ ២០២២* [Khmer dictionary 2022].

Shih, R. H., Chiang, J. Y., & Tien, F. (2000, Augst). Part-of-speech sequences and distribution in a learner corpus of English. *In Proceedings of Research on Computational Linguistics Conference XIII* (pp. 171–177).

Simsek, R., & Gun, M. (2021). A corpus-based research: The vocabulary content and parts of speech of A1 level textbooks used in teaching Turkish as a foreign language. *Journal of Language and Linguistic Studies*, *17*(15), 236–237. https://doi.org/10.3316/informit.166852590053499

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins Publishing Company.

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479.

https://doi.org/10.1093/applin/ams074

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis*. John Wiley & Sons.

Wiegand, V., & Mahlberg, M. (Eds.). (2019). *Corpus linguistics, context and culture*. De Gruyter. https://doi.org/10.1515/9783110489071