Sense Tagged CLE Urdu Digest Corpus

Saba Urooj, Sana Shams,
Sarmad Hussain, Farah Adeeba

Centre for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology Lahore,
Pakistan

Contents

- What is WSD & Sense Tagged Corpus?
- Why It is useful?
- An overview of existing Wordnet Tagged Corpora
- The process of Developing Urdu Sense Tagged Corpus
- Current State of Urdu Sense Tagged Corpus
- Future Work

What is WSD?

- "To define the correct meaning of a word in its respective context" [1]
- A word's meaning comes from patterns of use
- To see this usage pattern, corpora needs to be sense tagged

WSD

بول چال میں زبان کی صحت اور صفائی کا خیال رکھنا چاہیئے۔
AUXM VBI NN PSP NN CC NN PSP NN PSP NN NN
لحاظ
توجہ
بیگانہ
حرف عطف
درستی
ڈھنگ
دھوکا
چلنے کا عمل
شعر
جملہ
بولنے کی طاقت
صیغہ نکاح
طعنہ
بولنے کا انداز
بولی
بولنے کا عضو
شعلہ
جھاڑ پونچھ
خالی ہونا
ہمواری
عمدگی
یاد
فکر
شاعری کی صنف
رائے
آرام
تندرستی
رواج

Why is Sense Tagged Corpus Useful?

- Natural Language Application
  - traditional and statistical machine translation
  - computer-assisted lexicography [2]
- Statistic data extracted from sense tagged corpus can be implemented in
  - Information Retrieval (IR)
  - Information Extraction
  - Text Summarization [3]

An Overview of Wordnet Tagged Corpora

- Hector (20 Million words) [4]
- English SemCor (200, 000 words) [5]
- DSO Corpus (192,800 words) [4]
- Dutch SemCor (400,000 words) [6]
- Bulgarian Word sense tagged corpus (99,480 words) [7]
- Chinese Word sense tagged corpus (60,895 words) [8]

DSO Corpus

- 192,800 words tagged
- Tagged with WordNet 1.5 senses
- It is distributed on the Linguistic Data Consortium (LDC)
- The DSO corpus uses the targeted tagging approach

Need

- For the development of a successful and accurate WSD
- Attention has been paid to develop
  - Urdu Digest Corpus [9]
  - Urdu Wordnet [10]
- Sense annotation is unexplored

Developing Urdu Sense Tagged Corpus

- Four linguistic resources have been used
  - Urdu Wordnet 1.0 Wordlist[1]
  - CLE Urdu Digest Corpus [9]
  - Urdu WordNet [10]
  - Urdu Morphological Analyzer [11]

[1] http://www.cle.org.pk/software/ling_resources/UrduWordNetWordlist.htm

Urdu Wordnet 1.0 Wordlist

141	ارتقا	153	ارفع	165	اساطیر	177	استدلال
142	ارتقائی	154	اڑانا	166	اسامی	178	استر
143	ارتکاز	155	اڑتالیس	167	اسباب	179	استری
144	ارجمند	156	اڑتیس	168	اسپرٹ	180	استصواب
145	اردو	157	اڑنا	169	اسپیکر	181	استطاعت
146	ارز	158	ازبر	170	استاد	182	استعداد

CLE Urdu Digest Corpus

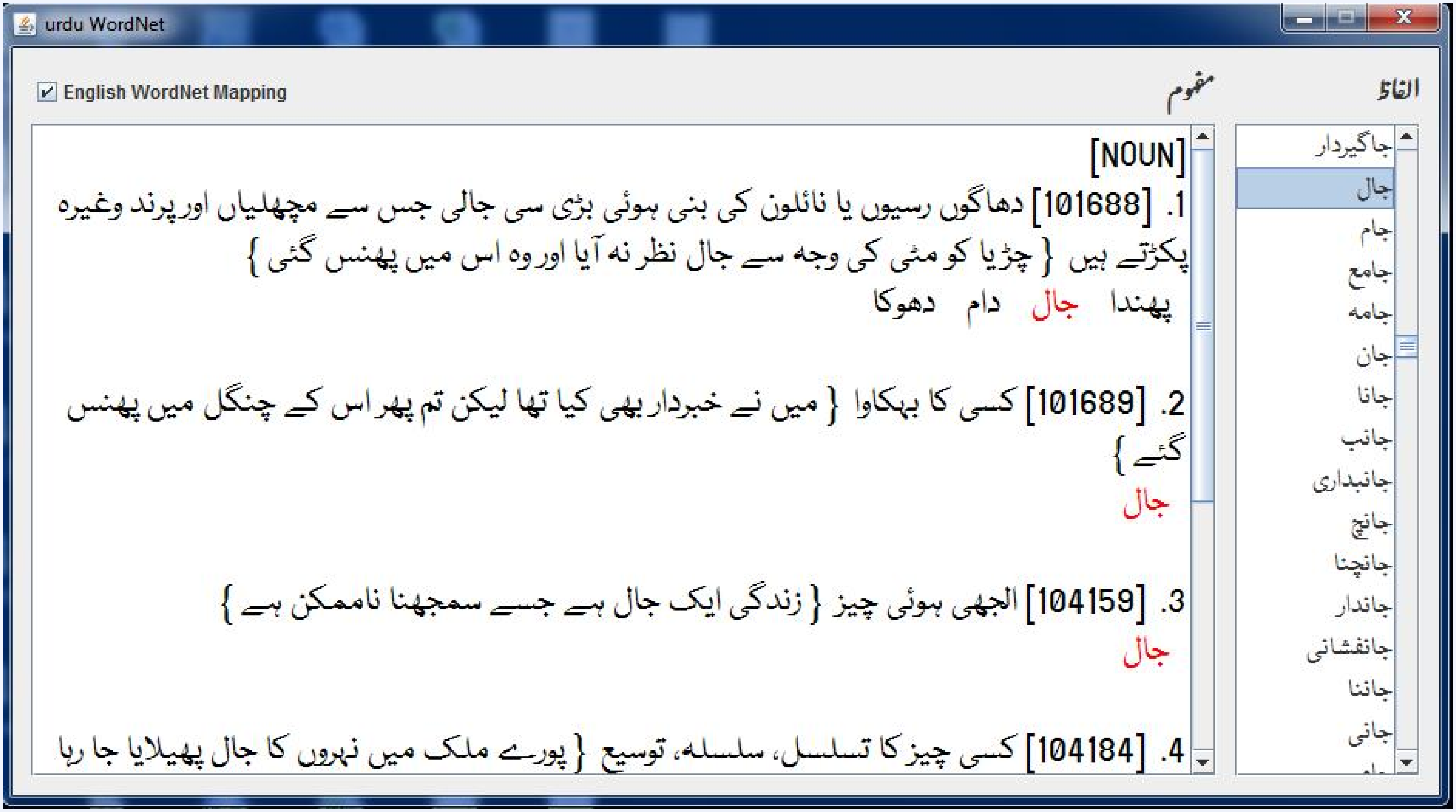
دنیا کا ہر فرد کامیابی کا آرزومند ہے۔ ناکامی سے سب گھبراتے ہیں۔ عزت، دولت، راحت اور عافیت کی زندگی کے بھی شیدائی ہیں۔ لیکن اصل کامیابی کیا چیز ہے؟ اور حقیقی عزت و راحت کس طرح نصیب ہوتی ہے؟ اس بھید سے بہت کم لوگ واقف ہیں۔ اگر آپ حقیقی کامیابی کے گر جاننا چاہتے ہیں تو ڈاکٹر زاہد منیر عامر کی تازہ تصنیف آئینہ کردار پڑھیے۔ ۱۱۲ صفحوں کی اس کتاب کا ایک ایک حرف بصیرت کے دریچے کھولنے پر مامور ہے۔

راقم نے اس کتاب کا مطالعہ کیا تو لفظ و معنی کی کہکشاں دیکھ کر مسحور ہو گیا جس چیز نے خاص طور پر متاثر کیا وہ ڈاکٹر صاحب کا فہم قرآن ہے۔ بظاہر ڈاکٹر صاحب پنجاب یونیورسٹی کے معلم ادبیات ہیں لیکن درحقیقت وہ ایک داعی، ایک عارف، ایک محقق، ایک مدبر، ایک مقرر، ایک آموزگار اخلاق اور قلم و قرطاس کے فرمانروا ہیں۔ ٹی وی پر ان کی تقریریں بڑے ذوق و شوق سے سنی جاتی ہیں۔ ان کی باتیں عید کی سویاں ہیں۔ بے اختیار دل میں اترتی چلی جاتی ہیں۔ سب سے زیادہ گرانمایہ خوبی یہ ہے کہ ان کی گفتگو قرآن کریم کی برمحل آیات اور ارشادات رسالت مآبؐ سے یوں جگمگاتی ہے جیسے

ع پرتوِ آفتاب کے ذرے میں جان ہے

Sense Tagged CLE Urdu Digest Corpus

Urdu WordNet



Urdu Morphological Analyzer

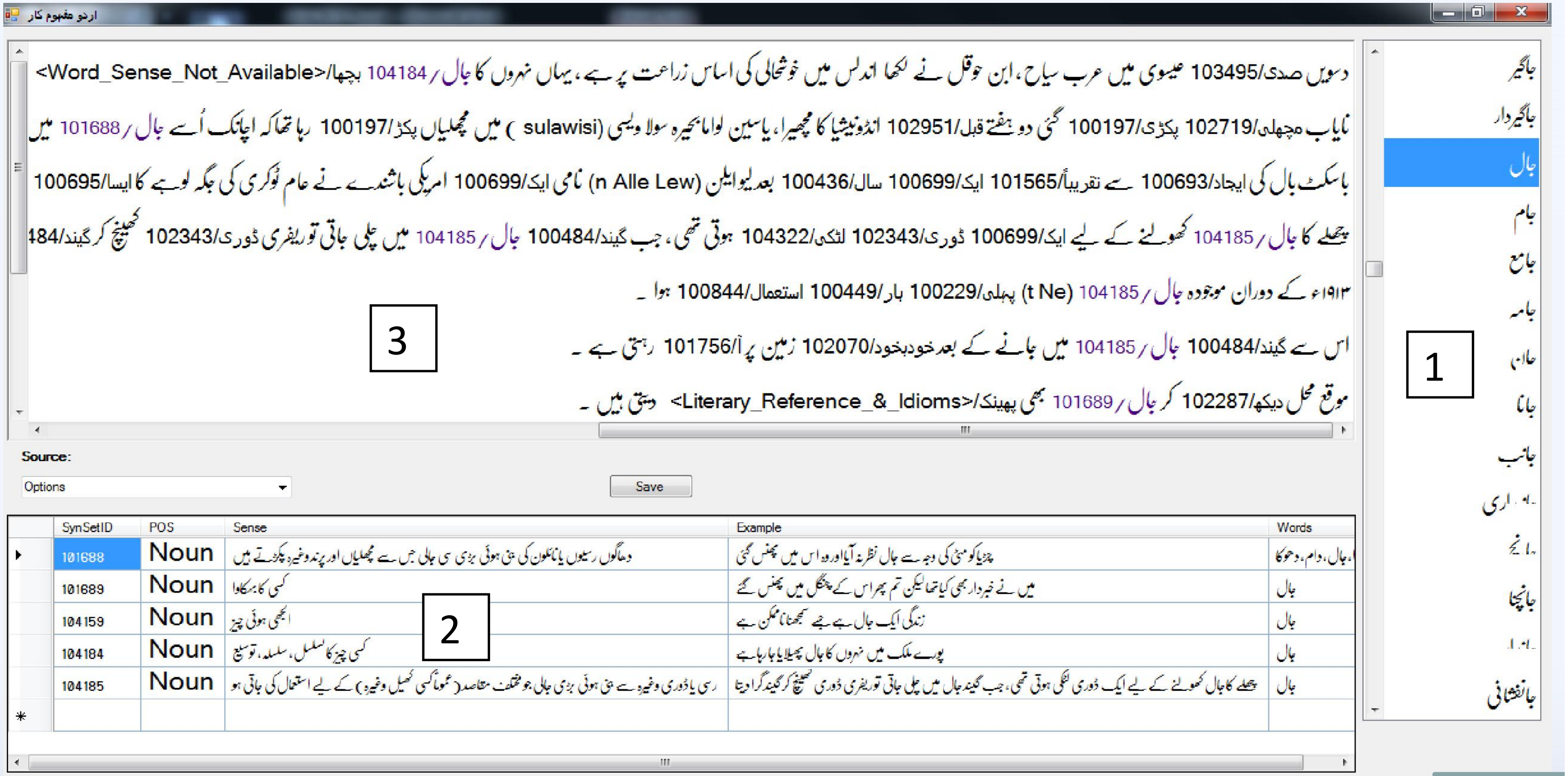
	A	B	C	D	E
1	ابا		ابالنا		اتر
2	enumerator>> string not found		enumerator>> ابلنا		enumerator>> اترنا
3	ابال		enumerator>> ابلنے		enumerator>> اترنے
4	enumerator>> ابلنا		enumerator>> ابلنی		enumerator>> اترنی
5	enumerator>> ابلنے		enumerator>> ابلنی		enumerator>> اترنی
6	enumerator>> ابلنی		enumerator>> ابل		enumerator>> اتر
7	enumerator>> ابلنی		enumerator>> ابلتا		enumerator>> اترتا
8	enumerator>> ابل		enumerator>> ابلتے		enumerator>> اترتے
9	enumerator>> ابلتا		enumerator>> ابلتی		enumerator>> اترتی
10	enumerator>> ابلتے		enumerator>> ابلتیں		enumerator>> اترتیں
11	enumerator>> ابلتی		enumerator>> ابل		enumerator>> اتر
12	enumerator>> ابلتیں		enumerator>> ابلا		enumerator>> اترا
13	enumerator>> ابل		enumerator>> ابلے		enumerator>> اترے
14	enumerator>> ابلا		enumerator>> ابلی		enumerator>> اتری
15	enumerator>> ابلے		enumerator>> ابلیں		enumerator>> اتریں
16	enumerator>> ابلی		enumerator>> ابلے		enumerator>> اترے

Sense Annotation Method

- Word forms were linked to corresponding word senses in Urdu Wordnet
- Targeted tagging approach was used
- Benefit of using this approach

Sense Tagged CLE Urdu Digest Corpus

Urdu Word-Sense Annotation tool (اردو مفہوم کار)



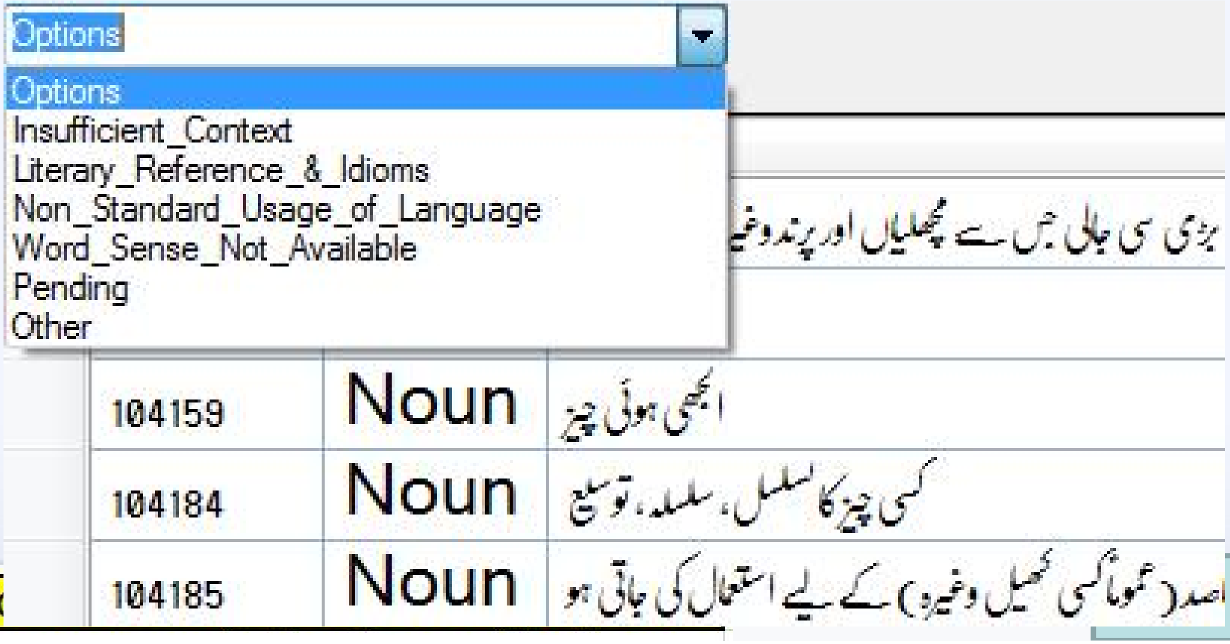
CLT14 | Center for Language Engineering (CLE), 2014

Three views of the interface

1. Selection view: displays the list of high frequency words and enables the annotator to select a target word
2. Wordnet view: displays the linguistic information available in WordNet for the selected lexical item
3. Corpus view: displays the corpus with all occurrences of the target word

Corpus annotation tags

- Exact Match
- Insufficient Context
- Literary Reference/symbolic Sense
- Non-standard Usage of language
- Word Sense Not Available
- Other



Current state of Urdu sense tagged corpus

Sense Tagged Corpus	
Total no. Of sentences in the corpus	5611
Total no. of words in the corpus	100,000
Tagged total word types	2225
Tagged total sense types	2285
Tagged total word tokens	17006

	No. of senses tagged					
Words	1	2	3	4	5	6
	1522	345	118	49	21	14
	No. of senses tagged					
Words	7	8	9	10	11	12
	3	2	3	3	1	1

Improving Wordnet via sense tagging

- Consistency of sense definition with POS
  - Interpretive definition doesn't match with the POS

Words	POS	Sense	Modified
افسردہ	ADJ	تھکن یا غم اور دکھ ہونے کی وجہ سے بیزار ہوجانا	تھکن یا غم اور دکھ ہونے کی وجہ سے بیزار ہونے والا، بجھا بجھا سا

Consistency of sense example with POS

- Example associated with the synset doesn’t match with POS of the word

Words	POS	Sense	Example	Modified
خزندہ	Noun	رینگنے والاجانور	وہاں خزندہ جانوروں کی بھرمار تھی	وہاں خزندوں کی بھرمار تھی

Consistency of definition across synset

- It is very important that all the members of the synset have equal relationship with sense meaning

POS	Sense	Example	Words	Modified
Verb	پرت والی چیز کا کوئی حصہ الگ کرنا، ٹکڑے کرنا	اس نے تختہ توڑ کے کشتی کو پھاڑ دیا	پھاڑنا، چیرنا ادھیڑنا	پھاڑنا،چیرنا

Addition of senses available in the corpus

- During the process of annotation all the “word sense not available” tags were reported back to the WordNet team

Challenges in the process of sense tagging

- Tool specific
  - Normalization
- Language specific
  - Non-standardized translations
  - Foreign language borrowed words
  - Complex predicates

Tool specific

- Normalization
  - annotation tool was unable to match corpus in some contexts e.g. vow (و) with hamza above case (ؤ)
  - these combinations were typed in different formats in corpus and WordNet and hence requiring the process of normalization
  - the process of representing texts into consistent formats [23]

Language specific

- Non-standardized translations
  - It was ambiguous to tag non-standardized translations of English words which have become part of Urdu language
  - e.g. (بلند فشارِ خون) i.e. high blood pressure and
- Foreign language borrowed words
  - Mapping was not found for such words which are borrowed from foreign language and have been lexicalized for Urdu
  - e.g. test match, basket ball and interview

Complex Predicates

- Complex predicates
- Main Verb + Light Verb in Urdu [12]
  - اس میں پائےجانے والی حیاتین ہماری آنکھوں کو صحتمند رکھتی ہے ۔

Future Work

- The current Sense tagged CLE Urdu Digest Corpus has
  - 100k Words
    - Of which 17006 are sense tagged
- This manually data can act as seed data
- Development of semantically rich corpus

Acknowledgements

- This work has been supported through the collaboration of
  - University of Konstanz, Germany
  - German Research Exchange Program (DAAD)

Thank you

References

1. A. Kilgarriff, “Word senses”, *Word Sense Disambiguation*, Springer Netherlands, 2006.
2. P. Resnik, “WSD in NLP applications”*, Word Sense Disambiguation: Algorithms and Applications (2006).*
3. S. J. Ker, C. R. Huang, J. F. Hong, S. Y. Liu, H. L. Jian, I. L. Su & S. K. Hsieh, "Design and Prototype of a Large-scale and Fully Sense-tagged Corpus." Large-Scale Knowledge Resources. Construction and Application. *Springer* Berlin Heidelberg, 2008
4. A. Kilgarriff, “SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Program”, Proceedings of First International Conference on *Language Resources and Evaluation*, Granada, 1998.

References

5. S. Landes, C. Leacock and R. I. Tengi, “Building semantic concordances”, *WordNet: An electronic lexical database*, 1998: 199-216.
6. P. Vossen, A. Görög, R. Izquierdo, & A. van den Bosch, “DutchSemCor: Targeting the ideal sense-tagged corpus”, *LREC*, 2012.
7. S. Koeva, S. Leseva, E. Tarpomanova, B. Rizov, T. Dimitrova & H. Kukova, “Bulgarian Sense-Annotated Corpus–Results and Achievements”, *FASSBL7*, 2010: 41.
8. Y. Wu, P. Jin, Y. Zhang & S. Yu, “A Chinese corpus with word sense annotation” Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead, *Springer,* Berlin Heidelberg, 2006.

References

9. S, Urooj, S, Hussain, F, Adeeba, F. Jabeen and R. Perveen, “CLE Urdu Digest Corpus”, in the Proc. of *Conference on Language and Technology 2012* (CLT12), Lahore, Pakistan, 2012.
10. A. Zafar, A. Mahmood, F. Abdullah, S. Zahid, S. Hussain, and A. Mustafa, "Developing Urdu WordNet Using the Merge Approach ", in the Proceedings of *Conference on Language and Technology* 2012 (CLT12), Lahore, Pakistan.
11. Hussain**,** S., “*Finite-State Morphological Analyzer for Urdu*”, National University of Computer and Emerging Sciences, (2004), Lahore, Pakistan.
12. M. Butt, “The light verb jungle: Still Hacking Away”, in Workshop on Multi-Verb Constructions, 2003.