ಟೆಕ್ಸಿಪ್ - II / ವಿಶ್ವ ಬ್ಯಾಂಕ್ ಪ್ರಾಯೋಜಿತ ಉತ್ಕೃಷ್ಟತೆ ಕೇಂದ್ರ

ಕೆನೋ (KAnOE) – ಜ್ಞಾನ ವಿಶ್ಲೇಷಣೆ ಮತ್ತು ಮೂಲತತ್ತ್ವ ತಂತ್ರಶಾಸ್ತ್ರ ಸಂಶೋಧನಾ ಕೇಂದ್ರ

ಪಿಇಎಸ್ ತಾಂತ್ರಿಕ ಮಹಾವಿದ್ಯಾಲಯ / ಪಿಇಎಸ್ ವಿಶ್ವವಿದ್ಯಾಲಯ

"ಪಿಇಎಸ್ ಉಲ್ಲೇಖ ಮಾಲಾ"

ದಶ ಲಕ್ಷ ಕನ್ನಡ ವಿಕಿ ಪುಟಗಳು

http://kanoe.org/kannada.html

ಡಾ॥ ಕವಿ ಮಹೇಶ್

ಹಾಗೂ ಸ್ನಾತಕೋತ್ತರ ವಿದ್ಯಾರ್ಥಿಗಳಾದ ಶ್ರೀ ಅನಂತ ಕೃಷ್ಣ ತಂತ್ರಿ ಅವರಿಂದ,

ಶ್ರೀ ಉದಯ ಕಿರಣ ಪಿ. ಅವರ ಸಹಾಯದಿಂದ ರಚಿಸಲ್ಪಟ್ಟದ್ದು

ನವೆಂಬರ್ 2014

ಅಂತರ್ಜಾಲವು ಒಂದು ಉಪಯುಕ್ತವಾದ ಮಾಹಿತಿಯ ಕೋಶವಾಗಿ ಬೆಳೆಯುತ್ತಿದೆ. ಆದರೆ, ಅದರಲ್ಲಿ ಹೆಚ್ಚಾಗಿ ಆಂಗ್ಲ ಭಾಷೆಯಲ್ಲೇ ಎಲ್ಲಾ ವಿದ್ಯಮಾನಗಳೂ, ಲೇಖನಗಳೂ, ಮಾಹಿತಿ ಅಂಶಗಳೂ ಲಭ್ಯವಿವೆ. ಹೀಗಾಗಿ ಕನ್ನಡ ಮತ್ತಿತರ ಭಾರತೀಯ ಭಾಷೆಯವರಿಗೆ ಅಂತರ್ಜಾಲದಲ್ಲಿರುವ ವಿಶಾಲವಾದ ಮಾಹಿತಿ ಜಗತ್ತಿನ ಪ್ರಯೋಜನವು ಸ್ವಲ್ಪವೂ ಸಿಗುತ್ತಿಲ್ಲ. ಉದಾಹರಣೆಗೆ, ಆಂಗ್ಲ ಭಾಷೆಯ ವಿಕಿಪೀಡಿಯಾದಲ್ಲಿ (http://en.wikipedia.org) ಸುಮಾರು 46,43,000 ಪುಟಗಳಿವೆ; ಕನ್ನಡದ ವಿಕಿಪೀಡಿಯಾದಲ್ಲಿ (http://kn.wikipedia.org) ಹನ್ನೆರಡು ವರ್ಷಗಳ ನಂತರ ಕೇವಲ 17,000 (ಅಂದರೆ ಆಂಗ್ಲದ ಕೇವಲ 0.37%) ಪುಟಗಳಿವೆ. ಇತರ ಕನ್ನಡದ ಕೋಶಗಳೂ ಇದೇ ಸ್ಥಿತಿಯಲ್ಲೇ ಇವೆ. ಉದಾಹರಣೆಗೆ, ಕಣಜ ಎಂಬ ಕೋಶದಲ್ಲಿ (http://kanaja.in) ಸುಮಾರು 13,000 ಪುಟಗಳಿವೆ.

ಏತನ್ಮಧ್ಯೆ, ಹಲವಾರು ದಶಕಗಳ ಸಂಶೋಧನೆಯ ನಂತರವೂ ಕಂಪ್ಯೂಟರ್ ಭಾಷಾಂತರ (Machine Translation) ದ ಮೂಲಕ ಕನ್ನಡಕ್ಕೆ ಆಂಗ್ಲದಿಂದ ಯಾಂತ್ರಿಕ ತರ್ಜುಮೆ ಮಾಡುವಂತಹ ಉಪಯುಕ್ತವಾದ ತಂತ್ರಾಂಶವೇನೂ ನಮಗೆ ದೊರೆತಿಲ್ಲ. ಗೂಗಲ್ ನವರ ತರ್ಜುಮೆ ಯಂತ್ರದ ಗುಣಮಟ್ಟವು ಆಂಗ್ಲದ ವಿಕಿಪೀಡಿಯ ಲೇಖನಗಳನ್ನು ಕನ್ನಡಕ್ಕೆ ಯಾಂತ್ರಿಕವಾಗಿ ಭಾಷಾಂತರಿಸುವ ಮಟ್ಟಕ್ಕೆ ಈವರೆಗಂತೂ ಬಂದಿಲ್ಲ.

ಈ ಸಮಸ್ಯೆಯನ್ನು ಬಗೆಹರಿಸಲು *ಉಲ್ಲೇಖ ಮಾಲಾ* ದಲ್ಲಿ ಒಂದು ಅಚ್ಚ ಹೊಸದಾದ ಮಾರ್ಗದಿಂದ ಮಾಹಿತಿ ಪುಟಗಳನ್ನು ಕನ್ನಡದಲ್ಲಿ ಸೃಷ್ಟಿಸಲಾಗಿದೆ. ಇಲ್ಲಿನ ಹೊಸ ಯೋಚನೆ ಏನೆಂದರೆ: ಮಾಹಿತಿಯಿಂದ ನೇರವಾಗಿ ಕನ್ನಡ ಪುಟಗಳ ತಯಾರಿಕೆ; ಆಂಗ್ಲದ ಪುಟಗಳಿಂದಲ್ಲ! ಅತ್ಯಂತ ಶ್ರೇಷ್ಠಮಟ್ಟದ ಯಾಗೋ (*YAGO*) ಮತ್ತು ಡಿಬಿಪೀಡಿಯ (*Dbpedia*) ಮೊದಲಾದ ಮಾಹಿತಿ ಕೋಶಗಳು ಸುಲಭವಾಗಿ ಸಿಗುತ್ತವೆ. ಆದರೆ ಇದರಲ್ಲಿರುವ ಮಾಹಿತಿಯ

ಅಂಶಗಳೂ ಕೂಡ ಆಂಗ್ಲ ಭಾಷೆಯಲ್ಲೇ ಇವೆ. ಯಾಗೋ (*Yet Another Great Ontology*, "ಮತ್ತೊಂದು ಶ್ರೇಷ್ಠ ಮೂಲತತ್ತ್ವ") ಎನ್ನುವುದು ಜರ್ಮನಿ ದೇಶದ ಮ್ಯಾಕ್ಸ್ ಪ್ಲಾಂಕ್ ಇನ್ಸ್ಟಿಟ್ಯೂಟ್ ಫಾರ್ ಇನ್ಫಾರ್ಮ್ಯಾಟಿಕ್ಸ್ (*Max Plank Institute for Informatics*) ನಲ್ಲಿ ವಿಕಿಪೀಡಿಯಾದಿಂದಲೇ ಶೋಧಿಸಿ ತಯಾರಿಸಲ್ಪಟ್ಟ 19 ಕೋಟಿ (19,00,00,000) ಅಂಶಗಳ ಒಂದು ಬೃಹತ್ ಕೋಶ. ಇದರಲ್ಲಿ ವ್ಯಕ್ತಿಗಳು, ಸ್ಥಳಗಳು, ರಾಷ್ಟ್ರಗಳು, ಮುಂತಾದ ವಿವಿಧ ವಿಷಯಗಳ ಬಗ್ಗೆ ಶುದ್ಧವಾದ ಮಾಹಿತಿ ದೊರಕುತ್ತದೆ. ಹೀಗಾಗಿ ಉಲ್ಲೇಖ ಮಾಲಾ ವನ್ನು ಹೆಚ್ಚಾಗಿ ಯಾಗೋದಲ್ಲಿರುವ ಮಾಹಿತಿಯಿಂದಲೇ ತಯಾರಿಸಲಾಗಿದೆ.

ಮೊದಲನೆಯದಾಗಿ, ಯಾಗೋ ದಲ್ಲಿರುವ ಮಾಹಿತಿ ಅಂಶಗಳನ್ನು ಕನ್ನಡಕ್ಕೆ ಮಾಡಿ, ಜಗತ್ತಿನಲ್ಲೇ ಪ್ರಪ್ರಥಮ ಕನ್ನಡದ ಲಿಂಕ್ಡ್ ಓಪನ್ ಡಾಟಾಸೆಟ್ (*Linked Open Dataset*) ಅನ್ನು ತಯಾರಿಸಲಾಯಿತು. ಕನ್ನಡದಲ್ಲೇ ಏಕೆ, ನಮಗೆ ತಿಳಿದ ಮಟ್ಟಿಗೆ, ಯಾವುದೇ ಭಾರತೀಯ ಭಾಷೆಯಲ್ಲೂ ಹಿಂದೆ ಯಾರೂ ಇಂತಹ ಮಾಹಿತಿ ಕೋಶವನ್ನು ತಯಾರು ಮಾಡಿಲ್ಲ. ಯಾಗೋದಲ್ಲಿರುವ ಎಲ್ಲ 2.5 ಕೋಟಿ ಹೆಸರುಗಳನ್ನೂ ಯಾಂತ್ರಿಕವಾಗಿ ಲಿಪ್ಯಾಂತರ ಅಥವಾ ಶಬ್ದಾಂತರ ಮಾಡುವುದರ ಮೂಲಕ, ಹಾಗೂ ಕೇವಲ ಸುಮಾರು ಒಂದು ನೂರು ಅರ್ಥಗರ್ಭಿತ ಶಬ್ದಗಳನ್ನು ಕೈಯ್ಯಾರೆ ಭಾಷಾಂತರಿಸುವ ಮೂಲಕ, ಈ ಕನ್ನಡ ಕೋಶವನ್ನು ತಯಾರು ಮಾಡಲಾಯಿತು. ಉದಾಹರಣೆಗೆ *wasBornOnDate* ಅನ್ನು " .. ರು .. ಅಂದು ಜನಿಸಿದರು." ಎಂದು ಭಾಷಾಂತರಿಸಲಾಯಿತು.

ಈ ರೀತಿ ಸೃಷ್ಟಿಸಲ್ಪಟ್ಟ ಕನ್ನಡದ ಕೋಶದಲ್ಲಿ ಸುಮಾರು ಒಂದು ಕೋಟಿ (1,00,00,000) ವಿವಿಧ ವಿಷಯಗಳ ಬಗ್ಗೆ 2.5 ಕೋಟಿ (2,50,00,000) ಮಾಹಿತಿಯ ಅಂಶಗಳಿವೆ. ಇವುಗಳಿಂದ 2.5 ಕೋಟಿ ಸರಳವಾದ ಕನ್ನಡ ವಾಕ್ಯಗಳನ್ನು ರಚಿಸಲಾಯಿತು. ನಂತರ ಈ ವಾಕ್ಯಗಳನ್ನು ಮಾನವ ಭಾಷಾ ಉತ್ಪತ್ತಿ (Natural Language Generation) ತಂತ್ರಶಾಸ್ತ್ರದ ಬಳಕೆಯಿಂದ ವಾಕ್ಯವೃಂದ (paragraph) ಗಳಾಗಿಯೂ, ತದನಂತರ ಅವುಗಳ ಸೂಕ್ತವಾದ ಸರಣಿಗಳಿಂದ ಇಡೀ ಪುಟಗಳನ್ನೂ ಸಹ ರಚಿಸಲಾಯಿತು. ಒಂದೇ ವಿಷಯವನ್ನು ಕುರಿತ ವಾಕ್ಯಗಳನ್ನು ಒಟ್ಟುಗೋಡಿಸಿ ಉದ್ದವಾದ ವಾಕ್ಯಗಳನ್ನೂ ಸಹ ಅಲ್ಲಲ್ಲಿ ರಚಿಸಲಾಯಿತು. ಈ ಎಲ್ಲಾ ಹಂತಗಳಲ್ಲಿಯೂ, ಕೆನೋ (KAnOE) ದಲ್ಲೇ ಸಂಶೋಧನೆ ಮಾಡಿ ಉತ್ಪತ್ತಿಮಾಡಿದ ಒಂದು ಮೂಲತತ್ತ್ವವು (ontology) ಅತಿ ಮುಖ್ಯವಾದ ಪಾತ್ರವನ್ನು ವಹಿಸಿತೆಂಬುದು ನಮಗೆ ಹೆಮ್ಮೆಯ ವಿಷಯ.

ಅತಿ ಚಿಕ್ಕದಾದ ಪುಟಗಳನ್ನು ತೆಗೆದುಹಾಕಿದ ನಂತರ ಹತ್ತು ಲಕ್ಷ (10,00,000) ಕನ್ನಡ ಪುಟಗಳ *ಉಲ್ಲೇಖ ಮಾಲಾ* ಸಿದ್ಧವಾಯಿತು. ಈ ಪ್ರತಿಯೊಂದು ಕಾರ್ಯಕ್ಕೂ ಅತಿ ಹೆಚ್ಚಿನ ಶಕ್ತಿ ಮತ್ತು ವೇಗದ ಕಂಪ್ಯೂಟರ್ ಗಳು (servers with 2 CPU, 12 core, 96 GB RAM and 8x500 GB RAID disks) ಕೆನೋ ದಲ್ಲಿ ಲಭ್ಯವಿದ್ದರೂ ಸಹ ದಿನಗಟ್ಟಲೆ, ವಾರಗಟ್ಟಲೆ ಹಿಡಿಯಿತು. ಒಂದು ಉದಾಹರಣೆಯಿಂದ ಇದು ಸ್ಪಷ್ಟವಾಗುತ್ತದೆ: ಸರಾಸರಿ 800 ಕನ್ನಡ ಪದಗಳಲ್ಲಿರುವ ತಿದ್ದುಪಡಿಯನ್ನು ಎಲ್ಲ ದಶ ಲಕ್ಷ ಪುಟಗಳಲ್ಲೂ ಮಾಡಲು 10,000 ಕೋಟಿ (10,000,00,00,000 ಅಥವಾ 100 Billion) ಸಲ ಪದಗಳ ತುಲನೆ ಮಾಡಬೇಕಾಗುತ್ತದೆ; ಇದಕ್ಕೆ ಏನಿಲ್ಲವೆಂದರೂ ಒಂದು ವಾರ ಕಂಪ್ಯೂಟರ್ ಅನ್ನು ತಡೆಯಿಲ್ಲದೆ ಚಾಲನೆ ಮಾಡಬೇಕಾಯಿತು.

ಉಲ್ಲೇಖ ಮಾಲಾ ದಲ್ಲಿರುವ ಕನ್ನಡ ಪುಟಗಳು ಅತ್ಯಂತ ಶುದ್ಧವಾಗಿಯೂ ಇಲ್ಲ; ಅಷ್ಟೇನೂ ಸಂಪೂರ್ಣವಾಗಿಯೂ ಇಲ್ಲ. ಆದರೆ, ಅವೆಲ್ಲವೂ ಎಲ್ಲರಿಗೂ ಶುಲ್ಕರಹಿತವಾಗಿ ಅಂತರ್ಜಾಲದಲ್ಲಿ ಲಭ್ಯವಿದೆ. ಈ ಪುಟಗಳನ್ನು ಯಾರು ಬೇಕಾದರೂ, ಬೇರೆ ಯಾವ ಸಾಫ್ಟ್ ವೇರ್ ನ (ಲಿಪಿ, ನುಡಿ, ಬರಹ ಮುಂತಾದ) ಅವಶ್ಯಕತೆಯಿಲ್ಲದೆ ತಿದ್ದಲು ಸಾಧ್ಯವಿದೆ; ಹೆಚ್ಚಿನ ಮಾಹಿತಿಗಳನ್ನು ಸೇರಿಸಬಹುದು; ವಾಕ್ಯಗಳ ಮತ್ತು ಕನ್ನಡ ಪದಗಳ, ಭಾಷೆಯ ಗುಣಮಟ್ಟವನ್ನು ಹೆಚ್ಚಿಸಬಹುದು. ಉಲ್ಲೇಖ ಮಾಲಾ http://kanoe.org/kannada.html ನಲ್ಲಿ ಲಭ್ಯವಿದೆ.

ಬಹಳಷ್ಟು ಜನ ಕನ್ನಡಿಗರು ಭಾಗವಹಿಸಿ ಉಲ್ಲೇಖ ಮಾಲಾವನ್ನು ತಿದ್ದಿ, ಬೆಳೆಸಿ, ಎಲ್ಲೆಲ್ಲಿರುವ ಕನ್ನಡದ ಬಂಧುಬಾಂಧವರಿಗೆ ಒಂದು ಅತ್ಯುಪಯುಕ್ತವಾದ ಮಾಹಿತಿ ಕೋಶವನ್ನಾಗಿ ಮಾಡುತ್ತಾರೆಂದು ನಂಬಿದ್ದೇವೆ.