

Lexical Resource

S. M. Murtoza Habib
CRBLP, BRAC University

Center for Research on

CRBLP

Bangla Language Processing



Topic

- ◆ **Wordnet**
- ◆ **Text Corpus**
- ◆ **POS Tagging**

Wordnet (শব্দজালিকা)

- ◆ Network of words
- ◆ Lexical and semantic relation between words

Wordnet (শব্দজালিকা)

Category of words:

- * noun
- * verb
- * adjective
- * adverb

Relation:

- * Hyponym and hypernymy (is a kind of)
- * Meronymy and holonymy (Part-whole relation)

Bangla wordnet:

- * Noun category
- * Hypernymy relation

Wordnet (Hypernyms and Hyponyms)

Hypernyms: (SKY is a kind of ...)

sky -- (the atmosphere and outer space as viewed from the earth)

=> atmosphere -- (the envelope of gases surrounding any celestial body)

=> gas -- (a fluid in the gaseous state having neither independent shape nor volume and being able to expand indefinitely)

=> fluid -- (a continuous amorphous substance that tends to flow and to conform to the outline of its container: a liquid or a gas)

=> substance, matter -- (that which has mass and occupies space; "an atom is the smallest indivisible unit of matter")

=> physical entity -- (an entity that has physical existence)

=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Hyponyms: (... is a kind of SKY)

sky -- (the atmosphere and outer space as viewed from the earth)

=> blue sky, blue, blue air, wild blue yonder -- (the sky as viewed during daylight; "he shot an arrow into the blue")

=> mackerel sky -- (a sky filled with rows of cirrocumulus or small altocumulus clouds)

Wordnet (Holonyms and Meronyms)

Holonyms: (SKY is a part of ...)

sky -- (the atmosphere and outer space as viewed from the earth)

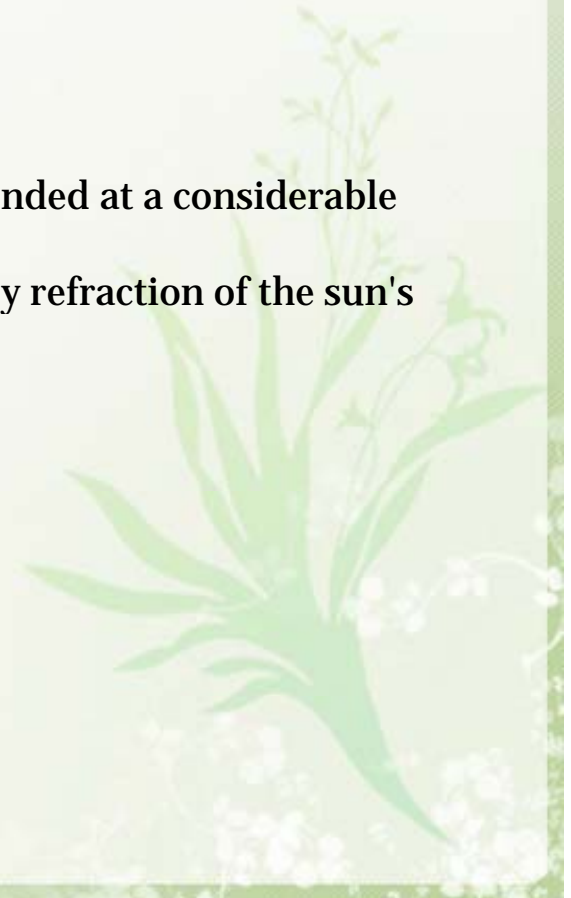
PART OF: Earth, world, globe -- (the 3rd planet from the sun; the planet we live on; "the Earth moves around the sun"; "he sailed around the world")

Meronyms: (... is a part of SKY)

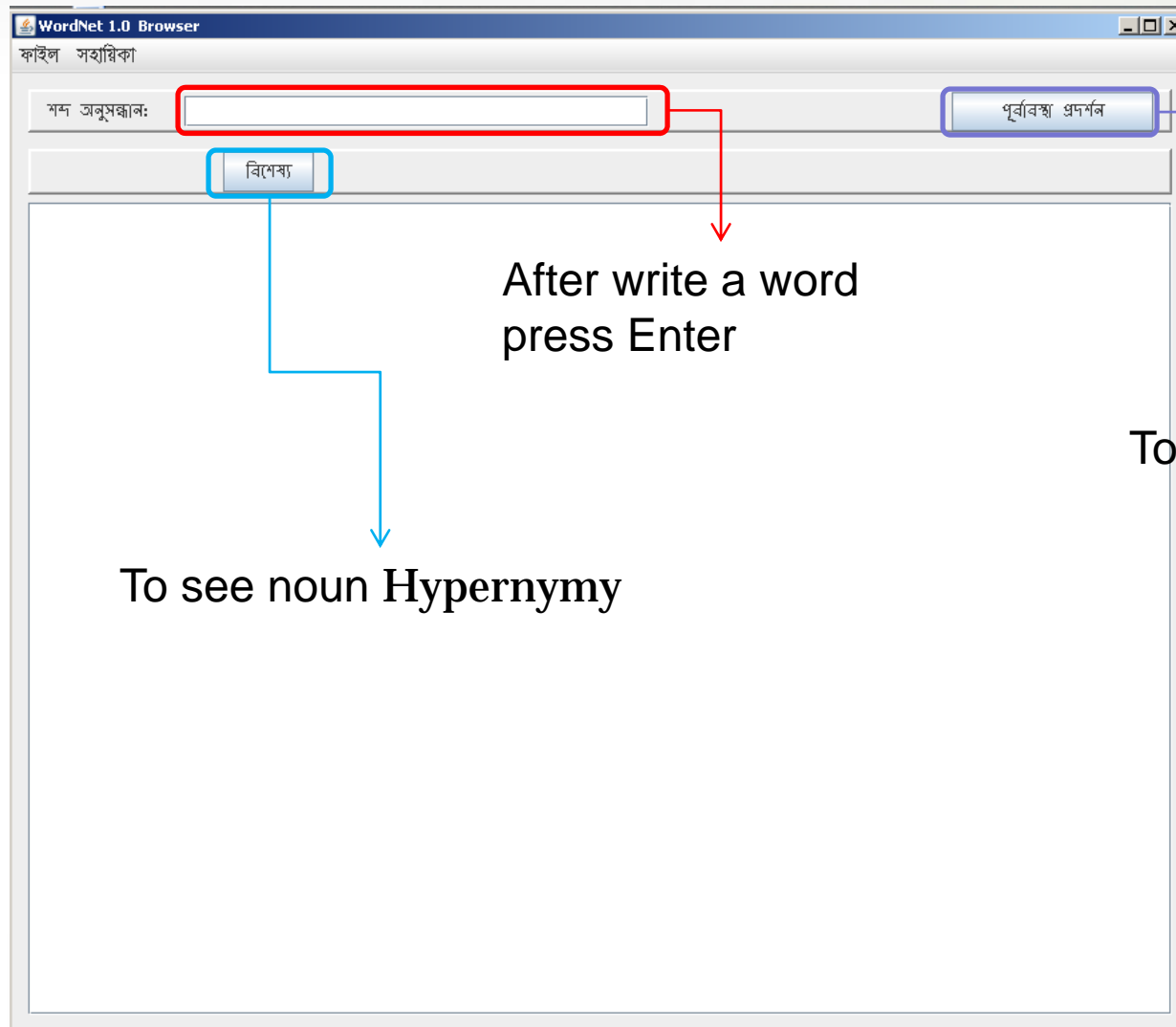
sky -- (the atmosphere and outer space as viewed from the earth)

HAS PART: cloud -- (a visible mass of water or ice particles suspended at a considerable altitude)

HAS PART: rainbow -- (an arc of colored light in the sky caused by refraction of the sun's rays by rain)



Wordnet (শব্দজালিকা)



Wordnet (শব্দজালিকা)

WordNet 1.0 Browser

ফাইল সহায়িকা

শব্দ অনুসন্ধান: পাখি

পাখি অনুসন্ধান:

There are 5 senses of পাখি

1. গগনগতি, খগ, খগম, খেচর, চিড়িয়া, নভোকা, পংখি, পখি, পখী, পক্ষধর, পক্ষালু, পক্ষী, পতঙ্গ, পতত্রি, পাক, পত্রনখ, পত্রী, পাখি, বিহগ, বিহঙ্গ, বিহঙ্গম, পাখি (পাখি) -- (পালকবৃত্ত এবং অগ্র-পদঙ্গ পাখায়ে নৃপাল্পিত হযেছে এমন উষ্ণরক্ত বিশিষ্ট ডিম প্রসবকারী মেম্ব্রুগী প্রাণী।)
2. জমির একক বিশেষ, ৩০ কানি ভূমি, ২৬/৩৩/৩৫ শতাংশ, পাখি, অঞ্চল একক -- (জমির পরিমাপ পদ্ধতির জন্য ব্যবহৃত একক।)
3. চক্র-অবলম্বক, স্পোক (spoke), পাখি (চক্র) -- (চক্রকেন্দ্র এবং চক্রের পরিধির মধ্যবর্তী অক্ষপ্রসারী সংযোজক দ্বারা সৃষ্ট অবলম্বক।)
4. পাখি (আড়াকার্ত) -- (মইয়ের ধাপ হিসাবে ব্যবহৃত আড়াকার্ত।)
5. পাখি (তন্নুগী) -- (অন্নবহসী নারী (নৃপকর্ষে)।)

Wordnet (শব্দজালিকা)

WordNet 1.0 Browser

ফাইল সহায়িকা

শব্দ অনুসন্ধান: পাখি পূর্ববস্থা প্রদর্শন

পাখি অনুসন্ধান: বিশেষ্য

5 senses of পাখি

Sense 1
গগনগতি, খগ, খগম, খেচর, চিড়িয়া, নভোকা, পংখি, পখি, পক্ষী, পক্ষধর, পক্ষালু, পক্ষী, পতঙ্গ, পতত্রি, পাক, পত্রনথ, পত্নী, পাখি, বিহগ, বিহঙ্গ, বিহঙ্গম, পা
=> মেনুদণ্ডী -- (যাদের মেনুদণ্ড- খণ্ড অস্ত্র বা তলুপাখি দিয়ে তৈরি এবং যা কলোটিতে যুক্ত থাকে।)
=> কর্ভেট -- (দেহকাঠামোতে নটোকর্ভ বা মেনুদণ্ড আছে, এমন কর্ভেটা পর্বের যে কোন প্রাণী।)
=> প্রাণী -- (জীবজগতের সদস্য, যারা স্বেচ্ছায় চলাচল করতে পারে।)
=> জীবসত্তা -- (আল্লা-উন্নয়নের সামর্থ্য আছে বা স্বাধীনভাবে কার্যক্রম সম্পন্ন করতে পারে এমন সত্তা।)
=> জীবন্ত বস্তু -- (জীবন্ত বা এক সময় জীবিত ছিল।)
=> দৈহিক লক্ষ্যবস্তু -- (যা স্পর্শ করা যায় এবং দেখা যায় এবং ছায়া প্রদান করে।)
=> ইন্দ্রিয়গ্রাহ্য সত্তা, দৈহিক সত্তা -- (দৈহিক অস্তিত্ব আছে এমন সত্তা।)
=> অস্তিত্ব, সত্তা -- (স্বতন্ত্র অস্তিত্ব আছে এমন ইন্দ্রিয়গ্রাহ্য বা জ্ঞাত বা সিদ্ধান্তকৃত এমন সত্তা কিছু (জীবিত বা জড়)।)

Sense 2
জমির একক বিশেষ, ৩০ কানি ভূমি, ২৬/৩৩/৩৫ শতাংশ, পাখি, অঞ্চল একক -- (জমির পরিমাপ পদ্ধতির জন্য ব্যবহৃত একক।)
=> পরিমাপ একক (নির্মাণ) -- (প্রমিত পরিমাপ বা বিনিময় মান অনুসারে গৃহীত যে কোনো পরিমাপের বিভাজন একক (নির্মাণ)।)
=> সুনির্দিষ্ট পরিমাপ -- (কোনো কিছু পরিমাপের জন্য সুনির্দিষ্ট পরিমাপ।)
=> পরিমাপ -- (কোন কিছু কি পরিমাপ আছে, তা পরিমাপ করতে পারে।)
=> বিমূর্তন -- (সুনির্দিষ্ট উদাহরণাদি থেকে গৃহীত সাধারণ নিদর্শনাদি দ্বারা স্ট্র সাধারণ ধারণা।)
=> অমূর্ত-সত্তা, অনূর্ণ-সত্তা, নিরূর্ণ-সত্তা, বিমূর্ত-সত্তা -- (শুধুমাত্র বিমূর্ত (দৈহিক নূর্ণযীন) অস্তিত্ব আছে এমন সত্তা।)
=> অস্তিত্ব, সত্তা -- (স্বতন্ত্র অস্তিত্ব আছে এমন ইন্দ্রিয়গ্রাহ্য বা জ্ঞাত বা সিদ্ধান্তকৃত এমন সত্তা কিছু (জীবিত বা জড়)।)

Sense 3
চক্র-অবলম্বক, স্পোক (spoke), পাখি (চক্র) -- (চক্রকেন্দ্র এবং চক্রের পরিধির মধ্যবর্তী অক্ষপ্রসারী সংযোজক দ্বারা স্ট্র অবলম্বক।)
=> অবলম্বক -- (অন্য কিছুর ভার ধারণ করে এমন যেকোন ডিভাইস।)
=> ডিভাইস -- (একটি সুনির্দিষ্ট উদ্দেশ্য সাধনের জন্য যান্ত্রিক উপায়ে উদ্ভাবিত।)
=> যান্ত্রিক-উপায়ে কৃত -- (একটি মানবসৃষ্ট (অথবা মানবসৃষ্টির পদ্ধতি), যা যান্ত্রিক উপায়ে সম্পন্ন হয়েছে।)
=> মানবসৃষ্টি -- (মানুষের সৃষ্ট একক (নির্মাণ) সমগ্র লক্ষ্যবস্তু।)
=> সমগ্র -- (উপকরণাদি দ্বারা সংযোজিত অবস্থায় যা একক (নির্মাণ) সত্তা হিসাবে পরিচিত।)
=> দৈহিক লক্ষ্যবস্তু -- (যা স্পর্শ করা যায় এবং দেখা যায় এবং ছায়া প্রদান করে।)
=> ইন্দ্রিয়গ্রাহ্য সত্তা, দৈহিক সত্তা -- (দৈহিক অস্তিত্ব আছে এমন সত্তা।)

Wordnet (শব্দজালিকা)

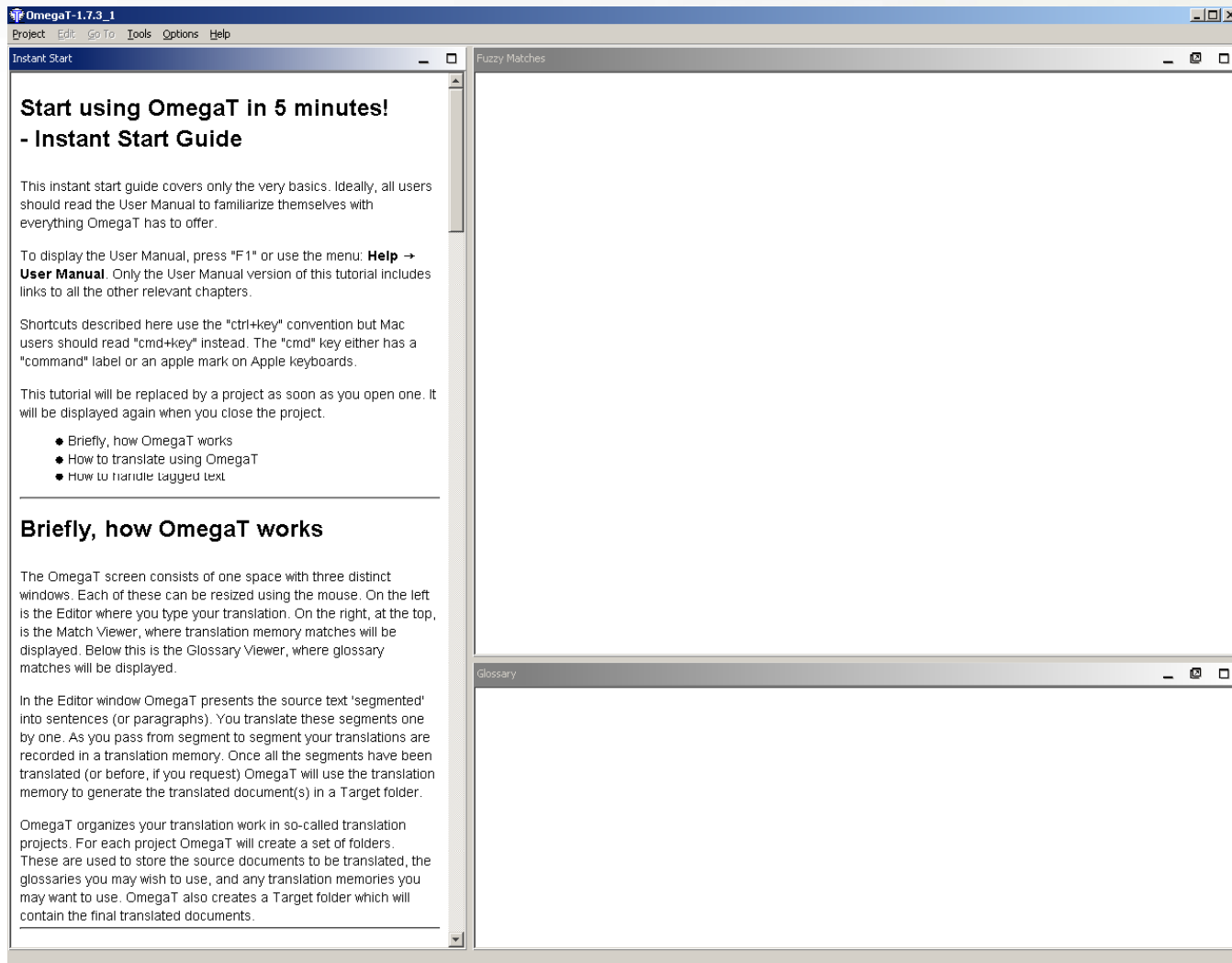
- Words 521
- Synsets 1227

Text Corpus

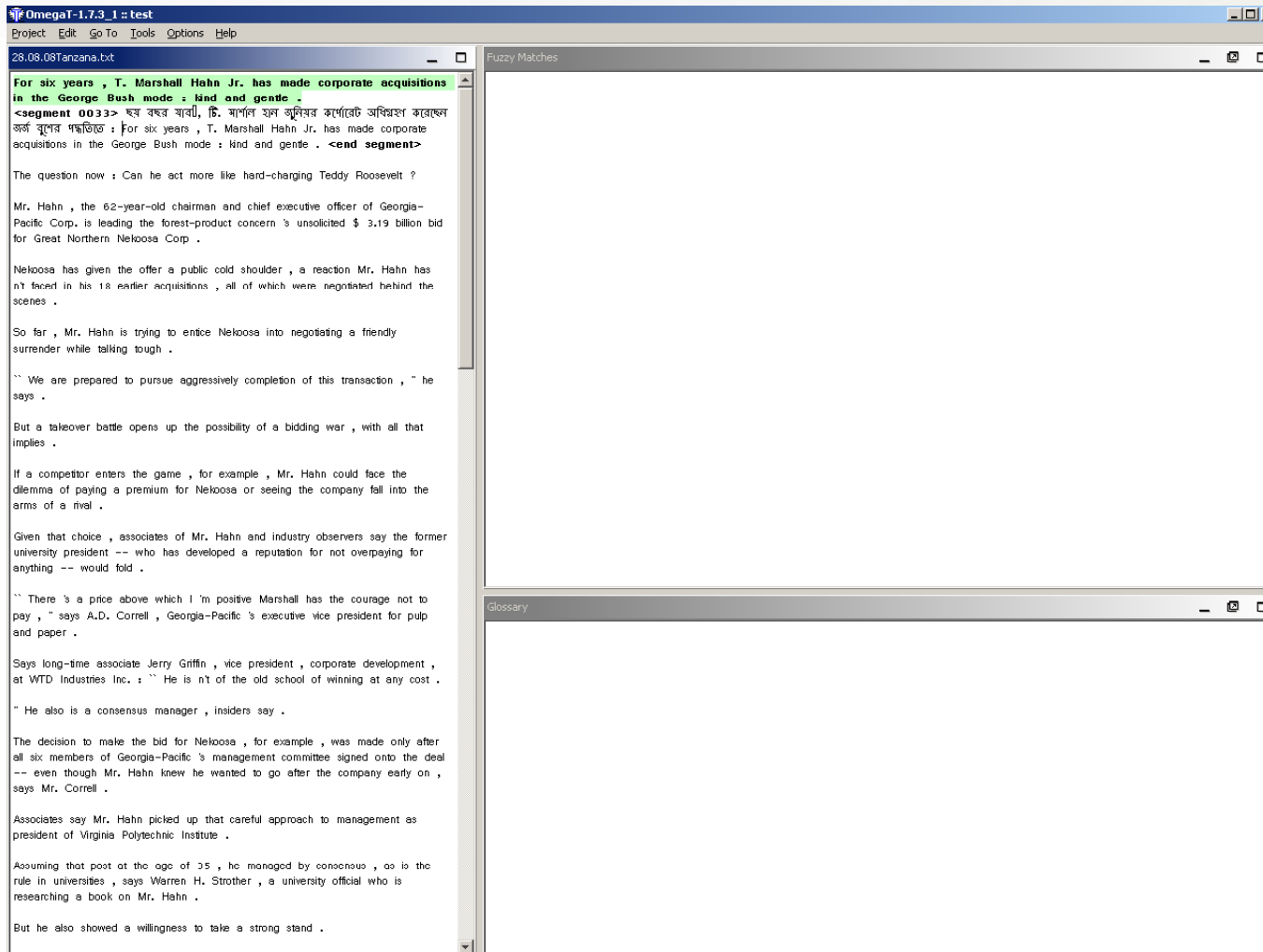
- Bangla text corpus
 - News corpus (prothom-alo)
 - 18,067,470 word tokens
 - 386,639 token types
- Parallel text corpus
 - Brown corpus
 - 1,00,000 words
 - OmegaT (for translation)
 - 1970 words translated



Text Corpus (OmegaT)



Text Corpus (OmegaT)



Text Corpus (OmegaT)

The screenshot displays the OmegaT-1.7.3_1 application window. The main text area shows a file named '28.08.08Tanzana.txt' with the following content:

ছয় বছর যাবা, টি. মার্শাল হান জুনিয়র কর্পোরেট অধিগ্রহণ করেছেন জর্জ বুশের পক্ষতিকে : দয়াসু
এবং ভদ্র ।

এখন প্রশ্ন হচ্ছে : উনি কি টেডি বুজভেল্টের মতো অদম্য হতে পারবেন ?

জর্জীয়া-প্যাসিফিক কর্পোরেশনের এই ৬২-বছর-বয়স্ক চেয়ারম্যান এবং প্রধান নির্বাহী কর্মকর্তা , হান
মাহেব নর্দার্ন লেকুসা কর্পোরেশনের বনজ-পণ্যের প্রতিষ্ঠানের অস্ট্রিভিকর ৩.১৯ বিলিয়ন ডলারের নিলামের
ডাকে এগিয়ে আছেন ।

**Nekoosa has given the offer a public cold shoulder , a reaction Mr. Hahn
has n't faced in his 18 earlier acquisitions , all of which were negotiated
behind the scenes .**

<segment 0004> তাঁর এই প্রশ্নাব লেকুসা জনসম্মুখে প্রত্যখ্যান করেছে , যে প্রতিক্রিয়ার মুখোমুখি
হান সাহেবকে তাঁর পূর্ববর্তী ১৮টি অধিগ্রহণে হতে হয়নি । **<end segment>**

So far , Mr. Hahn is trying to entice Nekoosa into negotiating a friendly surrender while
talking tough .

“ We are prepared to pursue aggressively completion of this transaction , ” he says .

But a takeover battle opens up the possibility of a bidding war , with all that implies .

If a competitor enters the game , for example , Mr. Hahn could face the dilemma of
paying a premium for Nekoosa or seeing the company fall into the arms of a rival .

Given that choice , associates of Mr. Hahn and industry observers say the former
university president -- who has developed a reputation for not overpaying for anything --
would fold .

“ There 's a price above which I 'm positive Marshall has the courage not to pay , ”
says A.D. Correll , Georgia-Pacific 's executive vice president for pulp and paper .

Says long-time associate Jerry Griffin , vice president , corporate development , at WTD
Industries Inc. : “ He is n't of the old school of winning at any cost .

“ He also is a consensus manager . Inside say

The right sidebar contains a 'Fuzzy Matches' section which is currently empty, and a 'Glossary' section with the following entries:

- Hahn = হান
- acquisitions = অধিগ্রহণ
- Nekoosa = লেকুসা
- public = জনসম্মুখে
- cold shoulder = প্রত্যখ্যান
- reaction = প্রতিক্রিয়া
- earlier = পূর্ববর্তী
- all = সব

POS Tagging

- the information about each word
- Important for statistical analysis
- 55 tags and 17 categories

POS Tagging

Level 1	Level 2	Tag	Example
Pronoun	Personal Pronoun	PRP	আমি, আমরা, তুমি, তোমরা, সে, তারা, আপনি, তিনি, তুই
	Question Pronoun	QPR	কে, কারা, যে, যারা
Adjective	Simple	JJ	সুন্দর, লাল, গরম, শ্রেষ্ঠ, শ্রেষ্ঠতর, শ্রেষ্ঠতম
	Verb Root	JJV	লাল, দুর্বল
	Question Adjective	QJJ	কেমন, যেমন
Vocative	Vocative	VOC	ওগো, ওরে, ওহে
Verb	Main Finite Verb	VB	করি, কর, করে, করাই, করলাম, করলে, করেছিস, করব, করাব
	Nonfinite Nominal	VBM	করা, করানো, পরা, পরানো
	Nonfinite Conditional	VBC	করলে, করালে
	Nonfinite Perfective	VBT	করে, গিয়ে
	Nonfinite	VBF	করতে, করাতে
	Finite Existential	VBE	হয়, হবে
	Nonfinite Existential	VBEF	হতে

POS Tagging

Level 1	Level 2	Tag	Example
Adverb	Adverb	RB	দ্রুত, হয়তো, অবশ্য, না, নাই, খুব, বেশী, অনেক
	Question Adverb	QRB	কেন, কিভাবে, যেভাবে
Conjunction	Coordinating	CC	এবং, ও, কিংবা, অথবা, নতুবা
	Compound Coordinating	CCC	না/CCC হয়/CC
	Suspicion	CN	যদি, পাছে
	Eternal Joining	CET	যেমন/CET ... তেমন/CET, যেই/CET ... সেই/CET, যখন/CET ... তখন/CET
	Subordinating	CS	যে, কেননা, বলে, এইজন্য
Postposition	Compound Subordinating	CSC	তাই/CSC বলে/CS, এই/CSC কারণে/CS
	Postposition	ON	দ্বারা, কর্তৃক, হতে, থেকে, জন্য, চেয়ে, চাইতে
Interjection	Interjection	UH	বাহ্!, ওহ্!, হায়!
Particle	Particle	RP	না, তো, বটে
	Question Particle	QRP	কি

POS Tagging

Level 1	Level 2	Tag	Example
Determiner	Common	DT	ওসব, তাবৎ, কোন, যেকোন, এই, ঐ
	Singular	DTS	এটি, ওটি
	Question Determiner	QDT	কোনটা, যেটা, কোনগুলো, যেগুলো, কোনসব
Quantifier	Quantifier	QF	সব, সকল, আরও, কম, কিছু
	Quantifier Number	QFNUM	১, ২, এক, তিন, একটি, পাঁচটি
	Question Quantifier	QQF	কত, যত, কতটুকু
Foreign Word	Foreign Word	FW	যেকোন বিদেশী শব্দ
Symbol	Symbol	SYM	বৈজ্ঞানিক বা অংকশাস্ত্রীয় যেকোন চিহ্ন, অন্যান্য
List Item Marker	List Item Marker	LS	a, b, (a), 1, 2.3.1, ক, ৩.১৩
Suffix	Postpositional	SFON	এ, য়, তে
	Accusative	SFAC	কে, রে, এরে, দিগকে, দিগেরে
	Possessive	SF\$	এর, দের

POS Tagging

Level 1	Level 2	Tag	Example
Punctuation Mark	Sentence Final Punctuation	.	, ?, !
	Comma	,	,
	; Semi-colon	:	;
	Dash, Double-Dash	-	-, --
	Left Parenthesis	((([
	Right Parenthesis))}]
	Opening Left Quote	LQ	' , "
	Closing Right Quote	RQ	' , "

POS Tagging

রূপালি/NNP বর্তমানে/NNT+SFON ঢাকা/NNPC মেডিকেল/NNPC
কলেজ/NNPC হাসপাতালে/NNP+SFON চিকিৎসাধীন/JJ রয়েছে/VB ।/.

Level 1	Level 2	Tag	Examples
Noun	Proper	NNP	অক্টোবর, ঢাকা, রহমান
Noun	Temporal	NNT	গতকাল, আজ
Noun	Compound Proper Noun	NNPC	আব্দুর/NNPC রহমান/NNPC বিশ্বাস/NNP
Adjective	Simple	JJ	সুন্দর, শ্রেষ্ঠ, দ্রুততম
Verb	Main Finite Verb	VB	করি, করলাম, করব
Suffixes	Adpositional	SFON	এ, য়, তে
Punctuation Marks	Sentence Final Punctuation	.	

Application

- Statistical Machine Translation
- Word sense disambiguation
- Language model

Conclusion

- 521 words in Wordnet
- 1970 words translated for Parallel Corpus