The Planning and Production of A Hindi Digital Dictionary for NLP Specific Purpose

Gatha Sharma, Anaadi Sharma

Shiv Nadar University, India

Corresponding author: Gatha Sharma, Email: gatha.sharma@snu.edu.in

A digital humanity provides possibilities for creation of new pathways for research in the field of Humanities and Social Sciences. A complete digital infrastructure has to be created to realise these possibilities. Since 1960's computational linguists are trying to develop tools for better communication between human languages and machine. Digital dictionary is one such tool. It forms the core of natural language processing (NLP) based tasks. Digital dictionaries contain an enormous amount of information and are very useful, but their design is structured according to specific needs either of the language learners or the program for which they had been designed, so data-access for other NLP tasks is either limited or denied completely. The creation of a complete new digital dictionary becomes inevitable in such a scenario.

Keywords: Digital Dictionary, NLP, Digital Infrastructure, Text Mining Software.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 185–194. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-19

1 Introduction

"The dictionary is like a time capsule of all of human thinking ever since words began to be written down." Andrew Clements[1] Dictionary is a cultural universal- a unique text that contains not only linguistic information but also the social, political, economic and religious history of a culture. Dictionaries, as reference works, have always been very important to researchers and scholars. Dictionaries and encyclopedias are classic examples of how "we abstract information from primary sources of various kinds and marshal that information in terms of some kind of indexing and pointing system" (McArthur, 1986, p.11). Recently, with the digitalization of knowledge and information, a new type of dictionary, digital dictionary, has come into existence. "Just as humans have consulted dictionaries not only to locate information and references, but to reassemble information and hypothesize about it too similarly do computers and computational systems" (Moulin et al. 2014, pp.56-57). Digital dictionaries provide information and reference; at the same time, they also form the core of digital tools such as annotation and data visualization websites, text mining software, translation and pronunciation websites, etc. These digital tools help in enhancing the quality of academic research. Hence, Digital dictionaries are now part of research and development infrastructure. McGann said: As with the renaissance sped forward by the printing revolution of the fifteenth century, digital technology is driving a radical shift in humanities scholarship and education. The depth and character of the change can be measured by one simple but profound fact: the entirety of our cultural inheritance will have to be reorganized and re-edited within a digital horizon (as cited in Moulin et al. 2014, p.49)

A digital dictionary is needed for all the natural language processing (NLP) based tasks. Once a digital dictionary- a comprehensive annotated lexical data set with an apt algorithm- gets created for a language, it becomes the core of the digital infrastructure for that particular language. It helps in the creation of sophisticated digital tools, apps and websites for that language.

The production of Digital dictionaries is a new field. Gouws (2018) confirmed that "the last few years witnessed an extremely positive increase in the publication of theoretical contributions within the field of internet lexicography, e.g., De Schryver (2003), Engelberg&Lemnitzer (2009), Fuertes-Olivera&Bergenholtz (2011), Granger &Paquot (2012), Müller-Spitzer (2014), Fuertes-Olivera& Tarp (2014), a major component of the volume Gouws et al. (2013), Hildenbrandt&Klosa (2016) and Klosa& Müller-Spitzer (2016)" (p.217); still presently there is no exhaustive and updated documentation of specific methods used for the production of digital dictionaries.

Herbert Ernest Wiegand (1987), the best known theoretician in the field of lexicography, was of the opinion that dictionaries are produced for the purpose of being used. Therefore, the actions of a potential user of a dictionary are in the center of the considerations (as cited in Schierholz, 2015, p.325). This stands true for digital dictionary as well. Digital dictionaries have evolved slowly over the years- starting with transfer of dictionary data from print to digital, then linking of thesaurus with the dictionary leading to induction of visual thesaurus, linguistic games, etc. to cater to the different needs of learners, researchers and scholars. The digital dictionaries have indeed covered a long distance.

Digital dictionaries are project specific e.g. "a new project to create a dictionary; a dictionary derived from one or more existing print dictionaries; a translation of another dictionary; a revision of an existing digital edition; a retro-digitalization; a software specific dictionary, etc." (Schierholz, 2015, p.328). I use computer corpus linguistics and computational linguistics to analyse the data (i.e. the transcriptions of audio recordings) collected for my research projects. To derive a variety of information from my collected data, I collaborated with a software programmer to create a text mining software for Hindi2 language as my data was in Hindi language. We realized that we need an NLP specific digital dictionary of Hindi if we want to create a good parsing scheme for our Hindi text mining

software. To extract information from the digitized data, one has to first add interpretative information to that data. This is called corpus annotation or parsing scheme. To create an apt parsing scheme, a comprehensive digital dictionary is a must.

Engelberg& Müller-Spitzer have found three different types of dictionary portals on internet- "a dictionary net, a dictionary search engine and a dictionary collection" (as cited in Gouws, 2018, p.219). Hindi language has all the three lexicographical information systems, but these systems don't have universal usability. The design rationale of Hindi dictionaries is focussed on 'learners', and rightly so, but "knowledge is multi-dimensional, colorful, and allows for several views; thus a dictionary, being a specific compilation of knowledge, should be able to demonstrate its content from different points of view" (Sinitsa et al, 1999, p.23). These Hindi digital dictionaries are not of much help for the researchers who are involved in various projects related to natural language processing (NLP). The structure of these portals is such that comprehensive word-lists cannot be extracted from them. Researchers have long discussed the architectural need within the dictionary portals for "conversion of the information into lexical databases" (as cited in Patrick, J. et al., 2000, p.294).

The varying use of standards, technologies and publishing strategies (e.g. server side transformation to HTML with underlying XML not being made public) used by these works mean that discovery, search, reuse and systematic research of individual digital dictionaries is often very difficult, and certainly impossible at the aggregate level. (Moulin et al., 2014, p.59)

For my Hindi text mining software, the comprehensive word-lists or lexical data set was needed. Since I was not able to extract lexical data set from Hindi dictionary portals, I started working on the planning and production of a digital dictionary to aid my Hindi text mining software.

The creation of a digital dictionary is a tedious and long-drawn process. It is a new field within lexicography and lacks exhaustive updated documentation of specific methods used for the creation of digital dictionaries. The process of creating a digital dictionary from scratch is interdisciplinary- linguists, lexicographers and software programmers collaborate to create the final product.

Hindi language is fortunate to have numerous researchers working on different projects to create digital infrastructure for the language. The downside is that most of them are working in their individual capacities and there is no synchronization between them. TEI3 initiative is a consortium that has revolutionized the NLP based tasks for European languages. Taking cue from their joint initiative, we decided to create a digital dictionary for NLP based tasks for Hindi language that can be easily extracted and used by other Hindi researchers. After two years of hard-work, we have successfully created an approximately 70000 words long, digital dictionary titled 'Akshara' for Hindi language users and researchers. This paper attempts to explain the multi-dimensional process of the actual creation of a digital dictionary from the scratch. It presents a detailed account of the planning and production of Akshara,, a Hindi digital dictionary created for NLP specific purposes.

2 Compilation of *Akshara*- the Hindi Digital Dictionary

A quick literary survey of lexicography field shows that practical lexicographers always developed the model of a dictionary independently without any theoretical underpinning, while metalexicographers too often developed their lexicographic theory based on the past models of dictionaries. Same holds true today in the context of digital dictionaries. The only difference is that today metalexicographical theories can be utilised aptly by the researchers working on a digital dictionary. "Developments in digital research are the natural trajectory of innovations and enquiries begun in print" (Santella, 2016, p.220).

Lexicographic processes are complex. Wiegand (1983) classified lexicographic activities into three fields of activity- "dictionary plan, dictionary base and the writing of the dictionary" (p.14). We also planned our digital dictionary in the same manner.

2.1 Dictionary Plan

The creation of a dictionary is a huge project, so a 'needs analysis' is a must before the start of the actual work to determine- workflow; an approximation of total time that will be invested in the formulation of the dictionary; size of the dictionary; and micro and macro structures of the dictionary. During the brainstorming session, we decided to include all the words included in the latest edition of *Brihat Hindi Shabdkosh*[4], a Hindi dictionary in print. This dictionary is considered among the best *Hindi* dictionaries in print version. Also, it updates its lexical data every year. This also gave us a fair idea about the size of our dictionary. We also concluded that the typing or copy pasting of words with *Hindi* OCR *Aksharyan*[5]will take about one or one and a half year to complete. Of-course our assumption was wrong. We took more than two years to complete the entire work. The micro and macro structures of a dictionary depend on three integral components of the dictionary- functions, contents and structures. Usually a general print dictionary has the following macro-structure:

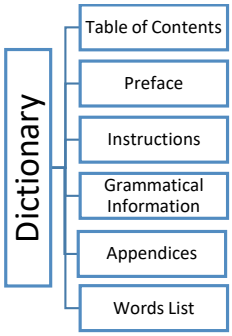

Figure 1: Macro-structure of a general dictionary

From the perspective of these three components, this *Hindi* digital dictionary is very different from other *Hindi* digital dictionaries. It is mono-lingual; doesn't include transcriptions, meanings or definitions of lemma-signs; and is created to serve only one function—to provide comprehensive digital lexical data set to scholars and researchers of *Hindi* language and literature. Following is its almost linear structure:

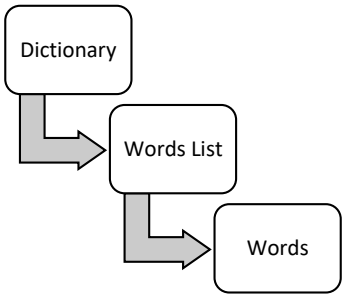

Figure 2: Macro-structure of *Akshara*, the *Hindi* digital dictionary

Weigand (1983) said, "Lexicographical activity has recourse to the results, methods and theories of various academic disciplines according to the type of reference work being produced" (p.14). This *Hindi* digital dictionary is formulated solely for NLP tasks that's why we used Corpus linguistic methods; philological method of introspection; and personal language competence for compiling lexical database for this *Hindi* digital dictionary.

The design of this digital dictionary is formulated by following the best practices of *Hindi* grammariansand the theoretical framework provided by Herbert Ernest Wiegand in his seminal research paper- 'On the structure and contents of a general theory of lexicography'(1983). The dictionary articles are organised alphabetically. This alphabetization scheme follows the standard arrangement of alphabets in Hindi *varnmala*[6] and*barahkhari matras*[7] . Following table is a small sample of the alphabetization scheme. It shows first ten words of first five alphabets of *Hindi Varnmala*.

Table 1: Alphabetization scheme of the digital dictionary

Alphabet /a/	Alphabet /aa/	Alphabet /i/	Alphabet /I:/	Alphabet /u/
अंक	आ	इँडुआ	ईंगुर	उँगनी
अंकगणित	आँ	इंक	ईंट	उँगलाना
अंकविधा	आँक	इंक-पैड	ईंत	उँगली
अंकसंबंधी	आँकड़ा	इंगला	ईंधन	उँघाई
अंकित	आँकड़ेबाज़	इंगित	ईकार	उँदरी
अंकुरित	आँकड़ेबाज़ी	इंग्लिश	ईकारांत	उँदरू
अंकुश	आँकना	इंग्लैंड	ईक्षक	उँह
अंखफोड़वा	आँख	इंच	ईक्षण	उँहूँ
अंखिया	आँखमिचौनी	इंची	ईक्षणिक	उंचन
अंग	आँखमिचौली	इंजन	ईक्षा	उंचना

The lexical data-set includes words, parts of words, abbreviations, and punctuation marks. Since this digital dictionary is primarily created for NLP tasks, only lexical data base is created; definitions, meanings, transcriptions, etc. are not included in the dictionary. The text mining software for which this particular digital dictionary is being created is also designed simultaneously. A document is created to record all the methods used, and procedures followed, to create this lexical database. Table 3 is an extract from the same document. This extract contains information about the typology of 'genitive pronoun' and the tags prescribed for each category.

Time and again procedures and methods are checked to test the practical suitability, and are sometimes changed to match the same.

2.2 Dictionary Base

The lexicographical corpus for this Hindi digital dictionary is created through primary, secondary and tertiary sources. This digital dictionay has two objectives, a) to create tagged corpus for Hindi text mining editor and, b) to provide lexical data to Hindi scholars.

Table 2: Documentation of Genitive Pronouns and the tags prescribed to them

Genitive/possessive ('s/of)				
Person	**Singular (masculine)**	**Singular (feminine)**	**Plural (masculine)**	**Plural (feminine)**
1st	मेरा/मेरे (my/mine)	मेरी (my/mine)	हमारा/हमारे (our/ours)	हमारी (our/ours)
2nd (intimate)	तेरा/तेरे (your/yours)	तेरी (your/yours)	तुम्हारा/तुम्हारे (your/yours)	तुम्हारी (your/yours)
2nd (honorary)	आपका/आपके (your/yours)	आपकी (your/yours)		
3rd (proximal)	इसका (of this/Its/his)	इसकी (of this/Its/her)	इन/इनका/इनके (their/theirs)	इनकी (their/theirs)
3rd (distal)	उसका (his)	उसकी (her)	उन/उनका/उनके (their/theirs)	उनकी (their/theirs)

Tags-

Pronoun	Tags
Pronoun Genitive (S, M)	PG1
Pronoun Genitive (P, M)	PG2
Pronoun Genitive (S, F)	PG3
Pronoun Genitive (P, F)	PG4
Pronoun Genitive (P, Gender Neutral)	PG5

> A dictionary base includes at least the lexicographical corpus as the set of all the primary sources: primary sources may be defined as all sources not themselves language dictionaries, the secondary sources as the set of all language dictionaries consulted, and other linguistic material (Wiegand ,1983, p. 14).

The primary source for Hindi lexical data is *Brihat Hindi Shabdkosh*. This dictionary has comprehensive lexical data that gets updated every year. To produce a digital copy of the print edition, its pages are scanned, and afterwards the obtained images are converted into text by means of an OCR software *e- Aksharyan*. However, these measures do not always ensure the high quality of the obtained data, and manual keyboarding has to be performed as well.

The secondary source is one of the oldest Hindi print dictionary- *Hindi Shabd Sagar*[8]. This dictionary has eight volumes and is considered as one of the best Hindi dictionary. I have used *Hindi ShabdSagar* primarily to evaluate each dictionary article selected from *Brihat Hindi Shabdkosh* . This task is performed mainly to reinforce the evaluation process of lexical data.

The tertiary source for this dictionary is the text of Hindi grammar- *Hindi Vyakaran*[9]. The grammar book is needed to determine the grammatical categories of word types. Clearly defined grammatical

categories help in determining the relevant tag-set. The tag-set is the base of an annotation scheme and is crucial for the formulation of an apt algorithm.

2.3 Writing of the Dictionary

The dictionary articles are entered manually-words are typed using 'quillpad'[10] or copied and pasted from images obtained through *Hindi* OCR *e-Aksharyan;* annotated, and encoded. The typology of Hindi lexicon is determined through different basis- origins of words, evolution of words, usage of words and meaning of words. We decided to categorize our lexical data according to the usage of words. All the decisions pertaining to the formulation of an annotation scheme aim for the compatibility between linguistic and computational elements. As leech (2015) confirmed, "While annotators are bound to face some theoretically sensitive decisions, their goal should be to adopt annotations which are as widely accepted and understood as can be managed" (p.7). The usage based typology divides words in the eight categories- nouns, pronouns, verbs, adjectives, adverbs, prepositions, conjunctions and interjections. This neat categorization is helpful in creating a comprehensive tag-set and parsing scheme. These eight categories form the core of the Tag-set. We have also added one more peripheral category to our lexical data- 'sentence endings'. In Hindi, 'sentence endings' stand on their own and have their own usage. So in total our tag-set has nine categories.

Our annotation scheme is inspired by SGML mark-up system created by TEI consortium. It conforms to the internationally accepted guidelines of Text Encoding Initiative (TEI). During the brainstorming session, we decided to keep our tags brief, intelligible and free of ambiguity. We decided each tag to have two Roman alphabets and one digit. The logic is that the first Roman alphabet connotes the word-category, the second Roman alphabet connotes sub-divisions within that particular word-category and the digit represents linguistic inflections. This tag design also helps in identifying homonyms; and maintaining structural consistency of the lexical data-set. Table 3 contains tag-sets for 'direct Pronouns'.

Table 3: Direct pronouns and their tag-set

Pronoun	**Tag-set**
Direct case (Singular)	PD1
Direct case (Plural)	PD2
Dative case (Singular)	PV1
Dative case (Plural)	PV2
Genitive (Singular, Masculine)	PG1
Genitive (Plural , Masculine)	PG2
Genitive (Singular, Feminine)	PG3
Genitive (Plural, Feminine)	PG4

Extra care and precaution are taken to evaluate each tag. The word categories, sub-categories and inflections are neatly classified through different tags. Table 4 shows lexical verb, its various inflections and their tags.

Table 4: Lexical verbs and their tag-set

Words	**Tag**
आ	LX0
आना	LX1
आओ	LX2

आइये	LX3
आएं	LX4
आया	LX5
आता	LX6
आते	LX7
आती	LX8
आऊंगा	LF1
आएँगे	LF2
आएगा	LF3
आएगी	LF4
आएंगी	LF5
आओगे	LF6
आओगी	LF7

3 Digital Location of *Akshara*

Lexicographic work has always been interdisciplinary and scientific but in the context of digital dictionary, information science plays a huge role as digital dictionary will "be situated within the broad spectrum of online information tools" (qtd.in Gouws, 2018. P.216). This digital dictionary is the result of individual efforts and fruitful collaboration with software programmer to create an infrastructural core for *Hindi* scholars. The dictionary is part of internal repository of Hindi text mining software *Shabdshilpi*,and is available with the text mining software as Txt. Files. These files can be easily downloaded and used by *Hindi* researchers.

The 'optimal retrieval of information' is one of the main objectives of this digital dictionary. The design of the data distribution structure and dictionary portal structure are in sync with each other to provide users better access to the data. A text containing guidelines for users has also been created for further convenience. The users can also provide their feedback to improve the services of this dictionary further.

4 Conclusion

Hindi is, digitally, a low-resource language although it is spoken by 52,83,47,193 speakers (Census, 2011). The linguistic space is digitizing at a fast pace world over. A sturdy digital infrastructure is needed not only for the development of NLP based digital tools but also for the survival of concerned languages in the future. The future belongs to the languages that can imbibe machine learning and deep learning to make their languages work on digital tools

Future research in the field of Humanities and Social Science will remain closely entwined with digital tools. The quality of digital infrastructure will determine the quality of research output. Production of a *Hindi* text mining software along with *Hindi* digital dictionary is first step towards digital analysis of *Hindi* texts. India has amazing linguistic diversity. The creation of digital infrastructure for Indian languages is the need of hour. It will strengthen not only the quality scholarship in Indian languages but will also nurture the languages for future generations.

References

[1] Gouws, Rufus H. (2014). Article structures: Moving from printed to e-dictionaries. Lexikos,24:155-177.

[2] Gouws, Rufus H.(2018).Internet lexicography in the 21st century. In: Stephen Engelberg. (eds.)Wortschatz: Theorie, empirie, documentation, 1: 215–236. De Gruyter, Berlin/Boston.

[3] Leech, Geoffery.(2013). Introducing corpus annotation. In: Roger Garside, Geoffery Leech & Tony McEnery (eds.)*Corpus annotation*, 1-18. Routledge, London/Newyork.

[4] Lew, R. (2013). From paper to electronic dictionaries: Evolving dictionary skills. In:D.A.Kwary, N. Wulan and L. Musyahda (eds.) *Lexicography and dictionaries in the information age*, 79-84. Selected papers from the 8th ASIALEX international conference. Airlangga University Press, Surabaya.

[5] McArthur, Tom. (1986)*Worlds of reference : Lexicography, learning, and language from the clay tablet to the computer*. Cambridge University Press, Cambridge [Cambridgeshire]; New York.

[6]O'riordan, T. (2019)The Irish world: How to revise a long-standing dictionary project. In: Karen Fox (eds.) *True biographies of nations?: The cultural journeys of dictionaries of national biography*pp. 37-56. ANU Press, Australia.

[7] Santella, Anna-Lise P. (2016).The ideal dictionary: Impossible tasks, frank adjustments, and Lexicoraphical Innovations in the Creation Of "Grove Music Online". *FontesArtisMusicae*, 63(3): 213–221.

[8] Schierholz, J.S. (2015). Methods in lexicography and dictionary research. Lexikos,25: 323-352.

[9] Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. Historical Social Research / HistorischeSozialforschung, 38 (4): 332-357.

[10]Wiegand, H.E. (1983).On the structure and contents of a general theory of lexicography. In: A. Sture, P. Corbin, R.K.H. Reinhard, F.J.Hausmann, Hans-PederKromann, Oskar Reichmann and LadislavZgusta (eds.) Proceedings of the 1st EURALEX international Congress,ELX83-007, 13-30. Exeter,United Kingdom, Sep. 9-12,1983. Exeter: Max Niemeyer VerlagTübingen, Exeter.

[11]Zimmer,Ben. (2014).Lexicography 2.0: Reimagining dictionaries for the digital age. Dictionaries: Journal of the dictionary society of North America, 35: 275-286.

[12] Arkhangelskiy, T., Usacheva, M., &Serdobolskaya, N. (2017). Corpus-oriented lexicographic database for Beserman Udmurt. Actalinguisticaacademica, 64(3): 397–415 .

[13]Moulin,C. and Nyhan, J. (2014). The dynamics of digital publications: an exploration of digital lexicography. In: Peter Dávidházi (eds.)*New publication cultures in the humanities: Exploring the paradigm shift*, 47–62. Amsterdam University Press, Amsterdam.

[14]Patrick, Jon., Zhang Jun, &Artola-Zubillaga, Xabier. (2000). An architecture and query language for a federation of heterogeneous dictionary databases. Computers and the humanities, 34(4): 393–407.

[15] Sinitsa, K. &Manako, A. (1999). Interactive dictionary as an information wish-maker. Educational Technology, 39(5): 22–25.

[16]Census 2011 https://censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf, last accessed on 21/10/17.

Endnotes

1. Andrew Clements https://www.brainyquote.com/quotes/andrew_clements_711948
2. *Hindi* is one of the main languages spoken in India. It belongs to 'Indo-European' language family and is written in 'Devnagari' script. According to 2011 Indian Census, *Hindi* is the mother tongue of approximately 44% people in India. It is also one of the official languages of India.
3. TEI Consortium is an international organization that formulates and maintains guiding principles for digital encoding of linguistic and literary texts.

4. *Brihat Hindi Shabkosh* is a *Hindi* dictionary, compiled by Kalika Prasad, Rajvallabh Sahay and Mukundi Lal Shrivastava, published by Jnana Mandal Ltd.,Varanasi, India (2016 edition).
5. e-aksharyan is a software created specifically for Indian languages by the Indian Ministry of Electronics and Information Technology (MeitY). This software converts any printed or scanned document into fully editable text. Currently it covers twelve Indian languages; *Hindi* is one of them.
6. Varnmala is *Hindi* alphabet. *Hindi* has eleven vowels and thirty three consonants.
7. Barahkharimatras are twelve spellings with which *Hindi* words are formed.
8. *Hindi ShabdSagar* is a *Hindi* dictionary compiled by Shyam Sunder Das and published by Kashi-Nagri-Pracharini Sabha (1925)
9. *Hindi Vyakaran* is a book on *Hindi* grammar, written by Kamta Prasad Guru and published by PrabhatPrakashan, New Delhi, India (2018)
10. Quillpad is an online typing tool for ten Indian languages including Hindi.