भाषा प्रौद्योगिकी विभाग में सॉफ्टवेयर विकास कार्यक्रम

विकासकर्ता :. डॉ. धनजी प्रसाद
असिस्टेंट प्रोफेसर, भाषा प्रौद्योगिकी विभाग
म.गा.अं.हिं.वि., वर्धा

विकसित सॉफ्टवेयर

1. **कुशल :** हिंदी वर्तनी जाँचक (यूनिकोड हेतु) (KUSHAL : Hindi Spell Checker)
2. **रूपविश्लेषक :** रूपवैज्ञानिक रूप विश्लेषक (ROOPVISHLESHAK : Morphological Form Analyzer)
3. **रूपसर्जक :** रूपवैज्ञानिक रूप प्रजनक (ROOPSARJAK : Morphological Form Generator)
4. **हिंटै :** संदर्भ-मुक्त पी.ओ.एस टैगर (HINTAI : Context-Free POS Tagger)
5. **खोजी :** संदर्भ में शब्द प्राप्तकर्ता (KHOJEE : Keyword in context Founder)
6. **अंतरक :** देवनागरी रोमन लिप्यंतरण प्रणाली (ANTARAK : Devanagari Roman Transliteration System)
7. **गणक :** शब्द आवृत्ति गणक (GANAK : Word Frequency Counter)
8. **सामान्यक** : विराम चिह्न सामान्यीकारक (SAMANYAK : Punctuation Mark Normalizer)
9. **अन्वेषक :** कोशीय इकाई (कोशिम) प्राप्तकर्ता (ANVESHAK : Lexical Entry (Lexeme) Finder)
10. **शब्दनिधि** : हिंदी-अंग्रेजी द्विभाषिक शब्दकोश(द्विदिशीय) (SHABDANIDHI : Hindi-English Bilingual Dictionary (bidirectional))

उपर्युक्त सॉफ्टवेयरों के अलावा 3 पर कार्य जारी है और भविष्य की योजना में दो सॉफ्टवेयरों को रखा गया है, जो निम्नलिखित हैं -

1. **संहिंटै : संदर्भयुक्त पी.ओ.एस टैगर (SANHINTAI : Context Sensitive POS Tagger)**
2. **विशेषज्ञ : हिंदी पदबंध चिह्नक (VISHESHGYA : Hindi Phrase Marker)**
3. **हंश : हिंदी शब्द विसंदिग्धिकारक (HANSH : Hindi Word Disambiguator)**
4. **ज्ञाता : हिंदी पद-विच्छेदक (GYATA : Hindi Parser)**
5. **पाणिनी : हिंदी का व्याकरण जाँचक (PANINI : Hindi Grammar Checker)**

विकसित सॉफ्टवेयरों का परिचय

1. कुशल : हिंदी वर्तनी जाँचक(यूनिकोड हेतु) (KUSHAL : Hindi Spell Checker)

हिंदी लेखन और टाइपिंग में वर्तनी अक्सर समस्या उत्पन्न करती है। वर्तमान में इस संबंध में कुछ टूल विकसित किए जा चुके हैं। यह टूल भी इसी शृंखला की एक कड़ी है। (अन्य टूलों के सापेक्ष) इसकी एक मुख्य विशेषता यह है कि इसे पूरी तरह से भाषावैज्ञानिक विधि से डिजाइन किया गया है। इसका विकास करते हुए मुख्य ध्यान इस बात पर केंद्रित किया गया है कि टाइपिंग में जो गलती टंकक द्वारा की जाएगी उसका सबसे सटीक सुझाव क्या हो सकता है। इस कारण यह टूल बहुत सारे विकल्प देने के बजाए कुछ सीमित विकल्प देता जो संबंधित शब्द के बिल्कुल करीब होते हैं।

साथ-ही इसमें एक विशेषता है कि जैसे ही आप जाँच करें बटन को क्लिक करते हैं इसमें सबसे पहले सभी त्रुटिपूर्ण वर्तनी शब्द लाल हो जाएँगे किंतु जिन शब्दों को आप डबल क्लिक करके राइट क्लिक करेंगे केवल उनके ही सुझाव यह टूल प्रस्तुत करेगा। इससे कार्य तेजी से हो जाता है और नाम पदों के लिए सॉफ्टवेयर को अनावश्यक मेहनत नहीं करनी पड़ती है।

इस टूल का अंतरापृष्ठ (Interface) इस प्रकार से डिजाइन किया गया है-

इसमें फाइल खोलें बटन से या कॉपी पेस्ट के माध्यम से बड़े टेक्स्टबॉक्स में वर्तनी जाँच हेतु पाठ दिया जाता है। अब 'जाँच करें' बटन को क्लिक करने पर इसी टेक्स्टबॉक्स में वे सभी शब्द लाल रंग के हो जाएँगे जिनकी वर्तनी में कुछ त्रुटि होगी। इनमें से किसी भी शब्द डबल क्लिक करके (सलेक्ट करके) राइट बटन क्लिक करने 'सुझाव' वाले लिस्टबॉक्स में संबंधित शब्द आ जाएँगे। अब उनमें से किसी एक शब्द को क्लिक करने पर वह शब्द मूल पाठ में संबंधित शब्द के स्थान पर आ जाएगा।

"हिन्दी सम्पूर्ण भारत की राजभाषा है। आज इसकि स्थिति पहले से बहुत सुधर गयी है। हमे इस पर गर्व होन चाहिये।" वाक्य का इनपुट देकर इसके परिणामों के कुछ नमूने नीचे दिए गए हैं-

यह सॉफ्टवेयर हमारे पूर्व के सॉफ्टवेयर की तुलना में कई दृष्टियों से अलग है। दो मुख्य बातें इस प्रकार हैं-

1. **गति :** 'सक्षम' में एक शब्द के सुझाव देनें 5-10 मिनट लगते हैं, इसमें 1 से 2 सेकेंड। (डाटाबेस में कुल शब्द : 1 लाख 40 हजार)

2. **सुझावों की संख्या** : इसमें प्रत्येक शब्द के लिए सीमित किंतु सबसे सटीक सुझाव सामान्यत: 1 से 5 और विशेष परिस्थितियों में 5 से 10 सुझाव दिए जाते हैं।

---

2. रूपविश्लेषक : रूपवैज्ञानिक रूप विश्लेषक (ROOPVISHLESHAK : Morphological Form Analyzer)

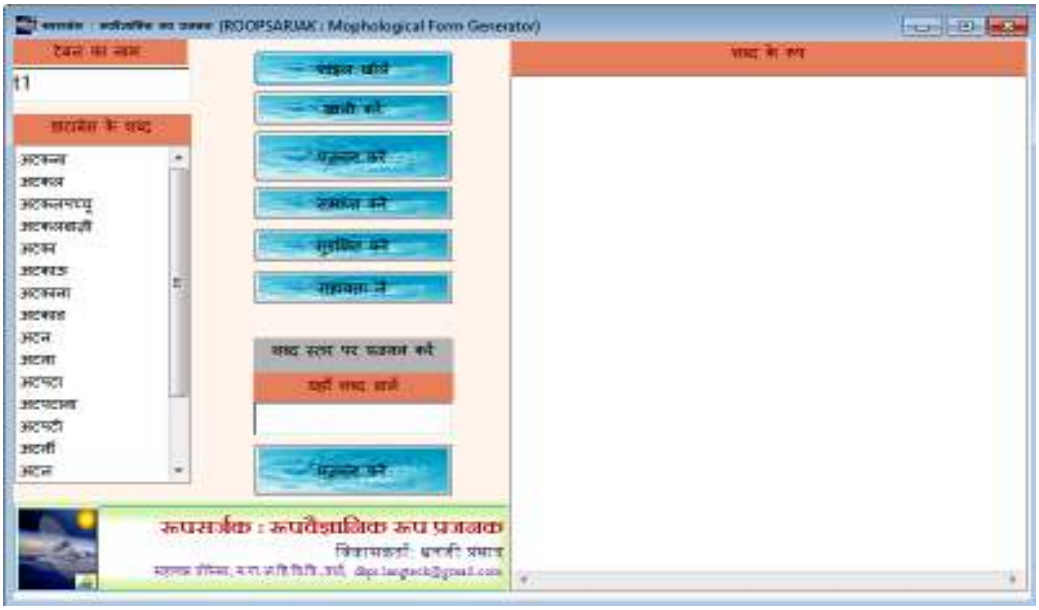
इनके विश्लेषण हेतु 'रूपविश्लेषक' का विकास किया गया है। इस टूल में वाक्य में आए शब्दों के विश्लेषण हेतु व्यवस्था है जहाँ शब्द के साथ-साथ उसके मूल रूप को भी प्रस्तुत किया जाता है। विकसित प्रणाली के अंतरापृष्ठ में इनपुट देकर 'विश्लेषण करें' बटन को क्लिक करने के पश्चात डाटा इस प्रकार दिखाई देता है -

इसे 'आउटपुट प्रिंट करें' बटन को क्लिक करके वर्ड फाइल में प्रिंट किया जा सकता है।

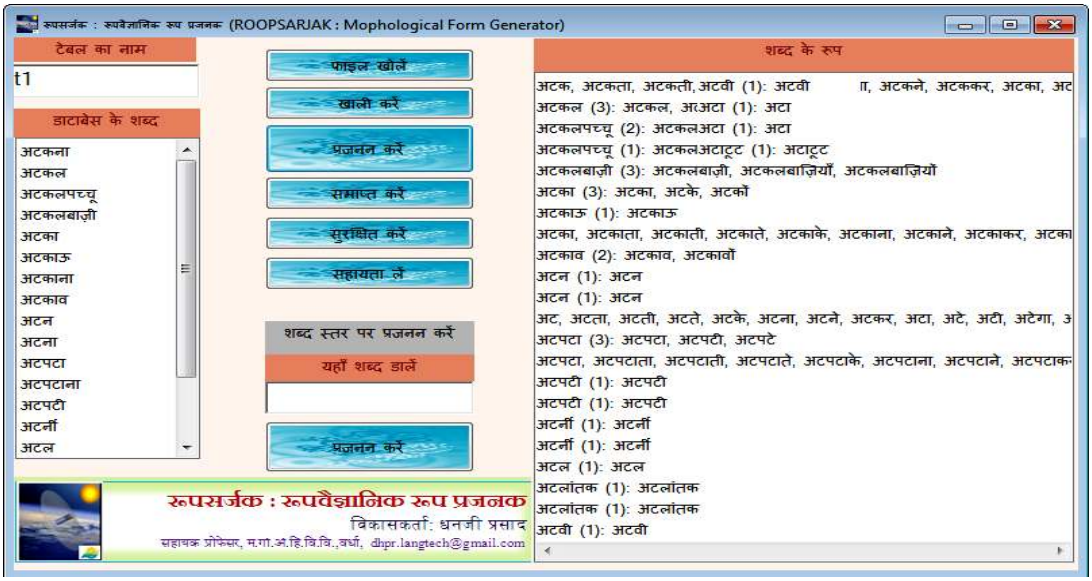
3. रूपसर्जक : रूपवैज्ञानिक रूप प्रजनक (ROOPSARJAK : Morphological Form Generator)

रूपसर्जक एक ऐसी प्रणाली है जो हिंदी के किसी भी कोशीय शब्द के बनने वाले सभी रूपों को निर्मित करती है। इससे हिंदी शब्दों के विभिन्न वाक्यात्मक रूपों को देखा जा सकता है। कोश में शब्दों के मूल रूप ही संग्रहीत होते हैं। जब उन शब्दों का वाक्य में व्यवहार होता है तो विभिन्न व्याकरणिक कोटियों (जैसे- लिंग, वचन, पुरुष, काल आदि) के आधार पर कुछ परिवर्तन होता है। यह परिवर्तन कई प्रकार का होता है। कभी मूलशब्द के साथ कुछ प्रत्यय जुड़ जाते हैं तो कभी पूरे के पूरे शब्द में ही परिवर्तन हो जाता है। उदाहरण के लिए क्रिया शब्दों के बनने वाले विभिन्न रूपों को देखा जा सकता है, जैसे- 'खा' धातु के भूतकालिक रूप निर्माण के लिए जब इसके साथ 'या' प्रत्यय का प्रयोग होता है तो 'खाया' रूप बनता है। किंतु 'जा' धातु का यही रूप निर्मित करने पर 'गया' बनता है। अत: हिंदी के कोशीय शब्दों के बनने वाले सभी रूपों का ज्ञान आवश्यक है। वैसे मुख्य रूप से संज्ञा, सर्वनाम, क्रिया और विशेषण शब्दों में ही विकार होते हैं।

यह प्रणाली हिंदी शब्दों के सभी बनने वाले रूपों को स्वचलित रूप से निर्मित करने के लिए विकसित की गई है। इसमें रूप-निर्माण हेतु इनपुट देने के लिए दो प्रकार की व्यवस्थाएँ दी गई हैं- प्रथम यदि बहुत सारे शब्दों के रूपों को एक साथ निर्मित करके देखना हो तो उन्हें किसी डाटाबेस में सुरक्षित करें और उसे खोलकर सीधे-सीधे टेबल से ही शब्दों को लेकर सभी रूप निर्मित किए जा सकते हैं। इसके अलावा दूसरी विधि है- यदि केवल एक शब्द का रूप निर्मित करना हो तो दिए हुए टेक्स्टबॉक्स में उसे टाइप करें और 'प्रजनन करें' बटन को क्लिक करने पर उसके सभी रूप निर्मित हो जाएँगे। इस टेक्स्टबॉक्स में एक से अधिक शब्द भी स्पेस से अलग करते हुए दिए जा सकते हैं। प्रणाली के अंतरापृष्ठ में 'फाइल खोलें' बटन द्वारा डाटाबेस फाइल को खोला जाता है, किंतु उसके पहले डाटाबेस में बने टेबल का नाम 'टेबल का नाम' नामक टेक्स्टबॉक्स में देना होगा इसके बाद यह उससे शब्दों को लेकर यह नीचे बने लिस्टबॉक्स में प्रदर्शित करता है, जैसे-

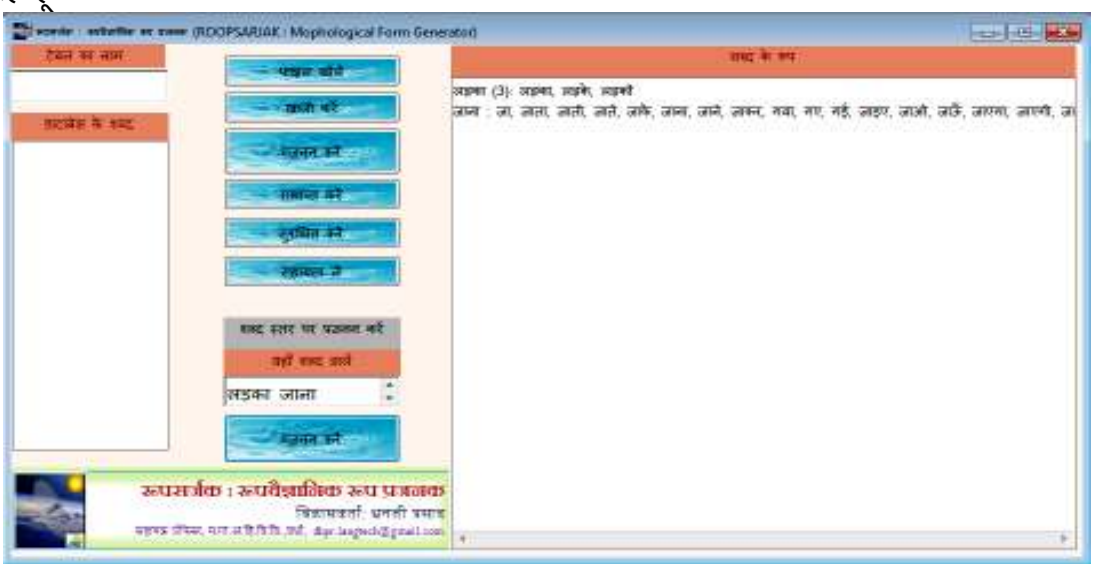


अब इसके बाद तीसरे नंबर पर बने 'प्रजनन करें' बटन को क्लिक करने पर सभी शब्दों के रूप सामने निर्मित होकर आ जाएँगे-



अब इन शब्दरूपों को 'सुरक्षित करें' बटन द्वारा वर्ड फाइल में प्रिंट किया जा सकता है।

इसके अलावा शब्द स्तर पर प्रिंट करने के लिए नीचे दिए गए छोटे से टेक्स्टबॉक्स में शब्द दिए जा सकते हैं। इसमें एक शब्द भी हो सकता है और एक से अधिक शब्द भी हो सकते हैं। उदाहरण के लिए मैं यहाँ पर दो शब्दों का इनपुट देकर उनके रूप प्रजनित कर रहा हूँ-



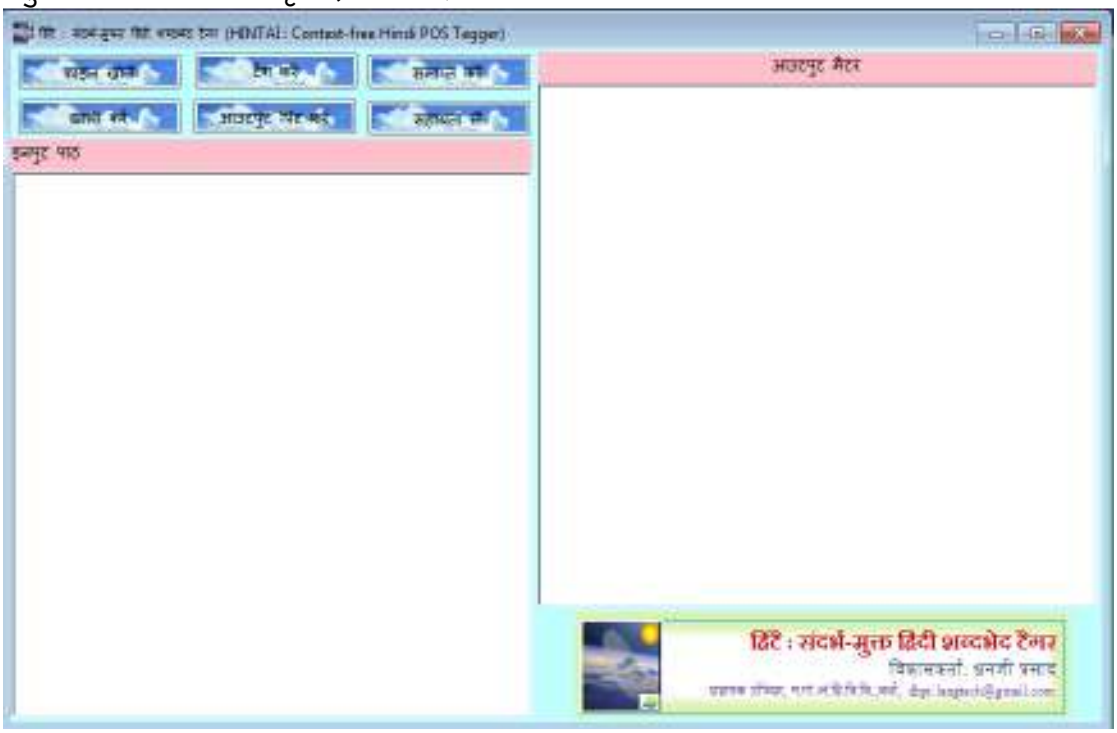
इसी प्रकार कितने भी शब्दों के विविध रूप निर्मित किए जा सकते हैं।

अत: रूपसर्जक एक अत्यंत ही उपयोगी सॉफ्टवेयर है जो हिंदी के कोशीय शब्दों के सभी व्याकरणिक रूपों को प्रजनित करता है। इस दृष्टि से यह हिंदी को तकनीकी के क्षेत्र में सर्जक दृष्टि से एक नया आयाम प्रदान करता है। इससे हिंदी की तकनीकी पूर्णता में एक कदम माना जा सकता है।

4. हिंटै : संदर्भ-मुक्त पी.ओ.एस. टैगर (HINTAI : Context-Free POS Tagger)

पी.ओ.एस. टैगिंग प्राकृतिक भाषा संसाधन की एक आधारभूत प्रक्रिया है। इसके बाद ही भाषिक विश्लेषण या प्रजनन से संबंधित कार्य संपन्न होते हैं। टैगिंग द्वारा किसी पाठ के वाक्यों में आए सभी शब्दों को उनके शब्दवर्ग के आधार पर एक टैग प्रदान किया जाता है। यह कार्य दो प्रकार से संभव है- संदर्भ-मुक्त और संदर्भ-युक्त। संदर्भ-मुक्त टूल द्वारा कोशीय स्थिति के आधार पर शब्द के टैग को प्रदान किया जाता है। ऐसी स्थिति में यदि किसी शब्द के दो कोशीय मूल प्राप्त होते हैं (जैसे- आम, आम या खाना, खाना आदि) तो यह प्रणाली दोनों टैगों को दिखाएगी। संदर्भ-युक्त प्रणाली वाक्यात्मक संदर्भ का विश्लेषण करते हुए वास्तविक टैग को ही प्रदान करती है। यह प्रणाली संदर्भ-मुक्त है। अत: यह शब्द के साथ उसके सभी संभव टैगों को प्रदान करती है।

इसमें इनपुट और आउटपुट के लिए सुविधा दी जाती है। संसाधन का निर्देश देने के लिए कोई बटन आदि के रूप में प्रतीक होता है। प्रस्तुत सॉफ्टवेयर का अंतरापृष्ठ इस प्रकार है-



इसमें किसी पाठ का इनपुट देने पर आउटपुट इस प्रकार से आएगा-

टैग निर्धारण के बाद टैगसेट बन जाता है। इसके पश्चात् टैग करने वाले सॉफ्टवेयर का निर्माण करना होता है। इसके दो पक्ष हैं- अंतरापृष्ठ एवं प्रक्रिया। अंतरापृष्ठ को ऊपर दिखाया जा चुका है। इसमें 'टैग करें' बटन को क्लिक करने के पश्चात् जो प्रक्रिया होती है, वह अत्यंत महत्वपूर्ण है। इसी में टैगिंग के भाषावैज्ञानिक नियम लगे होते हैं।

---

5. खोजी : संदर्भ में शब्द प्राप्तकर्ता (KHOJEE : Keyword in Context Finder)

भाषिक शोध में शब्दों के वाक्यात्मक या पदबंधीय संदर्भों का विशेष महत्व होता है। किसी संदिग्धार्थक शब्द का वास्तविक अर्थ उसके आस-पास के शब्दों को देखकर ही निर्धारित किया जाता है। इसके अलावा बहुत सारे शब्दों को उनके वाक्यात्मक प्रयोग द्वारा ही पूर्णत: समझा जा सकता है। वर्तमान में कार्पस भाषाविज्ञान के अंतर्गत भी संदर्भ में शब्द प्राप्त करने की पद्धति को विशेष महत्व प्राप्त हुआ है। इसके लिए प्रयुक्त टूल को तकनीकी रूप से कांकार्डेंस प्रोग्राम कहा जाता है। यह टूल भी इसी कार्य को संपन्न करता है।

'संदर्भ में शब्द' के लिए सर्वप्रथम Hans Peter Luhn द्वारा KWIC शब्द का प्रयोग किया गया। ऐसी पंक्तियाँ या वाक्य जिनमें कोई शब्द समान संदर्भ में आया हो, कांकार्डेंस वाक्य (या पंक्ति) कहलाता है। आरंभ में यह कार्य मुख्यत: धार्मिक ग्रंथों (जैसे- वेद, कुरान, बाइबिल आदि) में एक संदर्भ में किसी शब्द में विभिन्न वाक्यों में प्रयोग को देखने हेतु किया गया। किंतु धीरे-धीरे यह पद्धति अत्यंत उपयोगी सिद्ध हुई और व्यापक रूप से अन्य कार्यों में भी प्रयोग में लाई जाने लगी। जबसे कार्पस विश्लेषण का कार्य आरंभ हुआ है, कांकार्डेंस वाक्यों का महत्व बहुत अधिक बढ़ गया है। इसे देखते हुए प्रत्येक कार्पस हेतु एक कांकार्डेंस प्रोग्राम विकसित किया जा रहा है। किंतु स्वतंत्र रूप से भी कांकार्डेंस प्रोग्रामों की बहुत अधिक आवश्यकता है जिससे कि ये धीरे-धीरे लोकप्रिय होते जा रहे हैं।

इसे ही ध्यान में रखते हुए 'हिंदी' पाठों के लिए खोजी नाम से यह सॉफ्टवेयर तैयार किया गया है। इसके अंतरापृष्ठ में आप विशाल पाठ का इनपुट देकर खोज हेतु शब्द देते हैं। इसके पश्चात् आगे और पीछे की शब्द-संख्या को इंटर करने के बाद यह प्रोग्राम संबंधित आउटपुट को प्रदान करता है और उसे वर्ड फाइल में प्रिंट कर देता है। इसका अंतरापृष्ठ इस प्रकार है-

इसमें फाइल खोलें बटन से किसी भी वर्ड फाइल को इनपुट बॉक्स में खोल सकते हैं। इसके अलावा कहीं से डाटा कॉपी करके पेस्ट भी किया जा सकता है। तत्पश्चात् 'खोजशब्द' के नीचे दिए गए बॉक्स में वह शब्द डालें जिसे आप संदर्भ के साथ देखना चाहते हैं। फिर उसके नीचे पूर्व और पश्च शब्दों की संख्या दें। उदाहरण के लिए नीचे 'में' शब्द को 2 शब्द आगे और 2 शब्द पीछे से संदर्भ के साथ एक टेक्स्ट में इस प्रकार देखा गया है-

इसमें एक और विकल्प केवल वाक्य का भी है। इसे क्लिक करने पर केवल वे वाक्य प्रस्तुत किए जाएँगे जिनमें 'में' शब्द आया हो। इसके अलावा आप चाहें तो दोनों को भी देख सकते हैं। 'खाली करें' बटन द्वारा इनपुट और आउटपुट बॉक्सों को खाली कर दिया जाता है जिससे कि अगले मैटर पर कार्य किया जा सके। 'आउटपुट प्रिंट करें' बटन से आउटपुट को एक नई वर्ड फाइल में प्रिंट कर दिया जाता है।

6. अंतरक : देवनागरी रोमन लिप्यंतरण प्रणाली (ANTARAK : Devanagari Roman Transliteration System)

'लिप्यंतरण' में एक लिपि में लिखी गई पाठ सामग्री को दूसरी लिपि में अंतरित कर दिया जाता है। लिपि किसी भाषिक अभिव्यक्ति को भौतिक रूप से प्रस्तुत करने की एक व्यवस्था है जो भाषा निरपेक्ष होती है। अर्थात् किसी भी भाषा की अभिव्यक्ति को किसी भी लिपि में लिखा जा सकता है। यह हो सकता है कि संबंधित भाषा की कुछ ध्वनियों को प्रस्तुत करने के लिए उस लिपि में सटीक वर्ण न हो। ऐसी स्थिति में चिह्न विशेष के प्रयोग या कुछ वर्णों के समुच्चय का प्रयोग करते हुए वैकल्पिक व्यवस्था की जाती है। उदाहरण के लिए हिंदी की महाप्राण ध्वनियों को व्यक्त करने के लिए रोमन लिपि में वर्ण नहीं मिलते। ऐसी स्थिति में संबंधित वर्ण के समतुल्य वर्ण के साथ 'h' का प्रयोग किया जाता है, जैसे- हिंदी के 'क' के लिए रोमन में 'k' का प्रयोग किया जाता है। किंतु 'ख' के लिए कोई सीधे-सीधे वर्ण उपलब्ध नहीं है, इस कारण 'kh' का प्रयोग किया जाता है। यही स्थिति अन्य महाप्राण ध्वनियों के साथ देखी जा सकती है। इसी प्रकार रोमन के कैपिटल लेटर को दर्शाने के लिए देवनागरी में कोई व्यवस्था नहीं है। फिर भी काम चल जाता है।

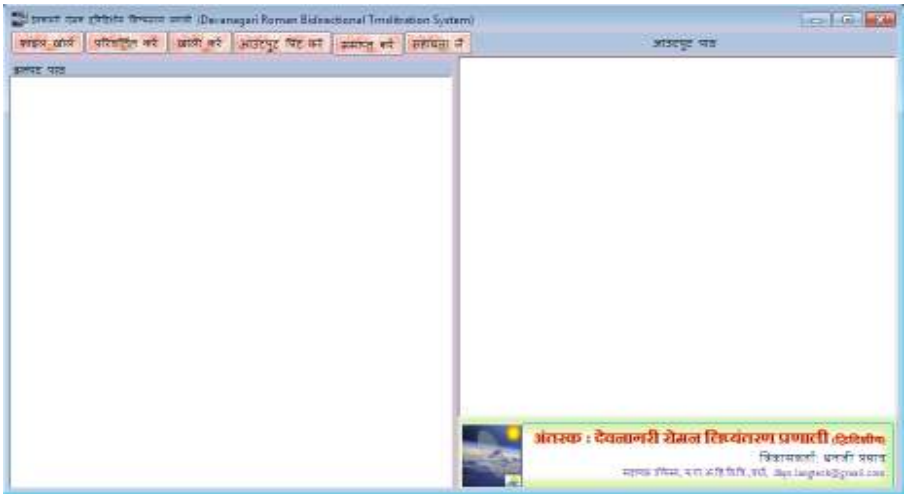
एक ही भाषा की सामग्री को एक से अधिक लिपियों में लिखा जा सकता है। बहुभाषिक समाजों में कई जगहों पर ऐसी स्थिति में देखी जा सकती है कि एक से अधिक भाषाओं के साथ-साथ एक से अधिक लिपियाँ भी प्रचलित होती हैं। ऐसे समाजों में कुछ व्यक्ति ऐसे भी होते हैं जो किसी भाषा के वाचिक रूप को तो समझ लेते हैं किंतु उन्हें लिपि का अभ्यास नहीं होता। उदाहरण के लिए हिंदी समझने वालों की संख्या बहुत अधिक है। परंतु इनमें बहुत सारे लोग ऐसे हैं जो देवनागरी में लिखी हिंदी को समझ नहीं सकते। ऐसी स्थिति में यदि उन्हें हिंदी की सामग्री उनकी संबंधित लिपि जैसे- रोमन, अरबी/फारसी या किसी दक्षिण भारतीय भाषा की लिपि में दे दी जाए तो वे समझ सकते हैं। अत: यदि सामग्री हाथ से लिखी हुई हो तो दूसरी लिपि में उसका पुन:लेखन करना होगा। किंतु यदि यह सामग्री कंप्यूटर के माध्यम से टाइप की गई हो और उन्हें किसी सॉफ्टवेयर की सहायता से परिवर्तित कर दिया जाए तो यह कार्य अत्यंत सरल हो जाएगा।

इसी प्रकार की आवश्यक्ताओं को ध्यान में रखते हुए 'अंतरक' नाम से यह लिप्यंतरण प्रणाली विकसित की गई है। यह प्रणाली 'देवनागरी से रोमन' और 'रोमन से देवनागरी' दो लिपियों में दोनों ही दिशाओं से लिप्यंतरण हेतु विकसित की गई है। यह प्रणाली देवनागरी पाठ होने पर यूनिकोड टाइपिंग में ही कार्य करती है। इसमें प्रयोक्ता को दिशा नहीं बताने की आवश्यकता है। जैसे ही इसे इनपुट के रूप में पाठ सामग्री प्राप्त होगी वैसे ही यह उसकी लिपि की पहचान कर लेगा और दूसरी लिपि में सामग्री को परिवर्तित कर देगा। इस प्रणाली के निर्माण में दोनों ही दिशाओं से कार्य करने के लिए अलग-अलग प्रकार के नियमों का प्रयोग किया गया है। इन्हें नीचे संक्षेप में दिया जा रहा है-

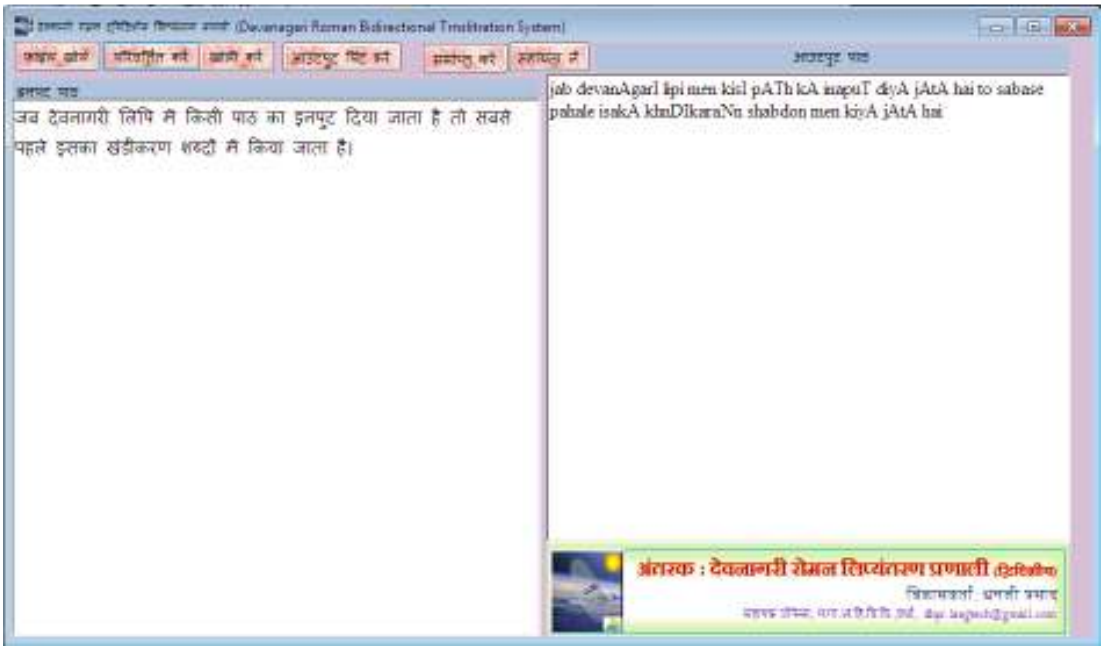
(1) देवनागरी से रोमन

जब देवनागरी लिपि में किसी पाठ का इनपुट दिया जाता है तो सबसे पहले इसका खंडीकरण शब्दों में किया जाता है। इसके पश्चात् प्रत्येक शब्द का खंडीकरण वर्णों में किया जाता है। अब प्रत्येक वर्ण के लिए डाटाबेस में संबंधित रोमन वर्ण से मिलान किया जाता है। मिलान के बाद जो भी रोमन वर्ण प्राप्त होते हैं उन्हें शब्द के रूप में संगठित किया जाता है और फिर शब्दों को जोड़ते हुए नए पाठ को प्रजनित किया जाता है। इस प्रक्रिया में वर्णों की विविधता के आधार पर अलग-अलग प्रकार के नियम प्रयुक्त होते हैं।

यद्यपि देवनागरी को वैज्ञानिक लिपि माना जाता है फिर भी इसका विश्लेषण कर सीधे-सीधे लिप्यंतरण संभव नहीं हो पाता है। इसमें वर्ण विश्लेषणात्मक नियमों की आवश्यकता पड़ती है। इस संबंध में एक बात कही जा सकती है कि कुछ कठिनाइयाँ लक्ष्य लिपि के स्वरूप के कारण भी आती हैं। इस प्रणाली का अंतरापृष्ठ इस प्रकार है-



इनमें देवनागरी पाठ का इनपुट देने पर आउटपुट इस प्रकार प्राप्त होता है-



इसमें देखा जा सकता है कि दीर्घ मात्राओं को प्रदर्शित करने के लिए कैपिटल लेटर्स का प्रयोग किया गया है।

(2) रोमन से देवनागरी

यूनिकोड के आगमन के पश्चात से ऐसे टूल्स की माँग बढ़ी है जिनकी सहायता से रोमन अक्षरों को टाइप करने पर देवनागरी अक्षर आ जाएँ। यह टूल इस सुविधा को प्रदान करता है। जैसे ही आप किसी शब्द को पूरा करके स्पेस दबाएँगे उसका देवनागरी रूप बगल के टेक्स्टबॉक्स में आ जाएगा। इसके अलावा आप अपने देवनागरी पाठ को भी इस टूल के माध्यम से रोमन में परिवर्तित कर सकते हैं।

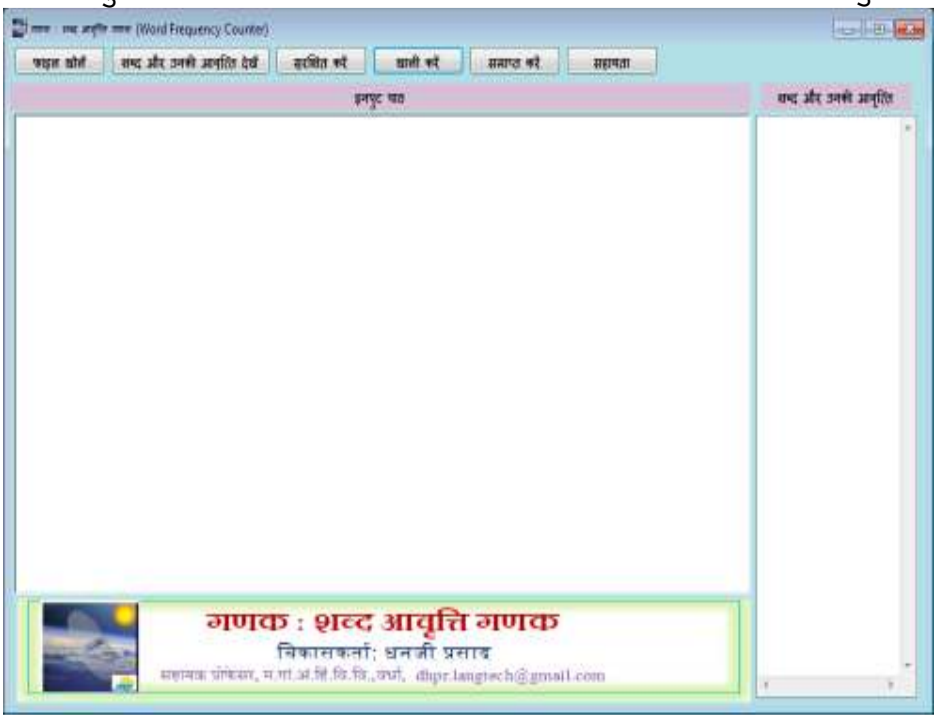
इसका विकास कार्य अभी जारी है।

7. गणक : शब्द आवृत्ति गणक (GANAK : Word Frequency Counter)

यह एक ऐसा टूल है जो किसी पाठ में आए प्रत्येक शब्द को केवल बार लिखकर सूचीबद्ध करता है और साथ-ही प्रत्येक शब्द के आगे यह भी सूचना प्रदान करता है कि वह शब्द उस पाठ में कितनी बार प्रयुक्त हुआ है। अत: किसी भाषा के विभिन्न प्रकार के पाठों में प्रयुक्त होने वाले शब्दों एवं उनकी आवृत्ति की सूची प्रदान करने की दृष्टि से यह एक अत्यंत ही उपयोगी सॉफ्टवेयर है।

अत: इसकी उपयोगिता को देखते हुए एक 'शब्द आवृत्ति गणक' (Word Frequency Counter) का विकास यहाँ पर किया गया है। इसका प्रयोग मुख्यत: हिंदी भाषा या देवनागरी लिपि में लिखी जाने वाली भाषाओं के लिए किया जा सकता है। साथ-ही यह अन्य भाषाओं (जैसे-अँग्रेजी आदि) के पाठों के साथ भी कार्य करने के लिए उपयुक्त है।

अंतरापृष्ठ डिजाइन : इस टूल में इनपुट हेतु एक टेक्स्टबॉक्स और आउटपुट हेतु एक लिस्टबॉक्स का प्रयोग किया गया है। संसाधन कार्य कुछ विशेष चरणों में संपन्न होता है जिसे ऊपर कार्यविधि में दिखाया जा चुका है। इसका अंतरापृष्ठ इस प्रकार है-



इसमें कुल 5 बटनों का प्रयोग किया गया है, जो इस प्रकार हैं-

1. **फाइल खोलें :** इसके द्वारा किसी भी वर्ड फाइल को सीधे-सीधे टेक्स्टबॉक्स में खोला जा सकता है। इस बटन को क्लिक करते ही एक OpenFileDialog आपके सामने आएगा जिसमें आप संबंधित फाइल को क्लिक करके ओपेन बटन पर क्लिक करते ही संबंधित फाइल का मैटर टेक्स्टबॉक्स में लोड हो जाएगा।

2. **शब्द और उनकी आवृत्ति देखें :** इस बटन को क्लिक करने पर यह टूल पाठ में शब्दों की आवृत्ति की गणना करेगा और उन्हें आउटपुट के रूप में निर्मित कर लिक्स्टबॉक्स में प्रदर्शित करेगा। इनपुट देकर इस बटन को क्लिक करने पर इस प्रकार परिणाम दिखेगा-

3. **सुरक्षित करें :** इस बटन को क्लिक करने पर एक नई वर्ड फाइल निर्मित हो जाएगी और आउटपुट के रूप में प्राप्त मैटर उसमें राइट कर दिया जाएगा। अब आप उस फाइल को कोई नाम देकर सुरक्षित कर सकते हैं।
4. **खाली करें :** इस बटन द्वारा इनपुट और आउटपुट स्थानों को खाली कर दिया जाता है।
5. **समाप्त करें :** इस बटन को क्लिक करके प्रोग्राम को बंद किया जाता है।
6. **सहायता लें :** इस बटन को क्लिक करने पर एक नई विंडो खुलती है जिसमें निम्नलिखित सूचनाएँ होती हैं-

---

8. सामान्यक : विराम चिह्न सामान्यीकारक (SAMANYAK : Punctuation Mark Normalizer)

यह एक छोटा किंतु महत्वपूर्ण टूल है। हिंदी (देवनागरी) में टाइपिंग के समय प्राय: लोगों द्वारा विराम चिह्नों के प्रयोग संबंधी अशुद्धियाँ या त्रुटियाँ हो जाती है। विशेष रूप से यह स्थिति तब देखने को मिलती है जब टंकक को देवनागरी लेखन या टंकण का उचित या पर्याप्त ज्ञान न हो या प्रयोक्ता ने अभी नई-नई टाइपिंग सीखी हो। यह टूल यूनीकोड में टाइपिंग करते हुए की जाने वाली विराम चिह्न संबंधी 50 प्रकार की त्रुटियों को सुधार सकता है।

किसी भी भाषा में लेखन अथवा टाइपिंग में विराम-चिह्नों के प्रयोग की महत्वपूर्ण भूमिका होती है। हिंदी में भी अनेक प्रकार के विराम चिह्नों का प्रयोग किया जाता है जिनका अपना संदर्भ होता है। साथ-ही उनके प्रयोग के कुछ नियम होते हैं। किंतु उनका ज्ञान नहीं होने या जल्दी-जल्दी टाइपिंग के क्रम में प्राय: कुछ लोगों द्वारा त्रुटियाँ हो ही जाती हैं। अत: केवल विराम-चिह्नों के प्रयोग को मानक बनाने हेतु एवं उनके प्रयोग में एकरूपता लाने हेतु सर्वप्रथम मैटर टाइप करके फिर इसमें उनका सामान्यीकरण किया जा सकता है। इस टूल का अंतरापृष्ठ इस प्रकार है-

कार्यविधि :

इस टूल/सॉफ्टवेयर में संसाधन कार्य को संपन्न कराने वाला मुख्य बटन 'सामान्यीकरण करें' है। इस बटन को क्लिक करने पर विराम-चिह्नों से संबंधित हो सकने वाली त्रुटियों का बारी-बारी से परीक्षण किया जाता है। उदाहरण के लिए कुछ विशेष प्रकार की हो सकने वाली त्रुटियाँ इस प्रकार हैं-

1. कुछ विराम-चिन्हों, जैसे- ! , ; । ? आदि से पूर्व स्पेस का प्रयोग।
2. कोष्ठक और उद्धरण चिह्न (' ' " ") के आरंभ अथवा अंत में स्पेस का प्रयोग, जैसे- (राम) या " राम "।
3. / के एक तरफ या दोनों तरफ स्पेस देना।
4. एक साथ दो विराम चिह्नों के आने पर उनके बीच स्पेस का प्रयोग, जैसे-) ; या ' ' ।
5. 'जैसे' के पूर्व अर्धविराम के अतिरिक्त किसी अन्य विराम चिह्न का प्रयोग, यथा- । जैसे – या ; जैसे-
6. एक स्थान पर एक से अधिक स्पेस का आना, जैसे – 'राम और मोहन'।

इसी प्रकार से हो सकने वाली त्रुटियों को नियमों के आधार पर यह सॉफ्टवेयर ठीक कर देता है। उदाहरण के लिए एक आउटपुट देखा जा सकता है-

इसके संदर्भ में एक बात उल्लेखनीय है कि यह 'डैश और कोलोन' (- :) के आगे पीछे के स्पेस को नहीं देखता है क्योंकि '–' का प्रयोग शीर्षक के बाद भी होता है और सामसिक शब्दों में भी। इसी प्रकार ':' का प्रयोग शीर्षक के बाद भी होता है और कुछ शब्दों के साथ भी, जैसे- अत:, सामान्यत: आदि।

9. अन्वेषक : कोशीय इकाई (कोशिम) प्राप्तकर्ता (ANVESHAK : Lexical Entry (Lexeme) Finder)

कोश वह इकाई है जिसमें किसी भाषा के सभी शब्द संग्रहीत होते हैं। इन्हीं शब्दों का प्रयोग करते हुए उस भाषा में व्यवहार किया जाता है। अर्थात् यदि किसी भाषा का वाचिक रूप या लिखित रूप हमें प्राप्त हो, तो हम उसमें प्रयुक्त शब्दों को कोश में देख सकते हैं और उनसे संबंधित सूचनाएँ प्राप्त कर सकते हैं। किंतु यहाँ पर एक समस्या देखी जा सकती है कि वाक्यात्मक व्यवहार में कोशीय शब्दों के साथ-साथ बहुत सारे शब्दों के व्याकरणिक रूपों का भी प्रयोग किया जाता है। ये रूप कोश में संग्रहीत नहीं होते। अत: इनके बारे में सूचना प्राप्त करने के लिए सबसे पहले इनके कोशीय रूप को प्राप्त करना होगा। इसके पश्चात् उसे कोश में देखकर संबंधित व्याकरणिक और आर्थी सूचना प्राप्त की जा सकती है।

व्याकरणिक रूप मुख्यत: चार शब्दवर्गों के शब्दों के बनते हैं- संज्ञा, सर्वनाम, विशेषण और क्रिया। इनमें क्रिया सबसे जटिल है। एक क्रिया के 20 से 25 रूप तक निर्मित होते हैं। अत: इन सभी का मूल रूप एक ही होगा। ऐसी स्थिति में कोशीय रूपों को समझना एक आवश्यक कार्य हो जाता है। इसे ही ध्यान में रखते हुए इस सॉफ्टवेयर को तैयार किया गया है। इसमें जब किसी पाठ का इनपुट दिया जाता है तो यह प्रणाली उसके प्रत्येक शब्द को अलग-अलग करते हुए उन शब्दों के कोशीय रूपों को उनके सामने प्रस्तुत कर देती है। इस प्रणाली का अंतरापृष्ठ इस प्रकार है-

इसमें इनपुट देकर 'विश्लेषण करें' बटन को क्लिक करने पर आउटपुट इस प्रकार प्रदर्शित किया जाता है-

आज के समय में इस प्रकार की प्रणालियों की आवश्यकता कोश निर्माण के क्षेत्र में भी है। आज किसी भी भाषा की बहुत अधिक सामग्री इंटरनेट पर प्राप्त हो जाती है। यदि सामग्री में आए शब्दों के आधार पर कोई कोश निर्मित करना हो तो इसके लिए इस सॉफ्टवेयर का प्रयोग किया जा सकता है, क्योंकि यह सॉफ्टवेयर कोशीय रूपों को प्रस्तुत कर देगा और इसके पश्चात् 'शब्द आवृत्ति गणक' में डालकर प्रत्येक शब्द को केवल एक बार प्राप्त किया जा सकेगा। या पहले 'शब्द आवृत्ति गणक' के माध्यम से प्रत्येक शब्द को केवल एक बार प्राप्त कर लें और फिर उनके कोशीय रूपों को प्राप्त कर लें। इससे हेडवर्ड चयन एवं संग्रह का कार्य अत्यंत सरल हो जाएगा। इसी प्रकार यह सॉफ्टवेयर किसी कार्पस के आधार पर शब्दकोश निर्माण में भी उपयोगी होगा।

---

10. शब्दनिधि : हिंदी-अंग्रेजी द्विभाषिक शब्दकोश(द्विदिशीय) (SHABDANIDHI : Hindi-English Bilingual Dictionary (bidirectional))

शब्दनिधि एक अंग्रेजी-हिंदी द्विभाषिक कोश है। इसे अभी आरंभिक स्तर पर लगभग 25 हजार बहुप्रचलित शब्दों को आधार बनाते हुए विकसित किया गया है। बाद में इसका विस्तार एक लाख शब्दों तक करने की योजना है। यह कोश द्विदिशीय है। अर्थात अंग्रेजी शब्द डालने पर हिंदी शब्द आएगा और हिंदी शब्द डालने पर अंग्रेजी शब्द। इसके अलावा इस कोश में अंग्रेजी शब्द के एक से अधिक हिंदी अर्थ भी देखे जा सकेंगे। आगे दूसरे चरण में इस कोश में मराठी और भोजपुरी के शब्दों जोड़ते हुए इसे चतुर्भाषिक बनाने की भी योजना है। फिर यदि संभव हो सका और पर्याप्त सहयोगी मिल सके तो इसमें संस्कृत और अन्य भाषाओं के शब्दों को भी सम्मिलित किया जाएगा।

प्रस्तुत प्रणाली के निर्मित अंतरापृष्ठ में इनपुट शब्द के बॉक्स में कोई शब्द देकर 'देखें' बटन क्लिक करें

-

इसी में हिंदी शब्द का भी इनपुट दिया जाता है। अंग्रेजी शब्द का इनपुट देने पर हिंदी शब्द और हिंदी शब्द का इनपुट देने पर अंग्रेजी शब्द आता है -

आगे दूसरे चरण में इस कोश में मराठी और भोजपुरी के शब्दों जोड़ते हुए इसे चतुर्भाषिक बनाने की भी योजना है। फिर यदि संभव हो सका और पर्याप्त सहयोगी मिल सके तो इसमें संस्कृत और अन्य भाषाओं के शब्दों को भी सम्मिलित किया जाएगा।