Rule Based Machine Translation

Pushpak Bhattacharyya
IIT Patna
GIAN course on NMT
4/12/17

Empiricism vs. Rationalism

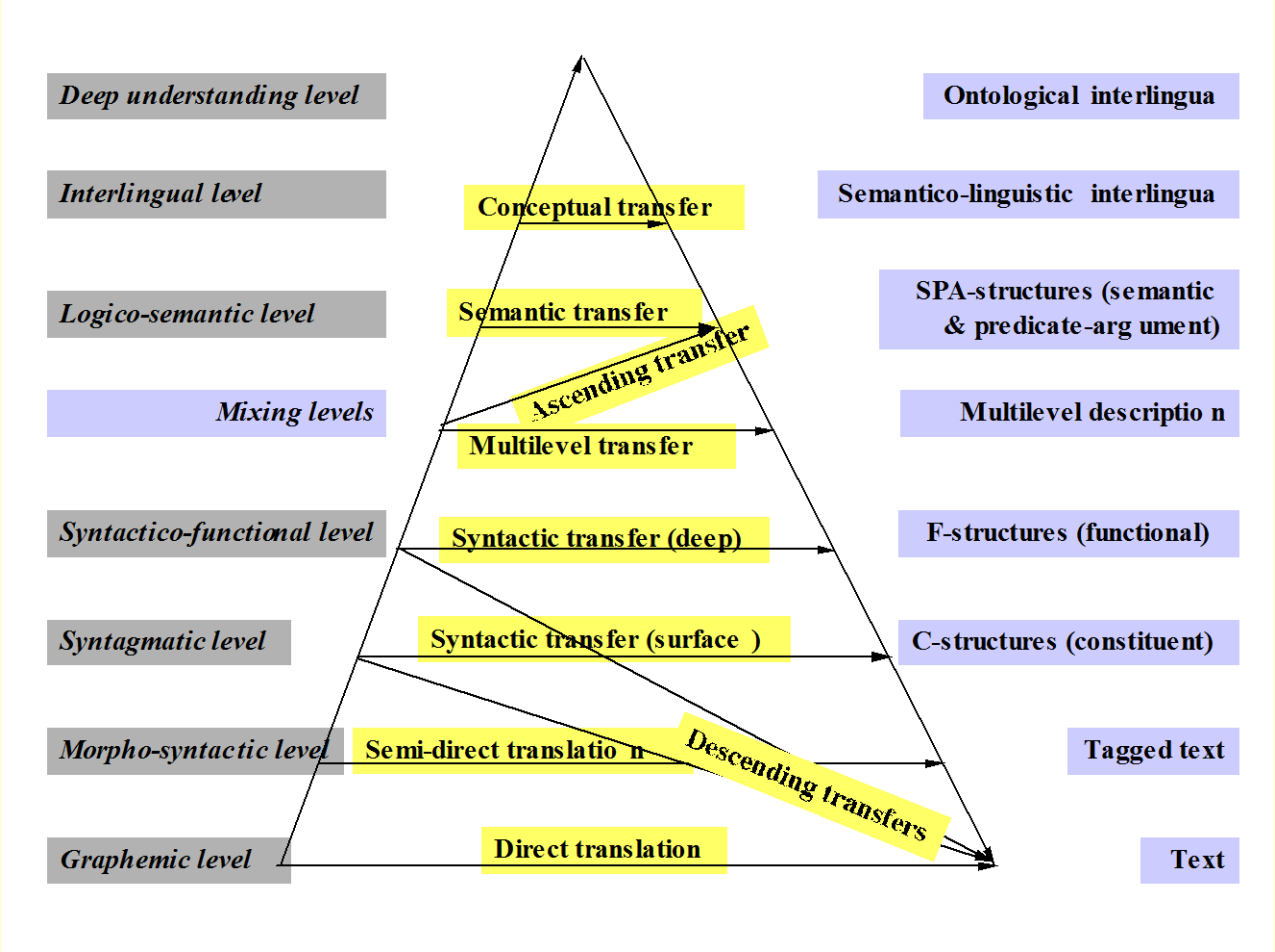
- Ken Church, “A Pendulum Swung too Far”, LILT, 2011
  - Availability of huge amount of data: what to do with it?
  - 1950s: Empiricism (Shannon, Skinner, Firth, Harris)
  - 1970s: Rationalism (Chomsky, Minsky)
  - 1990s: Empiricism (IBM Speech Group, AT & T)
  - 2010s: Return of Rationalism?

Introduction

- Machine Translation (MT) is a technique to translate texts from one natural language to another natural language using a machine
- Translated text should have two desired properties:
  - Adequacy: Meaning should be conveyed correctly
  - Fluency: Text should be fluent in the target language
- Translation between distant languages is a difficult task
  - Handling Language Divergence is a major challenge

Kinds of MT Systems

(point of entry from source to the target text)



Why is MT difficult: Language Divergence

- One of the main complexities of MT: *Language Divergence*
- Languages have different ways of expressing meaning
  - Lexico-Semantic Divergence
  - Structural Divergence

English-IL Language Divergence with illustrations from Hindi
(Dave, Parikh, Bhattacharyya, Journal of MT, 2002)

Language Divergence Theory:
Lexico-Semantic Divergences

- Conflational divergence
  - F: vomir; E: to be sick
  - E: *stab;* H: *churaa se maaranaa (knife-with hit)*
  - S: *Utrymningsplan;* E: *escape plan*
- Structural divergence
  - E: *SVO;* H: *SOV*
- Categorial divergence
  - Change is in POS category (many examples discussed)

Language Divergence Theory:
Lexico-Semantic Divergences (cntd)

- Head swapping divergence
  - E: *Prime Minister of India;* H: *bhaarat ke pradhaan mantrii (India-of Prime Minister)*
- Lexical divergence
  - E: *advise;* H: *paraamarsh denaa (advice give):* Noun Incorporation- very common Indian Language Phenomenon

Language Divergence Theory:
Syntactic Divergences

- Constituent Order divergence
  - E: *Singh, the PM of India, will address the nation today;* H: *bhaarat ke pradhaan mantrii, singh, … (India-of PM, Singh…)*
- Adjunction Divergence
  - E: *She will visit here in the summer;* H: *vah yahaa garmii meM aayegii (she here summer-in will come)*
- Preposition-Stranding divergence
  - E: *Who do you want to go with?;* H: *kisake saath aap jaanaa chaahate ho? (who with…)*

Language Divergence Theory:
Syntactic Divergences

- Null Subject Divergence
  - E: I will go; H: jaauMgaa (subject dropped)
- Pleonastic Divergence
  - E: *It is raining*; H: *baarish ho rahii haai* (*rain happening is*: no translation of *it*)

Language Divergence exists even for close cousins

(Marathi-Hindi-English: case marking and postpositions do not transfer easily)

- *संनिहित भूत (immediate past)*
  - *कधी आलास? हा येतो इतकाच ! (M)*
  - *कब आये? बस अभी आया । (H)*
  - *When did you come? Just now (I came) (H)*
  - *kakhan ele? ei elaam/eshchhi (B)*
- *निःसंशय भविष्य (certainty in future)*
  - *आता तो मार खातो खास ! (M)*
  - *अब वह मार खायगा ही ! (H)*
  - *He is in for a thrashing. (E)*
  - *ekhan o maar khaabei/khaachhei*
- *आश्वासन (assurance)*
  - *मी तुम्हाला उद्या भेटतो. (M)*
  - *मैं आप से कल मिलता हूँ। (H)*
  - *I will see you tomorrow. (E)*
  - *aami kaal tomaar saathe dekhaa korbo/korchhi (B)*

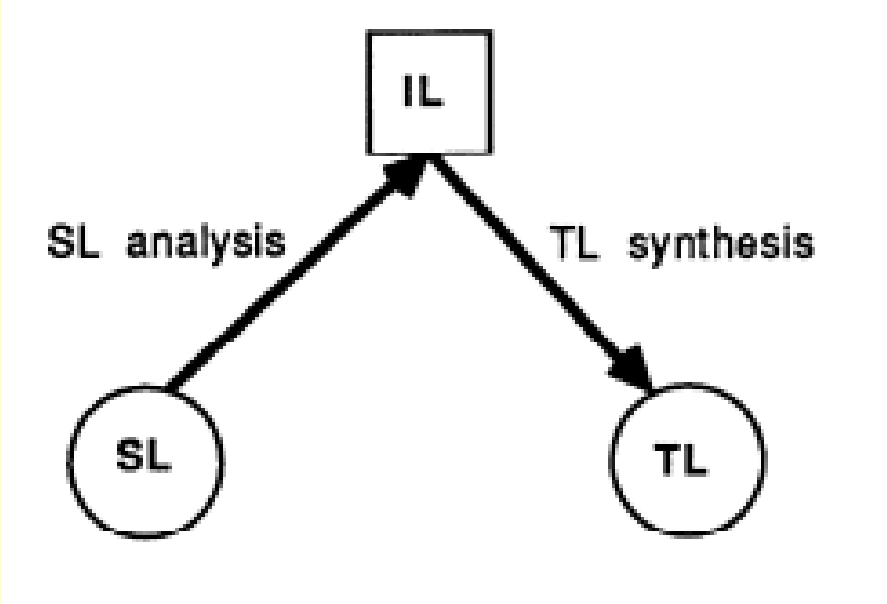
Interlingua Based MT

KBMT

- *Nirenberg, Knowledge Based Machine Translation, Machine Translation, 1989.*

- Forerunner of many interlingua based MT efforts

IL based MT: schematic



Specification of KBMT

- Source languages: English and Japanese
- Target languages: English and Japanese
- Translation paradigm: Interlingua;
- Computational architecture: A distributed, coarsely parallel system
- Subworld (domain) of translation: personal computer installation and maintenance manuals.

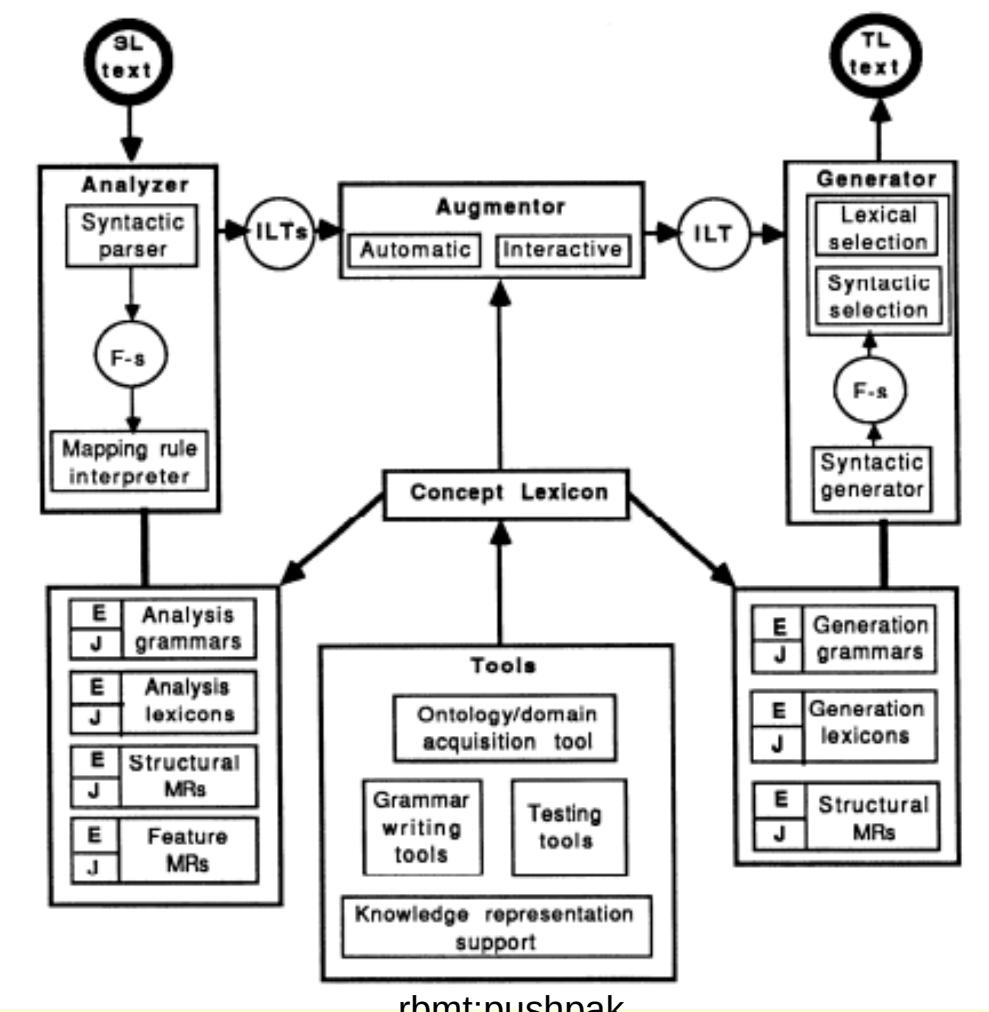
Knowledge base of KBMT (1/2)

- An ontology (domain model) of about 1,500 concepts
- Analysis lexicons: about 800 lexical units of Japanese and about 900 units of English
- Generation lexicons: about 800 lexical units of Japanese and about 900 units of English

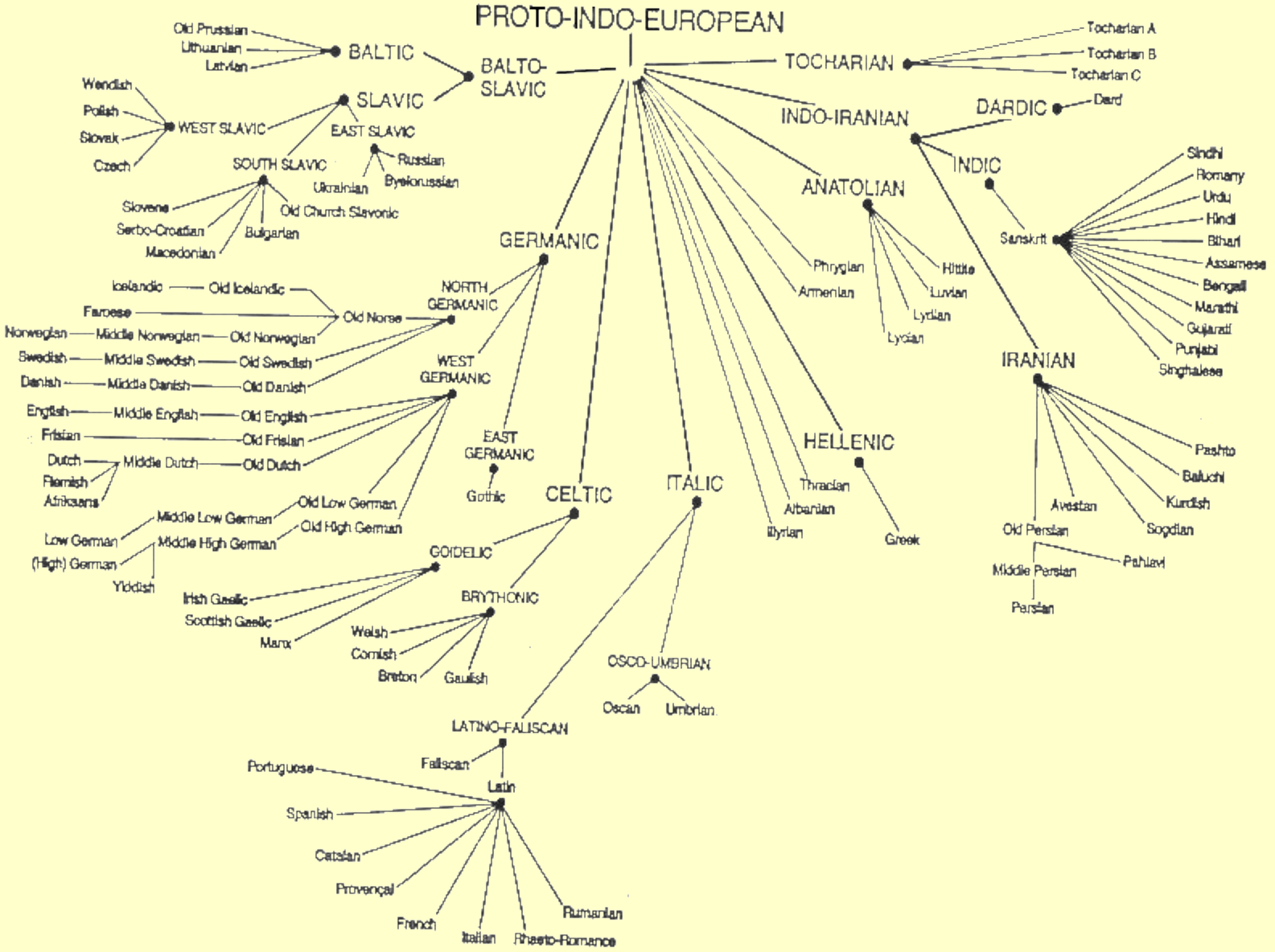
Knowledge base of KBMT (2/2)

- Analysis grammars for English and Japanese;
- Generation grammars for English and Japanese
- Specialized syntax-semantics structural mapping rules

Architecture of KBMT



Digression: language typology


PROTO-INDO-EUROPEAN
BALTO-SLAVIC
BALTIC
Old Prussian
Lithuanian
Latvian
SLAVIC
WEST SLAVIC
Wendish
Polish
Slovak
Czech
EAST SLAVIC
Russian
Ukrainian
Byelorussian
SOUTH SLAVIC
Slovene
Serbo-Croatian
Macedonian
Bulgarian
Old Church Slavonic
GERMANIC
NORTH GERMANIC
Icelandic
Old Icelandic
Faroese
Old Norse
Norwegian
Middle Norwegian
Old Norwegian
Swedish
Middle Swedish
Old Swedish
Danish
Middle Danish
Old Danish
WEST GERMANIC
English
Middle English
Old English
Frisian
Old Frisian
Dutch
Middle Dutch
Old Dutch
Flemish
Afrikaans
Low German
Middle Low German
Old Low German
(High) German
Middle High German
Old High German
Yiddish
EAST GERMANIC
Gothic
CELTIC
GOIDELIC
Irish Gaelic
Scottish Gaelic
Manx
BRYTHONIC
Welsh
Cornish
Breton
Gaulish
ITALIC
OSCO-UMBRIAN
Oscan
Umbrian
LATINO-FALISCAN
Faliscan
Latin
Portuguese
Spanish
Catalan
Provençal
French
Italian
Rhaeto-Romance
Rumanian
TOCHARIAN
Tocharian A
Tocharian B
Tocharian C
INDO-IRANIAN
DARDIC
Dard
INDIC
Sanskrit
Sindhi
Romany
Urdu
Hindi
Bihari
Assamese
Bengali
Marathi
Gujarati
Punjabi
Singhalese
IRANIAN
Old Persian
Avestan
Middle Persian
Pahlavi
Parsian
Pashto
Baluchi
Kurdish
Sogdian
ANATOLIAN
Hittite
Luvian
Lydian
Lycian
Phrygian
Armenian
HELLENIC
Greek
Thracian
Albanian
Illyrian

Language and dialects

- There are 6909 living languages (2009 census)
- Dialects far outnumber the languages
- Language varieties are called dialects
  - if they have no standard or codified form,
  - if the speakers of the given language do not have a state of their own,
  - if they are rarely or never used in writing (outside reported speech),
  - if they lack prestige with respect to some other, often standardised, variety.

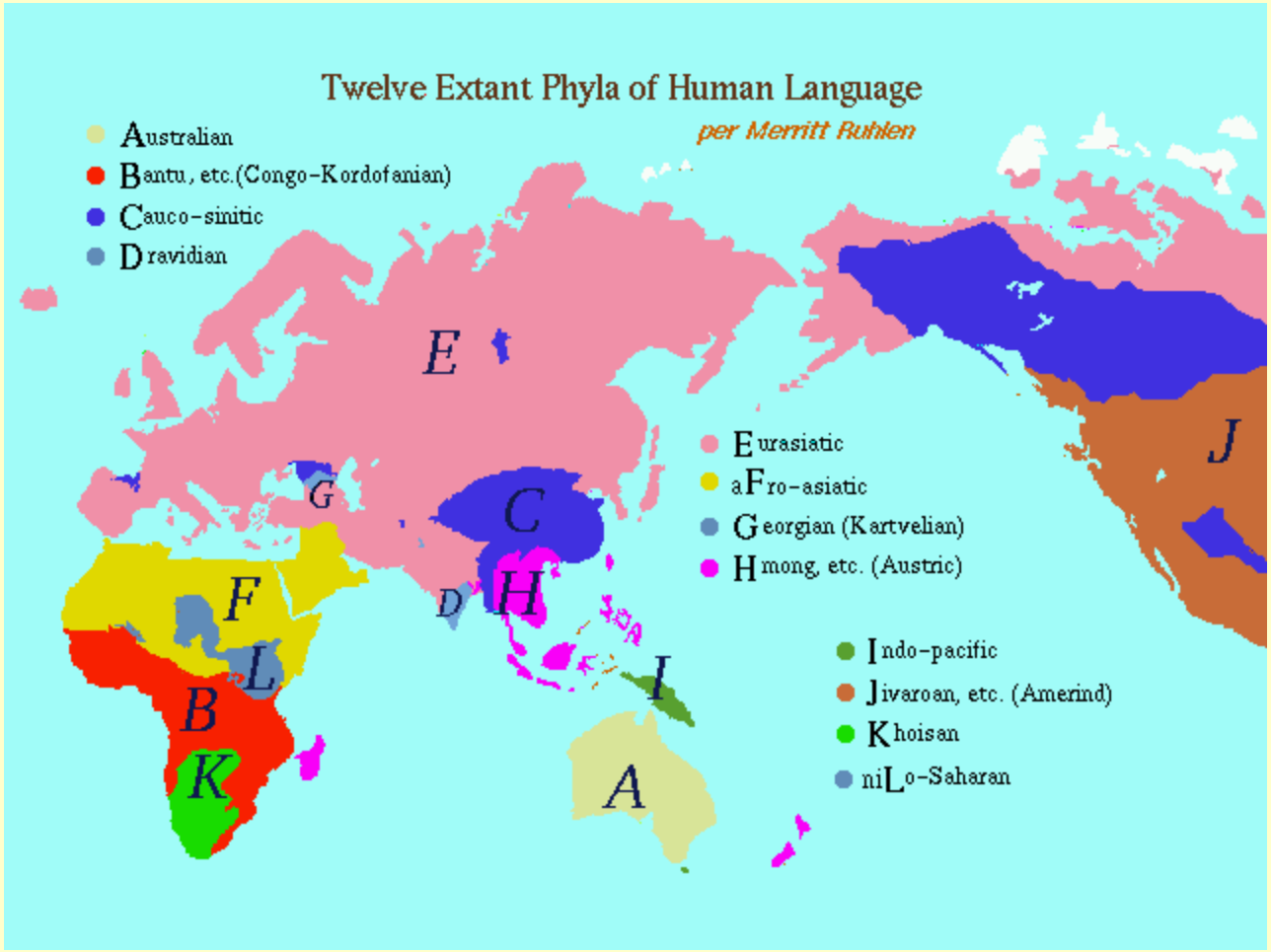
“Linguistic” interlingua

- Three notable attempts at creating interlingua
  - “Interlingua” by IALA (International Auxiliary Language Association)
  - “Esperanto”
  - “Ido”

The Lord's Prayer in Esperanto	Interlingua version	Latin version	English version (traditional)
1. Patro Nia, kiu estas en la ĉielo, via nomo estu sanktigita. 2. Venu via regno, 3. plenumiĝu via volo, kiel en la ĉielo, tiel ankaŭ sur la tero. 4. Nian panon ĉiutagan donu al ni hodiaŭ. 5. Kaj pardonu al ni niajn ŝuldojn, kiel ankaŭ ni pardonas al niaj ŝuldantoj. 6. Kaj ne konduku nin en tenton, sed liberigu nin de la malbono. Amen.	1. Patre nostre, qui es in le celos, que tu nomine sia sanctificate; 2. que tu regno veni; 3. que tu voluntate sia facite como in le celo, etiam super le terra. 4. Da nos hodie nostre pan quotidian, 5. e pardona a nos nostre debitas como etiam nos los pardona a nostre debitores. 6. E non induce nos in tentation, sed libera nos del mal. Amen.	1. Pater noster, qui es in cælis, sanctificetur nomen tuum. 2. Adveniat regnum tuum. 3. Fiat voluntas tua, sicut in cælo, et in terra. Panem nostrum 4. quotidianum da nobis hodie, 5. et dimitte nobis debita nostra, sicut et nos dimittimus debitoribus nostris. 6. Et ne nos inducas in tentationem, sed libera nos a malo. Amen.	1. Our Father, who art in heaven, hallowed be thy name; 2. thy kingdom come, 3. thy will be done. on earth, as it is in heaven. 4. Give us this day our daily bread; 5. and forgive us our debts as we have forgiven our debtors. 6. And lead us not into temptation, but deliver us from evil. Amen.

“Interlingua” and “Esperanto”

- *Control languages* for “Interlingua”: French, Italian, Spanish, Portuguese, English, German and Russian
- Natural words in “interlingua”
- “manufactured” words in Esperanto, using heavy agglutination
- Esperanto word for "hospital“: **mal·san·ul·ej·o: mal** (opposite), **san** (health), **ul** (person), **ej** (place), **o** (noun)


Twelve Extant Phyla of Human Language
per Merritt Ruhlen
Australian
Bantu, etc.(Congo-Kordofanian)
Cauco-sinitic
Dravidian
Eurasiatic
aFro-asiatic
Georgian (Kartvelian)
Hmong, etc. (Austric)
Indo-pacific
Jivaroan, etc. (Amerind)
Khoisan
niLo-Saharan
E
G
C
F
D
H
L
B
K
I
A
J

Language Universals vs. Language Typology

- “Universals” is concerned with what human languages have in common, while the study of typology deals with ways in which languages differ from each other.

Typology: basic word order

- SOV (Japanese, Tamil, Turkish etc.)
- SVO (Fula, Chinese, English etc.)
- VSO (Arabic, Tongan, Welsh etc.)

"Subjects tend strongly to precede objects."

Typology: Morphotactics of singular and plural

- No expression: Japanese *hito* 'person', pl. *hito*
- Function word: Tagalog *bato* 'stone', pl. *mga bato*
- Affixation: Turkish *ev* 'house', pl. *ev-ler*; Swahili *m-toto* 'child', pl. *wa-toto*
- Sound change: English *man*, pl. *men*; Arabic *rajulun* 'man', pl. *rijalun*
- Reduplication: Malay *anak* 'child', pl. *anak-anak*

Typology: Implication of word order

Due to its SVO nature, English has:

- preposition+noun (in the house)
- genitive+noun (Tom's house) or noun+genitive (the house of Tom)
- auxiliary+verb (will come)
- noun+relative clause (the cat that ate the rat)
- adjective+standard of comparison (better than Tom)

Typology: motion verbs (1/3)

- Motion is expressed differently in different languages, and the differences turn out to be highly significant.
- There are two large types:
  - verb-framed and
  - satellite-framed languages.

Typology: motion verbs (2/3)

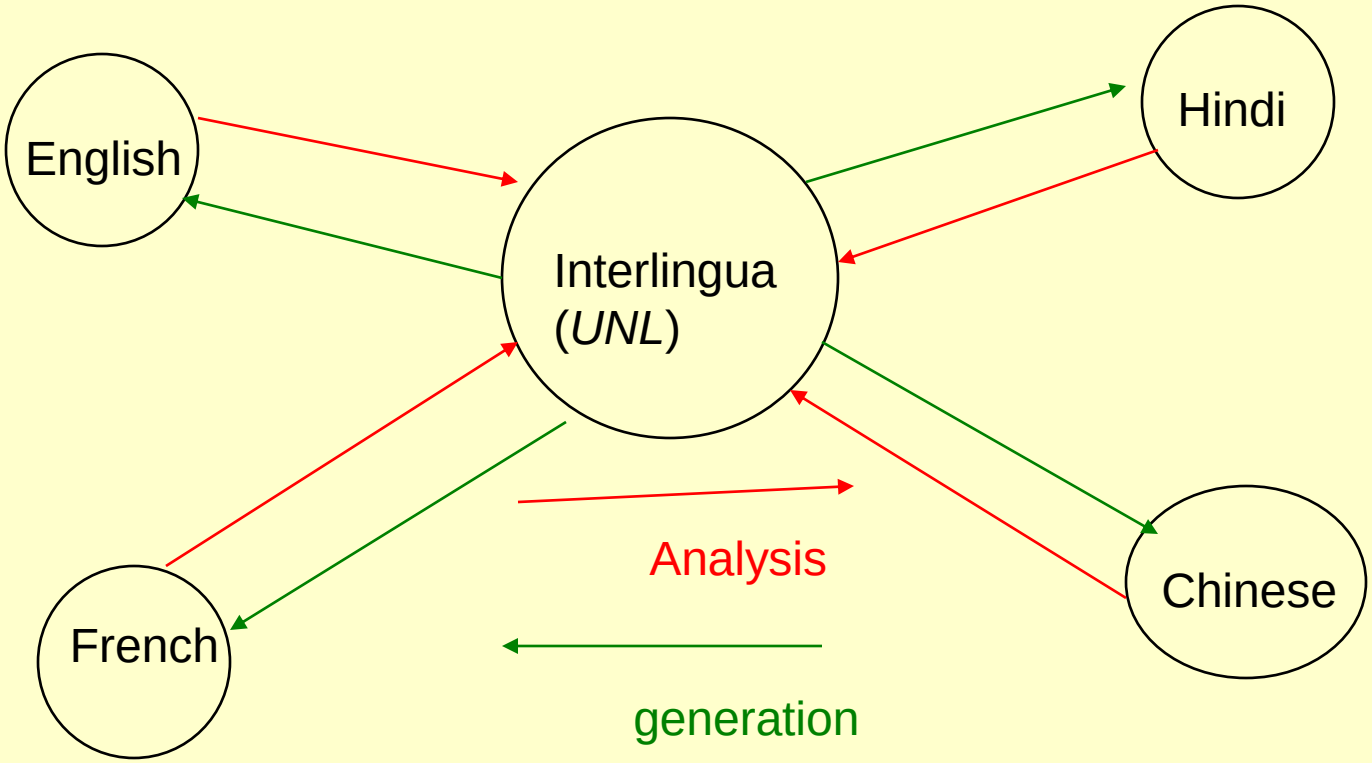
- In satellite-framed languages like English, the motion verb typically also expresses **manner** or **cause**:
  - *The bottle floated out of the cave. (Manner)*
  - *The napkin blew off the table. (Cause)*
  - *The water rippled down the stairs (Manner)*

Typology: motion verbs (3/3)

- Verb-framed languages express manner and cause, not in the verb, but in a more peripheral element:
- Spanish
  - *La botella salió de la cueva flotando.*
  - *The bottle exited the cave floatingly*
- Japanese
  - *Bin-ga dookutsu-kara nagara-te de -ta*
  - *bottle-NOM cave-from float-GER exit-PAST*

Back to Interlingua

MT: *EnConverion + Deconversion*



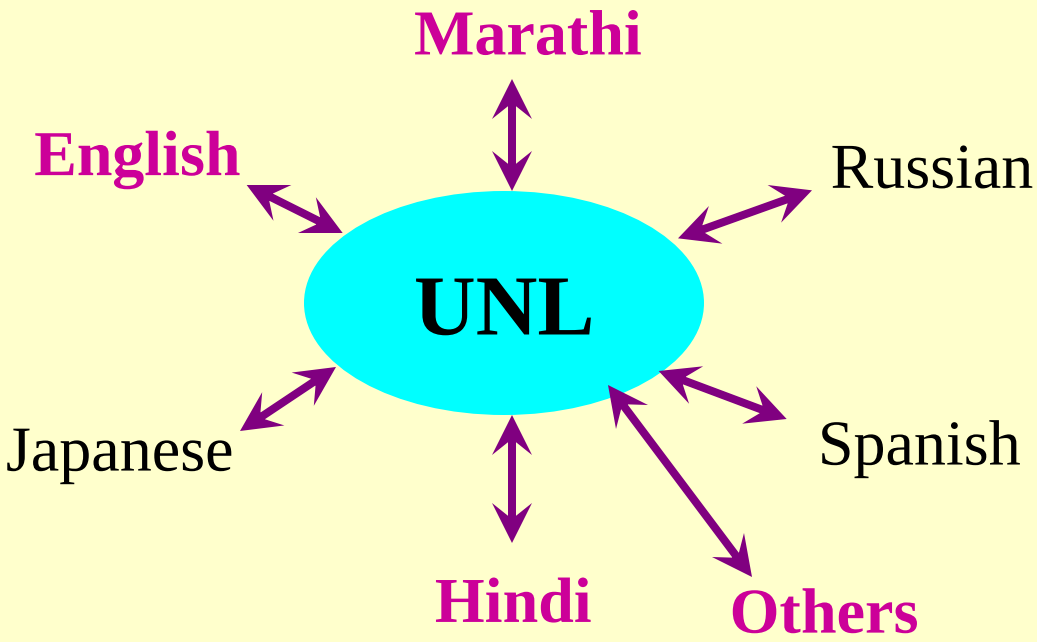
Challenges of interlingua generation

- Mother of NLP problems - Extract meaning from a sentence!
- Almost all NLP problems are sub-problems
  - Named Entity Recognition
  - POS tagging
  - Chunking
  - Parsing
  - Word Sense Disambiguation
  - Multiword identification
  - and the list goes on...

UNL: a United Nations project

- Started in 1996
- 10 year program
- 15 research groups across continents
- First goal: generators
- Next goal: analysers (needs solving various ambiguity problems)
- Current active language groups
  - UNL_French (GETA-CLIPS, IMAG)
  - UNL_Hindi (IIT Bombay with additional work on UNL_English)
  - UNL_Italian (Univ. of Pisa)
  - UNL_Portugese (Univ of Sao Paolo, Brazil)
  - UNL_Russian (Institute of Linguistics, Moscow)
  - UNL_Spanish (UPM, Madrid)

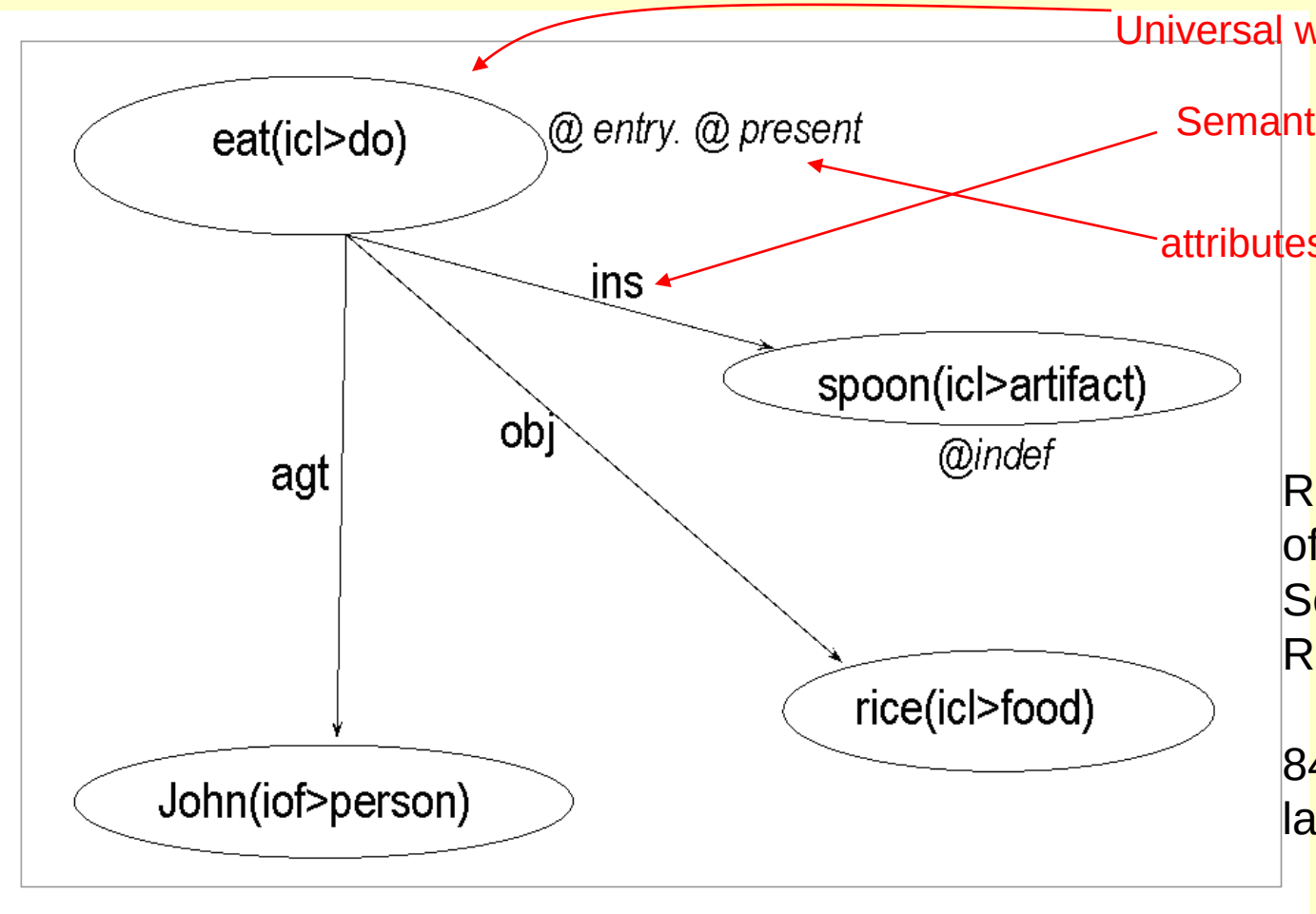
World-wide
Universal Networking Language (UNL) Project



- Language independent meaning representation.

UNL represents knowledge:

John eats rice with a spoon



Repository of 42 Semantic Relations and 84 attribute labels

Sentence embeddings

Deepa claimed that she had composed a poem.

[UNL]

agt(claim.@entry.@past, Deepa)

obj(claim.@entry.@past, :01)

agt:01(compose.@past.@entry.@complete, she)

obj:01(compose.@past.@entry.@complete, poem.@indef)

[\UNL]

The Lexicon

He forwarded the mail to the minister.

Content words:

[forward] {} “forward(icl>send)” (V,VOA) <E,0,0>;

[mail] {} “mail(icl>message)” (N,PHSCL,INANI) <E,0,0>;

[minister] {} “minister(icl>person)” (N,ANIMT,PHSCL,PRSN) <E,0,0>;

The Lexicon (cntd)

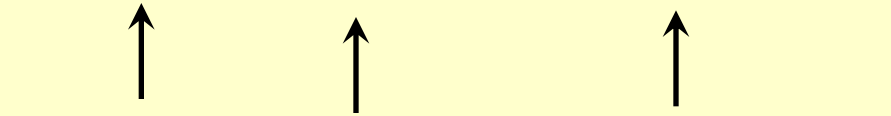
He forwarded the mail to the minister.

function words:

[he] {} “he” (PRON,SUB,SING,3RD) <E,0,0>;

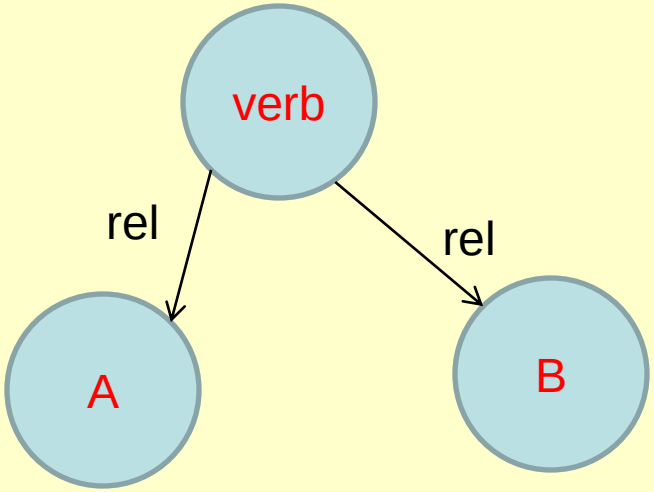
[the] {} “the” (ART,THE) <E,0,0>;

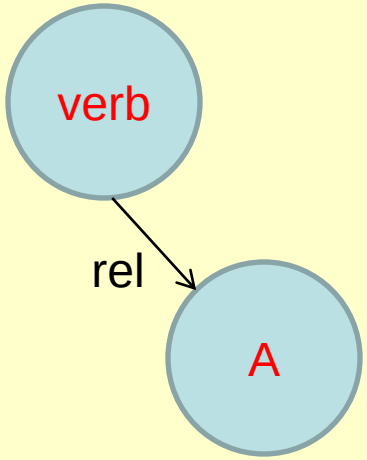
[to] {} “to” (PRE,#TO) <E,0,0>;



How to obtain UNL expresssions

- UNL nerve center is the verb
- English sentences:
  - *A <verb> B* or
  - *A <verb>*





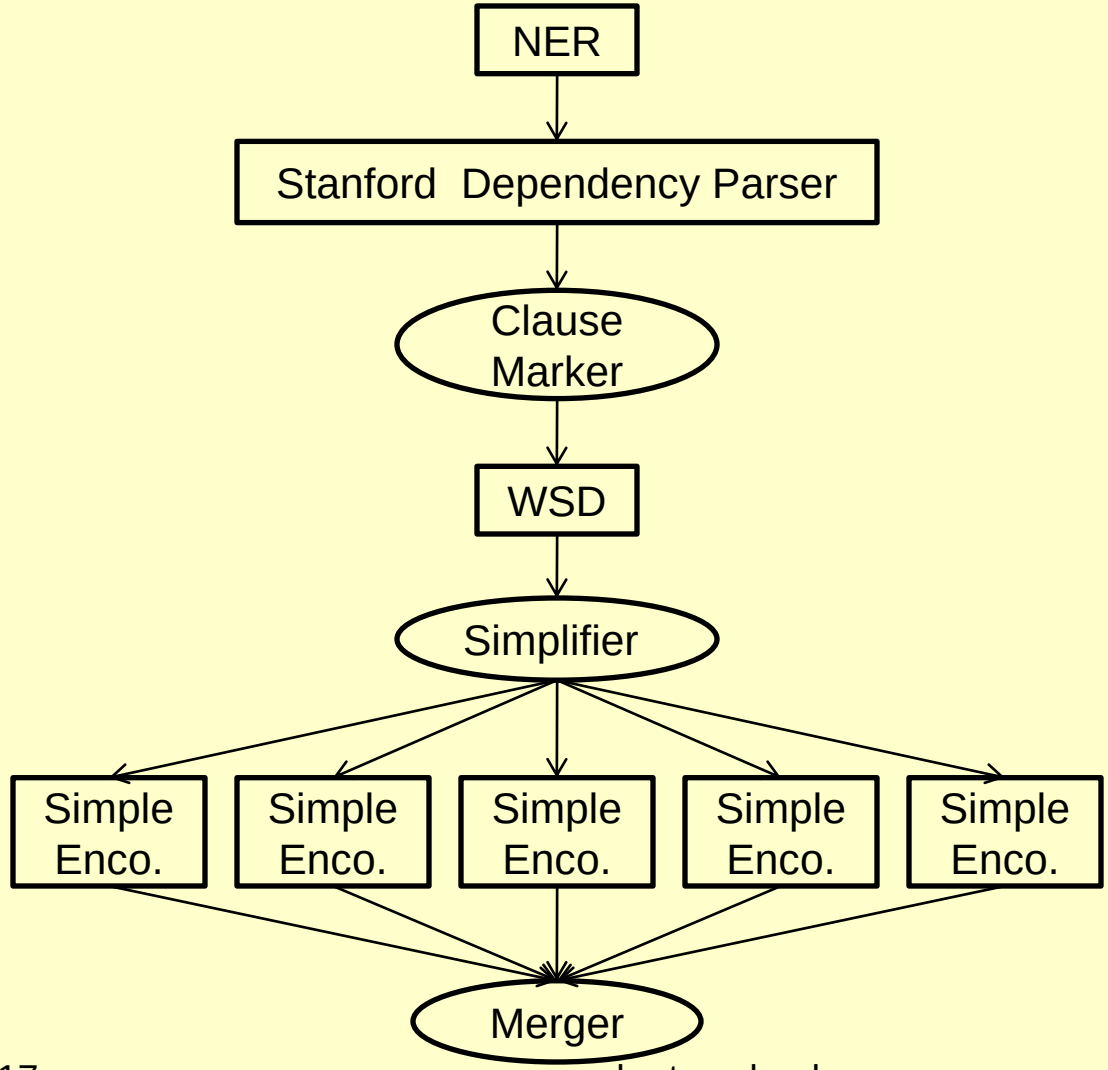
Recurse on A and B

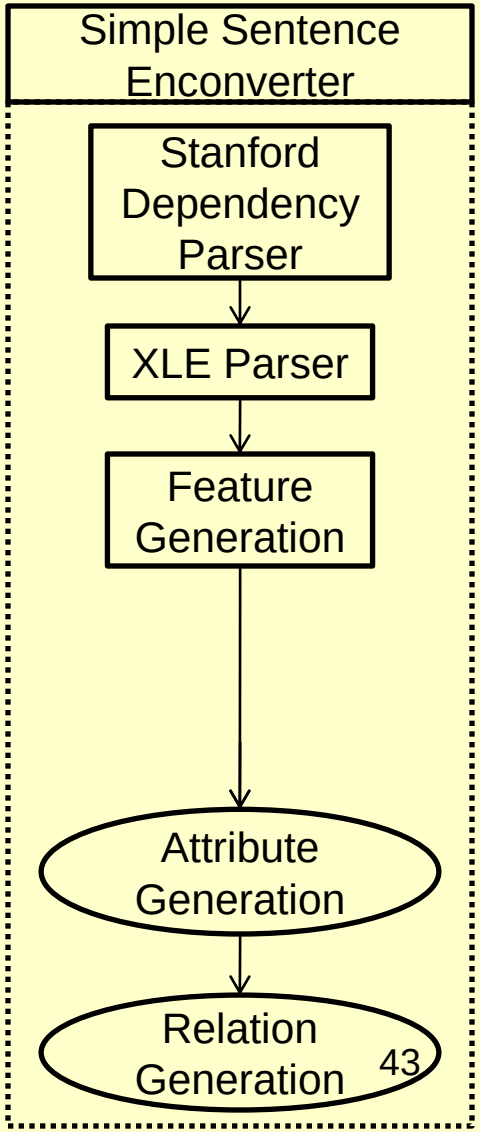
Enconversion

Long sequence of masters theses: *Anupama, Krishna, Sandip, Abhishek, Gourab, Subhajit, Janardhan* in collaboration with *Rajat, Jaya, Gajanan, Rajita*

Many publications

System Architecture





Complexity of handling long sentences

- Problem
  - As the length (number of words) increases in the sentence the XLE parser fails to capture the dependency relations precisely
  - The UNL enconversion system is tightly coupled with XLE parser
  - Resulting in fall of accuracy.

Solution

- Break long sentences to smaller ones
  - Sentence can only be broken at the clausal levels: Simplification
  - Once broken they needs to be joined: Merging
- Reduce the reliance of the Encoverter on XLE parser
  - The rules for relation and attribute generation are formed on Stanford Dependency relations

Text Simplifier

- It simplifies a complex or compound sentence.
- Example:
  - Sentence: *John went to play and Jane went to school.*
  - Simplified: *John went to play. Jane went to school.*
- The new simplifier is developed on the clause-markings and the inter-clause relations provided by the Stanford dependency parser

UNL Merging

- Merger takes multiple UNLs and merges them to form one UNL.

- Example-
  - Input:
    - *agt(go@entry, John), pur(go@entry, play)*
    - *agt(go@entry, Mary), gol(go@entry, school)*
  - *Output:*
    - *agt(go@entry, John), pur(go@entry, play)*
    - *and (go, :01)*
    - *agt:01(go@entry, Mary), gol:01(go@entry, school)*

- There are several cases based on how the sentence was simplified. Merger uses rules for each of the cases to merge them.

NLP tools and resources for UNL generation

Tools

- Stanford Named Entity Recognizer(NER)
- Stanford Dependency Parser
- Word Sense Disambiguation(WSD) system
- Xerox Linguistic Environment(XLE) Parser

Resource

- Wordnet
- Universal Word Dictionary(UW++)

System: Processing Units

Syntactic Processing

- NER
- Stanford Dependency Parser
- XLE Parser

Semantic Processing

- Stems of words
- WSD
- Noun originating from verbs
- Feature Generation
  - Noun features
  - Verb features

SYNTACTIC PROCESSING

Syntactic Processing: NER

- NER tags the named entities in the sentence with their types.

- Types covered
  - Person
  - Place
  - Organization

- Input: *Will Smith was eating an apple with a fork on the bank of the river.*
- Output: *<PERSON>Will Smith</PERSON> was eating an apple with a fork on the bank of the river.*

Syntactic Processing: Stanford Dependency Parser (1/2)

- Stanford Dependency Parser parses the sentence and produces
  - POS tags of words in the sentence
  - Dependency parse of the sentence

Syntactic Processing: Stanford Dependency Parser (2/2)

Input	*Will-Smith was eating an apple with a fork on the bank of the river.*

POS tags	Will-Smith/NNP was/VBD eating/VBG an/DT apple/NN with/IN a/DT fork/NN on/IN the/DT bank/NN of/IN the/DT river/NN ./.

Dependency Relations
nsubj(eating-3, Will-Smith-1) aux(eating-3, was-2) det(apple-5, an-4) dobj(eating-3, apple-5) prep(eating-3, with-6) det(fork-8, a-7) pobj(with-6, fork-8) prep(eating-3, on-9) det(bank-11, the-10) pobj(on-9, bank-11) prep(bank-11, of-12) det(river-14, the-13) pobj(of-12, river-14)

Syntactic Processing: XLE Parser

- XLE parser generates dependency relations and some other important information.

Input	*Will-Smith was eating an apple with a fork on the bank of the river.*

Dependency Relations
‘OBJ'(eat:8,apple:16)
'SUBJ'(eat:8,will-smith:0)
'ADJUNCT'(apple:16,on:20)
‘OBJ'(on:20,bank:22)
'ADJUNCT'(apple:16,with:17)
'OBJ'(with:17,fork:19)
'ADJUNCT'(bank:22,of:23)
‘OBJ'(of:23,river:25)

Important Information
'PASSIVE'(eat:8,'-')
'PROG'(eat:8,'+‘)
'TENSE'(eat:8,past)
'VTYPE'(eat:8,main)

SEMANTIC PROCESSING

Semantic Processing: Finding stems

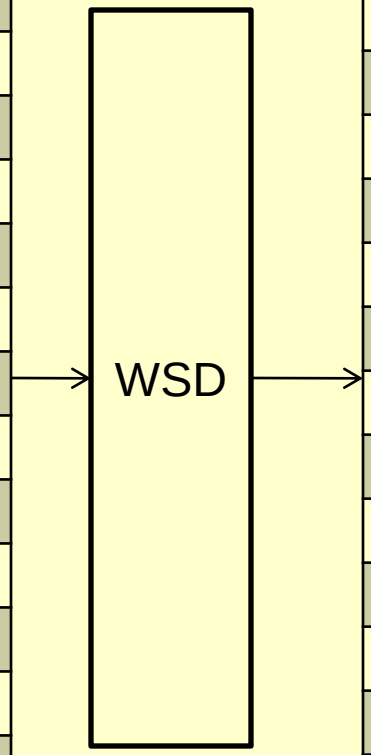
- Wordnet is used for finding the stems of the words.

Input	*Will-Smith was eating an apple with a fork on the bank of the river.*

Word	**Stem**
was	be
eating	eat
apple	apple
fork	fork
bank	bank
river	river

Semantic Processing: WSD

Word	POS
Will-Smith	Proper noun
was	Verb
eating	Verb
an	Determiner
apple	Noun
with	Preposition
a	Determiner
fork	Noun
on	Preposition
the	Determiner
bank	Noun
of	Preposition
the	Determiner
river	Noun



Word	Synset-id	Sense-id
Will-Smith	-	-
was	02610777	be%2:42:03::
eating	01170802	eat%2:34:00::
an	-	-
apple	07755101	apple%1:13:00::
with	-	-
a	-	-
fork	03388794	fork%1:06:00::
on	-	-
the	-	-
bank	09236472	bank%1:17:01::
of	-	-
the	-	-
river	09434308	river%1:17:00::

Semantic Processing: Universal Word Generation

- The sense-ids are used to get the Universal Words from the Universal Word dictionary (UW++).

Input	*Will-Smith was eating an apple with a fork on the bank of the river.*

Word	**Universal Word**
Will-Smith	Will-Smith(iof>PERSON)
eating	eat(icl>consume, equ>consumption)
apple	apple(icl>edible fruit>thing)
fork	fork(icl>cutlery>thing)
bank	bank(icl>slope>thing)
river	bank(icl>stream>thing)

Semantic Processing: Nouns originating from verbs

- Some nouns are originated from verbs.
- The frame *noun1 of noun2,* where *noun1* has originated from *verb1,* should generate relation *obj(verb1, noun2)* instead of any relation between *noun1* and *noun2.*

Algorithm
• Traverse the hypernym-path of the noun to its root. • Any word on the path is *"action",* then the noun has originated from a verb.

Examples
• *removal of garbage = removing garbage* • *Collection of stamps = collecting stamps* • *Wastage of food = wasting food*

Features of UWs

- Each UNL relation requires some properties of their UWs to be satisfied.

Relation	Word	Property
agt(uw1, uw2)	uw1	Volitional verb
met(uw1, uw2)	uw2	Abstract thing
dur(uw1, uw2)	uw2	Time period
ins(uw1, uw2)	uw2	Concrete thing

- We capture these properties as features
- We classify the features into **Noun** and **Verb Features**

Semantic Processing: Noun Features

Algorithm

- Add word.NE_type to word.features
- Traverse the hypernym-path of the noun to its root.
- Any word on the path matches a keyword, corresponding feature generated

Words in path to hypernym-root	Feature
time	TIME
time_unit	TIME UNIT
time_period	TIME PERIOD
measure, unit	MEASURE
quantity, number	QUANTITY
person	PERSON
living_thing	ANIMATE
part	PART
group	GROUP
abstract_entity	ABSTRACT
physical_entity	CONCRETE
room, structure, facilty, location, way	PLACE

Noun	Feature
Will-Smith	PERSON, ANIMATE
bank	PLACE
fork	CONCRETE

Types of verbs

- Verb features are based on the types of verbs
- There are three types
  - **Do**: Verbs that are performed volitionally
  - **Occur**: Verbs that are performed involuntarily or just happen without any performer
  - **Be**: Verbs which tells about existence of something or about some fact
- We classify these three into **do** and **not-do**

Semantic Processing: Verb Features

- The sentence frames of verbs are inspected to find whether a verb is *"do"* type of verb or not.
- *"do"* verbs are volitional verbs which are performed voluntarily
- Heuristic:
  - *Something/somebody …..s something/somebody* then "*do*" type
- This heuristic fails to mark some volitional verbs as *"do",* but marks all non-volitional as "*not-do*".

Word	Type	Feature
was	not do	-
eating	do	do

GENERATION OF RELATIONS AND ATTRIBUTES

Relation Generation (1/2)

- Relations handled separately
- Similar relations handled in cluster
- Rule based approach
- Rules are applied on
  - Stanford Dependency relations
  - Noun features
  - Verb features
  - POS tags
  - XLE generated information

Relation Generation (2/2)

Input	*Will-Smith was eating an apple with a fork on the bank of the river.*

Conditions	Relations Generated
nsubj(eating, Will-Smith) Will-Smith.feature=ANIMATE eat is *do* type and *active*	agt(eat, Will Smith)
dobj(eating, apple) *eat* is *active*	obj(eat, apple)
prep(eating, with), pobj(with, fork) fork.feature=CONCRETE	ins(eat, fork)
prep(eating, on), pobj(on, bank) *ba*nk.feature=PLACE	plc(eat, bank)
prep(bank, of), pobj(of, river) bank.POS = noun river.feature is not PERSON	mod(bank, river)

Attribute Generation (1/2)

- Attributes handled in clusters
- Attributes are generated by rules
- Rules are applied on
  - Stanford dependency relations
  - XLE information
  - POS tags

Attribute Generation (2/2)

Input	*Will-Smith was eating an apple with a fork on the bank of the river.*

Conditions	Attributes Generated
eat is verb of main clause *VTYPE(eat, main)* *PROG(eat,+)* *TENSE(eat,past)*	eat @entry @progress @past
det(apple, an)	apple@indef
det(fork, a)	fork@indef
det(bank, the)	bank@def
det(river, the)	river@def

EVALUATION OF OUR SYSTEM

Results

- Evaluated on EOLSS corpus
  - 7000 sentences from EOLSS corpus
- Evaluated on 4 scenarios
  - Scenario1: Previous system (tightly coupled with XLE parser)
  - Scenario2: Current system (tightly coupled with Stanford parser)
  - Scenario3: Scenario2 – (Simplifier + Merger)
  - Scenario4: Scenario2 – XLE parser
- Scenario3 lets us know the impact of Simplifier and Merger
- Scenario4 lets us know the impact of XLE parser

Results: Metrics

- $t_{gen} = relation_{gen}\ (UW1_{gen}, UW2_{gen})$
- $t_{gold} = relation_{gold}\ (UW1_{gold}, UW2_{gold})$

- $t_{gen} = t_{gold}$ *if*
  - $relation_{gen} = relation_{gold}$
  - $UW1_{gen} = UW1_{gold}$
  - $UW2_{gen} = UW2_{gold}$
- If $UW_{gen} = UW_{gold}$, let $a_{match}(UW_{gen}, UW_{gold})$ = {# attributes matched in UW_{gen} and UW_{gold}}
- $count(UW_x)$= {# attributes in UW_x}

Results: Relation-wise accuracy

$$\text{Precision } p_{relation} = \frac{\#(t_{gen,relation} = t_{gold,relation})}{\#t_{gen,relation}}$$

$$\text{Recall } r_{relation} = \frac{\#(t_{gen,relation} = t_{gold,relation})}{\#t_{gold,relation}}$$

$$\text{F-score } f_{relation} = \frac{2 * p_{relation} * r_{relation}}{(p_{relation} + r_{relation})}$$

Results: Overall accuracy

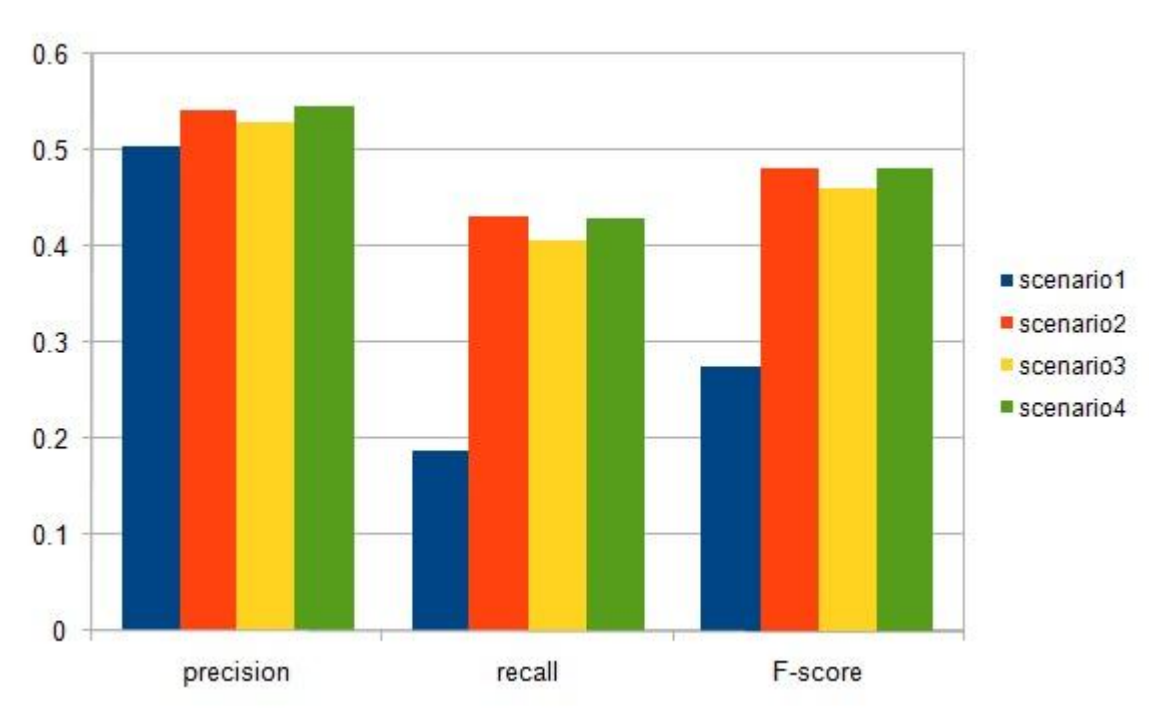
$$\text{Precision } p_{overall} = \frac{\#(t_{gen} = t_{gold})}{\#t_{gen}}$$

$$\text{Recall } r_{overall} = \frac{\#(t_{gen} = t_{gold})}{\#t_{gold}}$$

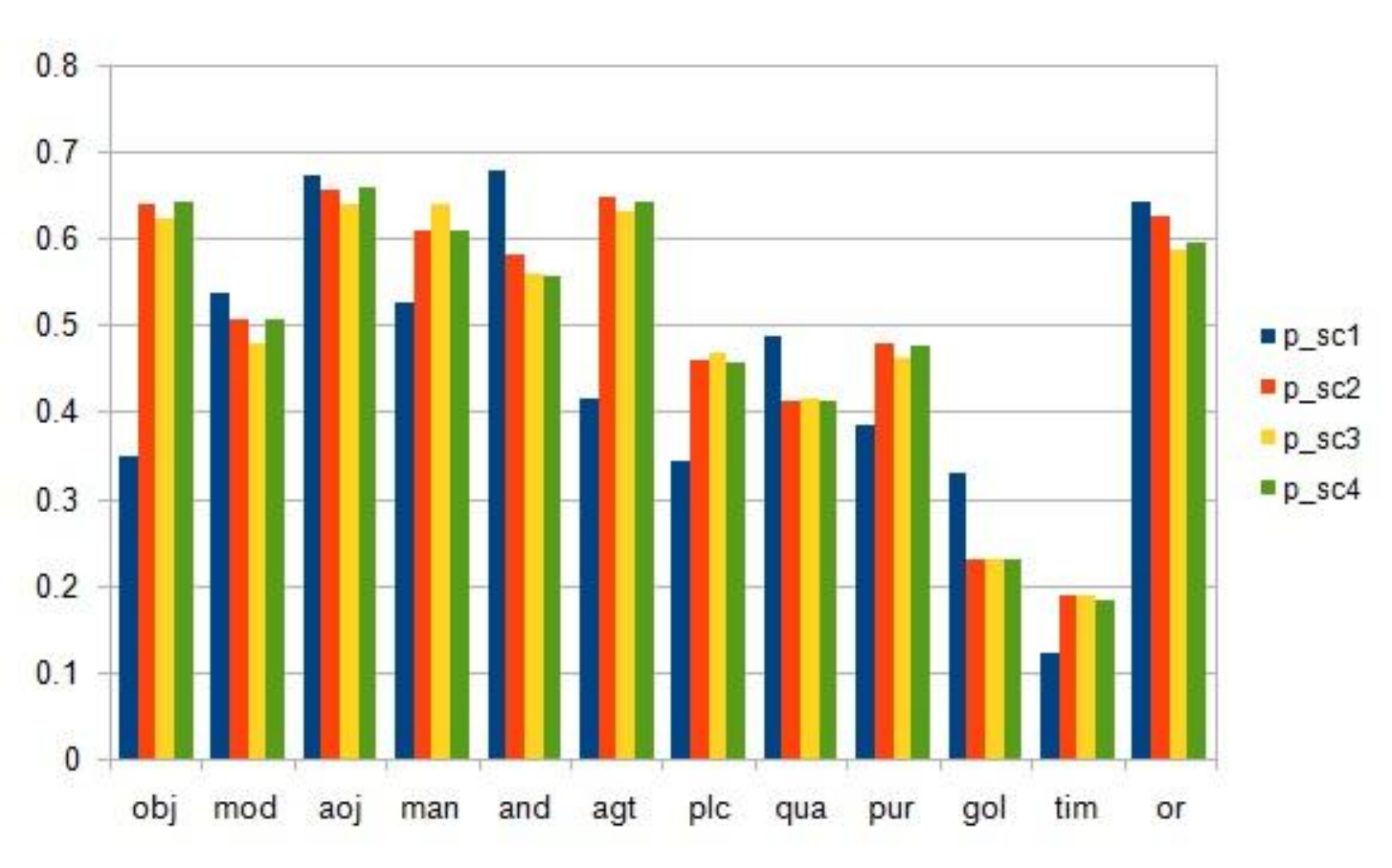
$$\text{F-score } f_{overall} = \frac{2 * p_{overall} * r_{overall}}{(p_{overall} + r_{overall})}$$

$$\text{Attribute accuracy } = \frac{\sum_{UW_{gold}=UW_{gen}, UW_{gold} \in GEN, UW_{gen} \in GOLD} a_{match}(UW_{gold}, UW_{gen})}{\sum_{UW_{gold}=UW_{gen}} count(UW_{gold})}$$

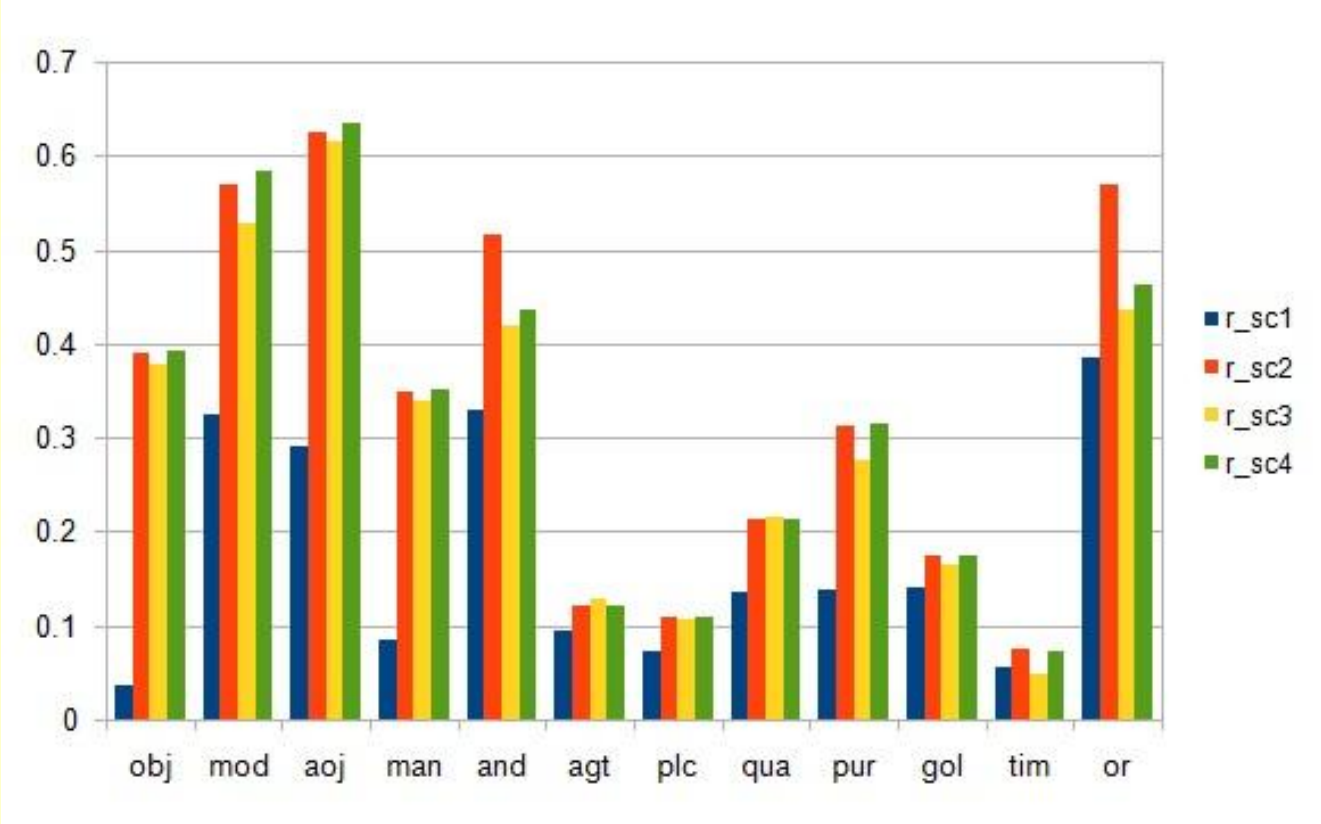
Results: Relation



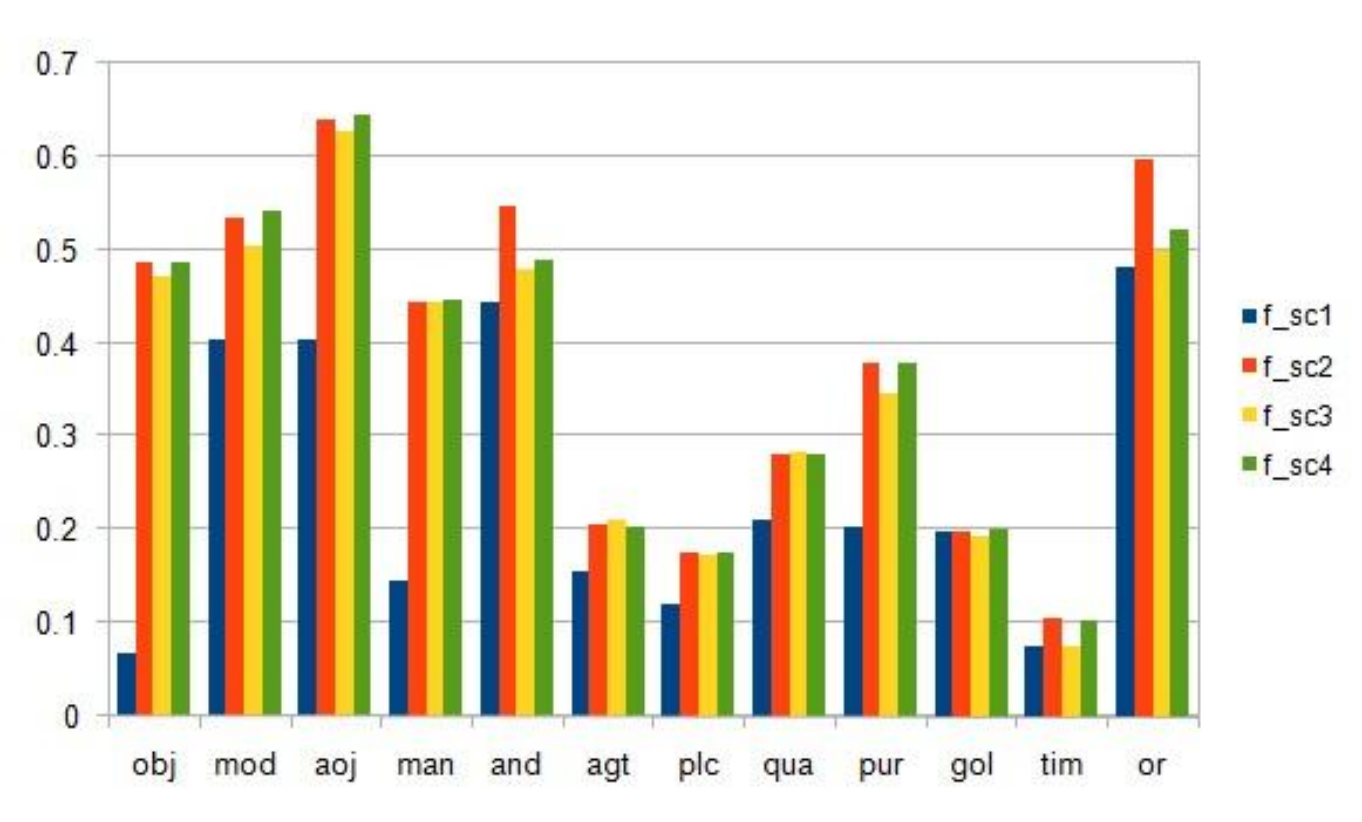
Results: Relation-wise: Precision



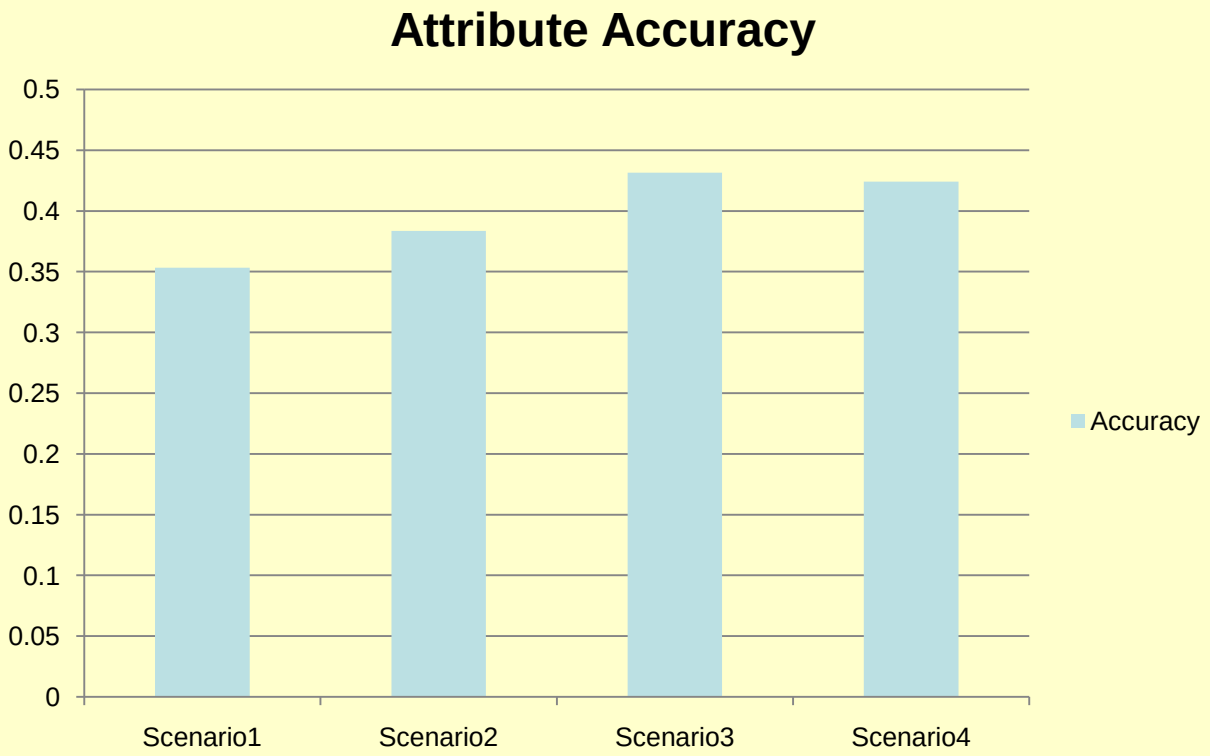
Results: Relation-wise: Recall



Results: Relation-wise: F-score

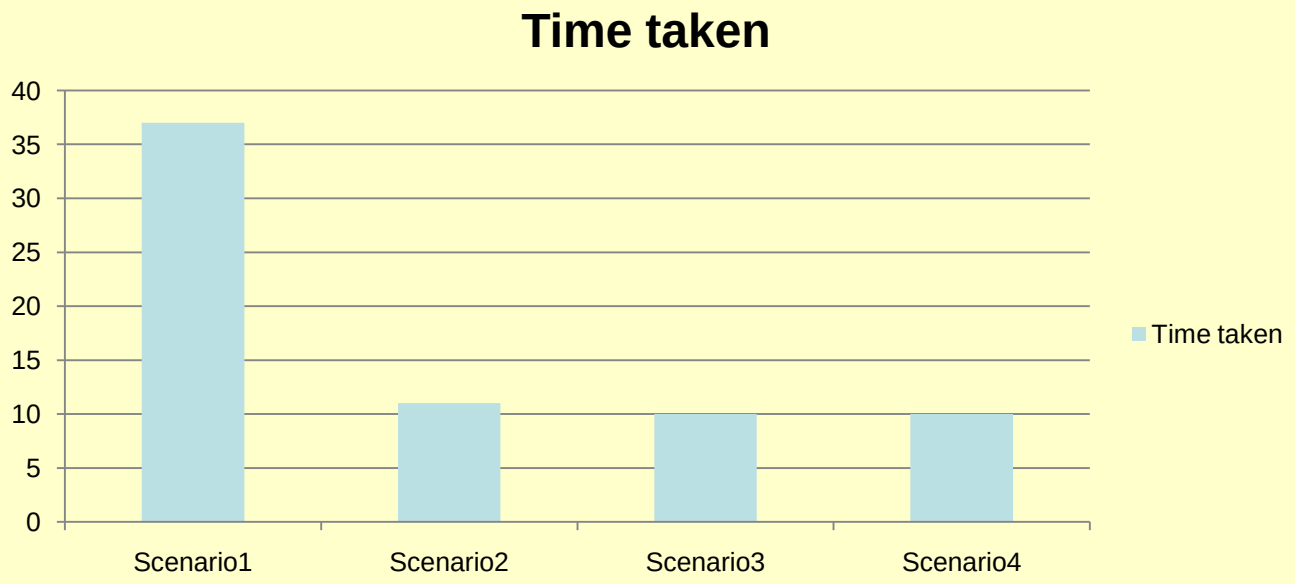


Results: Attribute



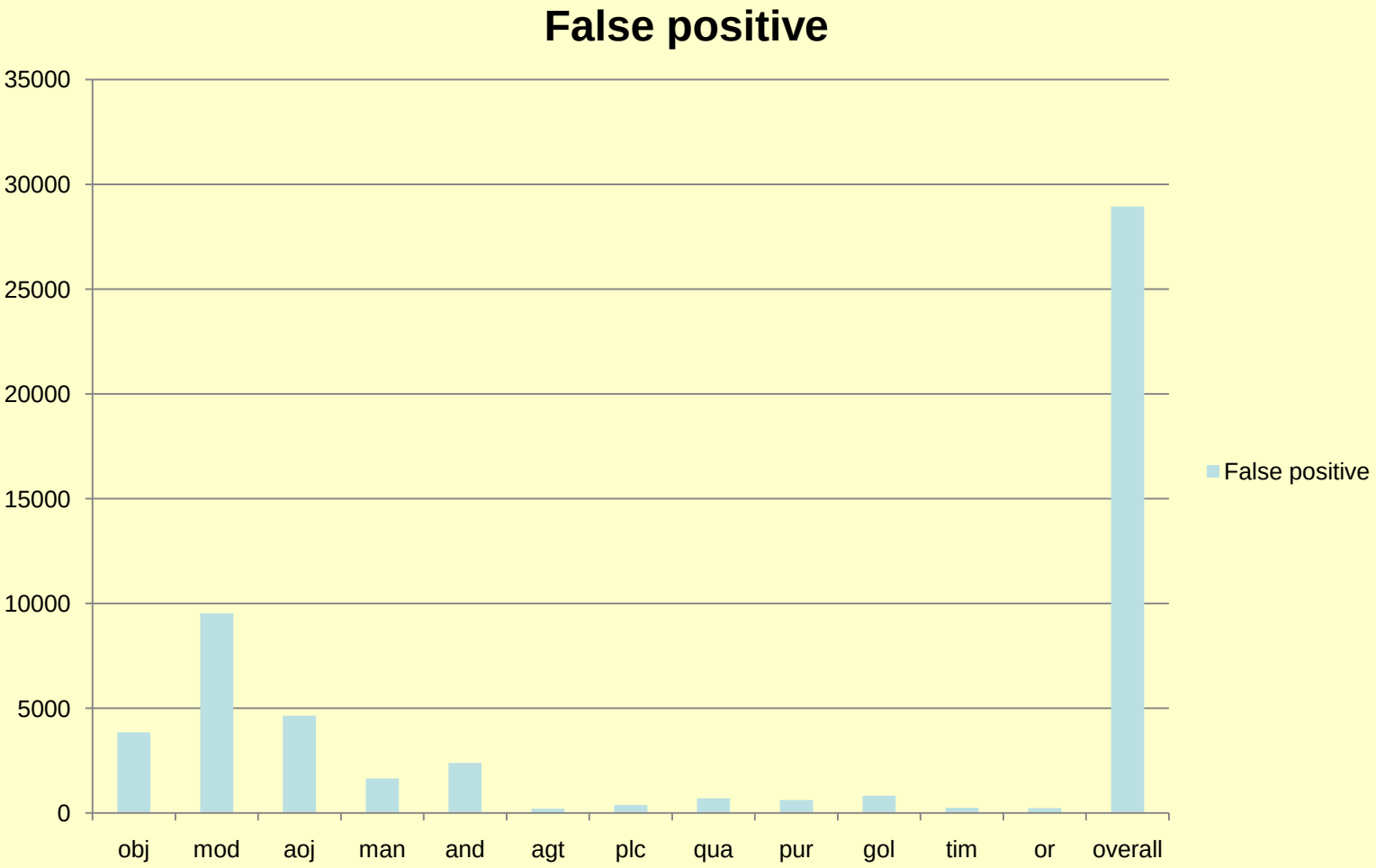
Results: Time-taken

- Time taken for the systems to evaluate 7000 sentences from EOLSS corpus.

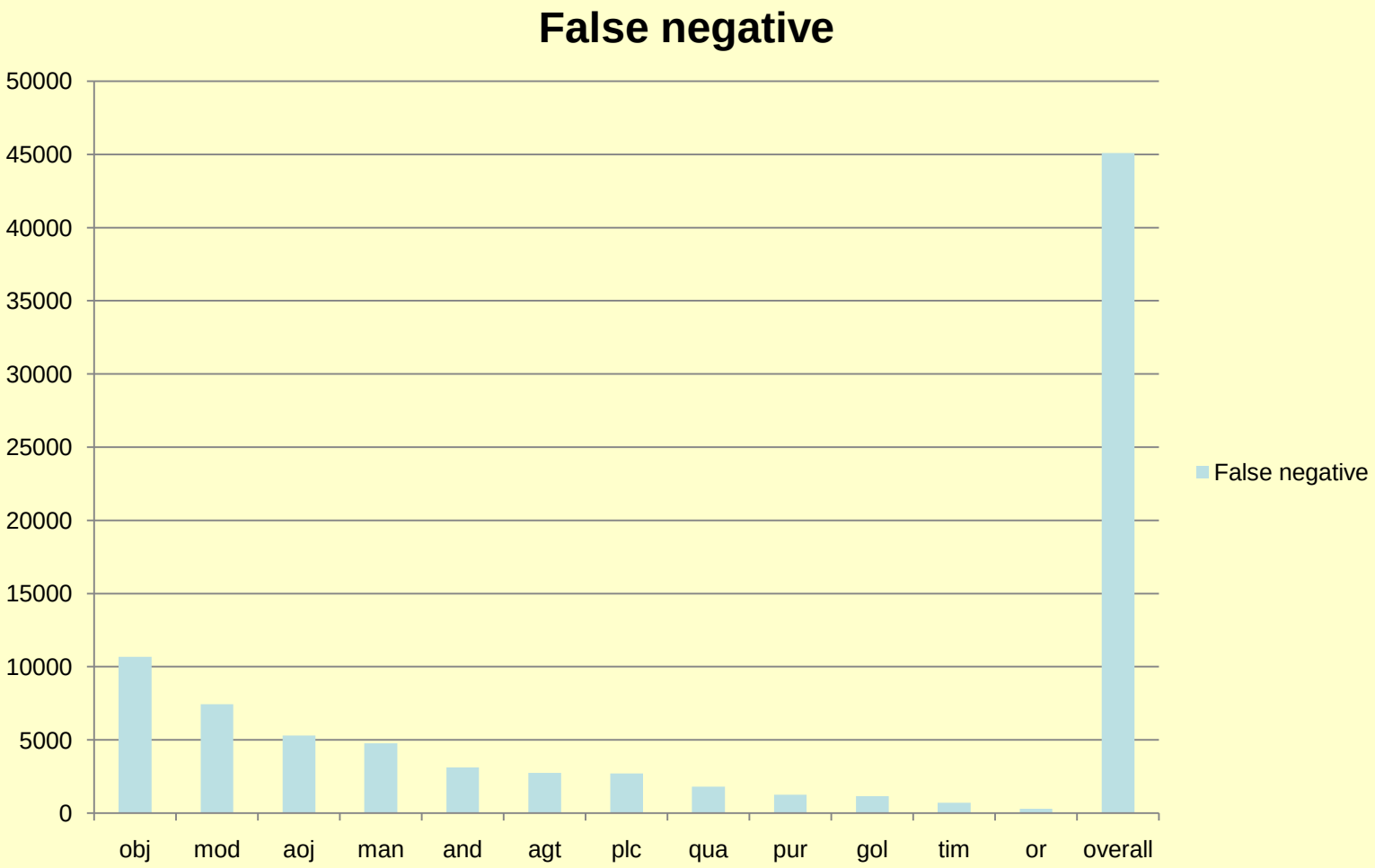


ERROR ANALYSIS

Wrong Relations Generated



Missed Relations



Case Study (1/3)

- Conjunctive relations
  - Wrong apposition detection
    - *James loves eating red water-melon, apples peeled nicely and bananas.*

amod(water-melon-5, red-4)
*appos(water-melon-5, apples-7)**
partmod(apples-7, peeled-8)
advmod(peeled-8, nicely-9)
cc(nicely-9, and-10)
conj(nicely-9, bananas-11)

Case Study (2/3)

- Conjunctive relations
  - Non-uniform relation generation
    - *George, the president of the football club, James, the coach of the club and the players went to watch their rival's game.*

appos(George-1, president-4)
conj(club-8, James-10)
conj(club-8, coach-13)*
cc(club-8, and-17)
conj(club-8, players-19)

Case Study (3/3)

- Low Precision frequent relations
  - mod and qua
    - More general rules
  - agt-aoj: non-uniformity in corpus
  - Gol-obj: Multiple relation generation possibility

The former, however, represents the toxic concentration in the water, while the latter represents the dosage to the body of the test organisms.

Corpus Relations
agt(represent, latter)
aoj(represent, former)
obj(dosage, body)

Generated Relations
aoj(represent, latter)
aoj(represent, former)
gol(dosage, body)

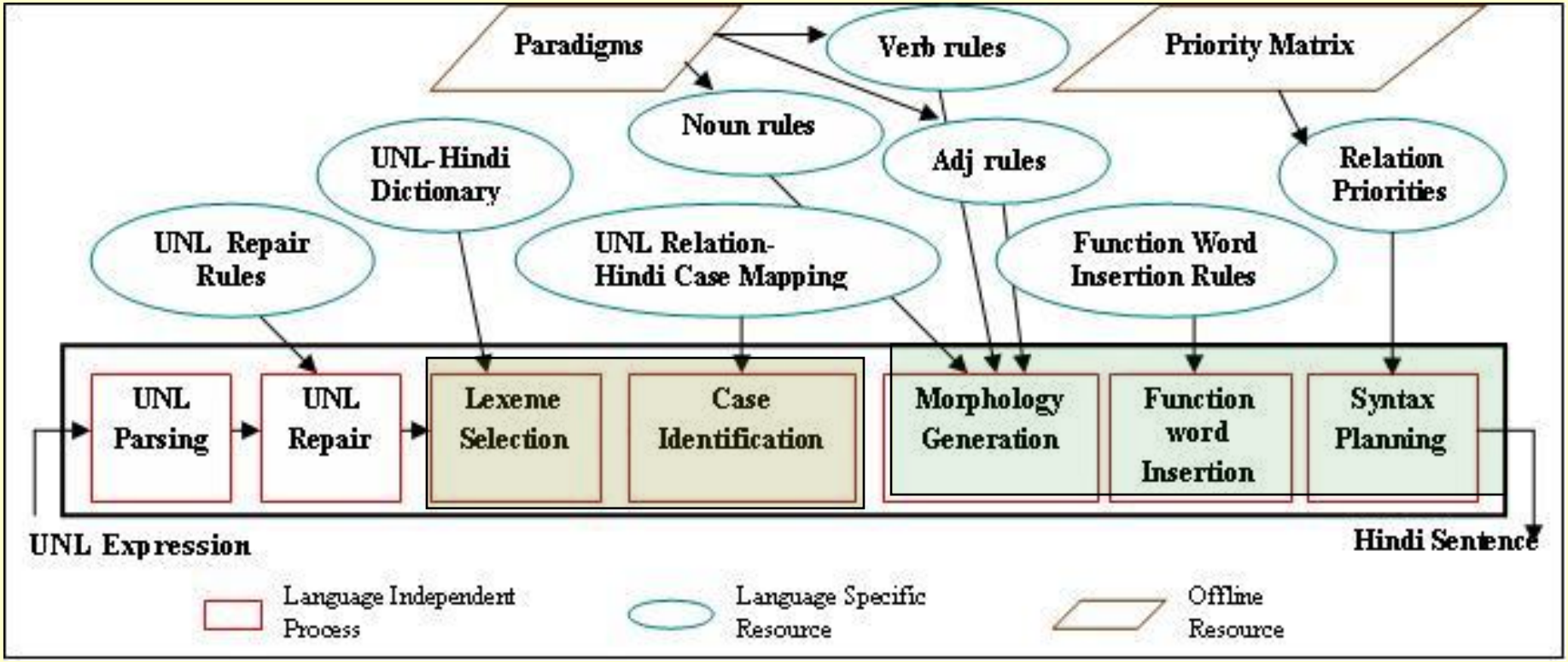
Hindi Generation from Interlingua (UNL)

(Joint work with S. Singh, M. Dalal, V. Vachhani, Om Damani

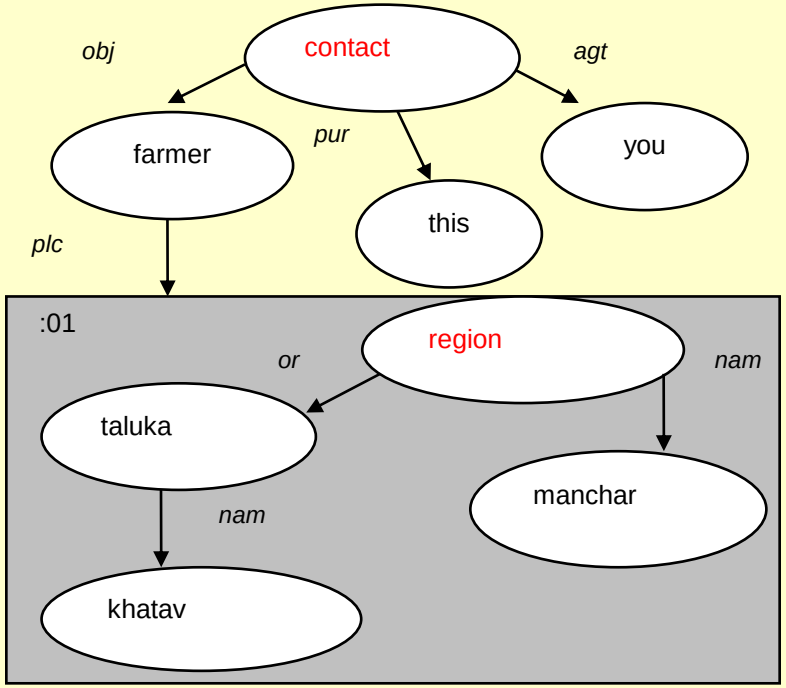
MT Summit 2007)

HinD Architecture

Deconversion = Transfer + Generation



Step-through the Deconverter



Module	Output
UNL Expression	obj(contact(…
Lexeme Selection	संपर्क किसान् यह आप क्षेत्र् तालुक् मंचर खटाव contact farmer this you region taluka manchar khatav
Case Identification	संपर्क किसान्* यह आप क्षेत्र्* तालुक्* मंचर खटाव contact farmer* this you region* taluka* manchar khatav
Morphology Generation	संपर्क कीजिए किसानों इस आप क्षेत्र contact .@imperative farmer.@pl this you region तालुके मंचर खटाव taluka manchar Khatav
Function Word Insertion	संपर्क कीजिए किसानों को इसके लिए आप क्षेत्र contact farmers this for you region या तालुके के मंचर खटाव or taluka of Manchar Khatav
Linearization	इसके लिए आप मंचर क्षेत्र या खटाव This for you manchar region or khatav तालुके के किसानों को संपर्क कीजिए | taluka of farmers contact

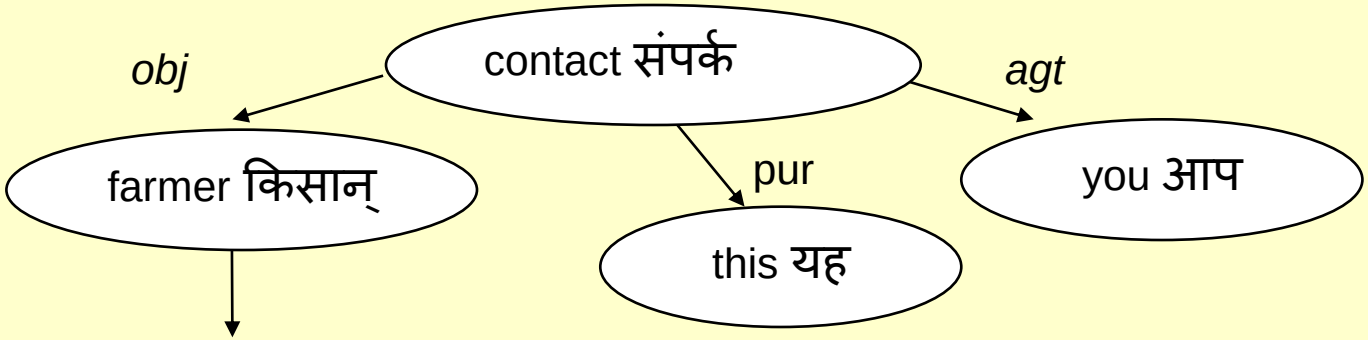
Lexeme Selection

[संपर्क]{}"contact(icl>communicate(agt>person,obj>person))" (V,VOA,VOA-ACT,VOA-COMM,VLTN,TMP,CJNCT,N-V,link,Va)

[पहचान काव्यक्ति]{}"contact(icl>representative)"

(N,ANIMT,FAUNA,MML,PRSN,Na)

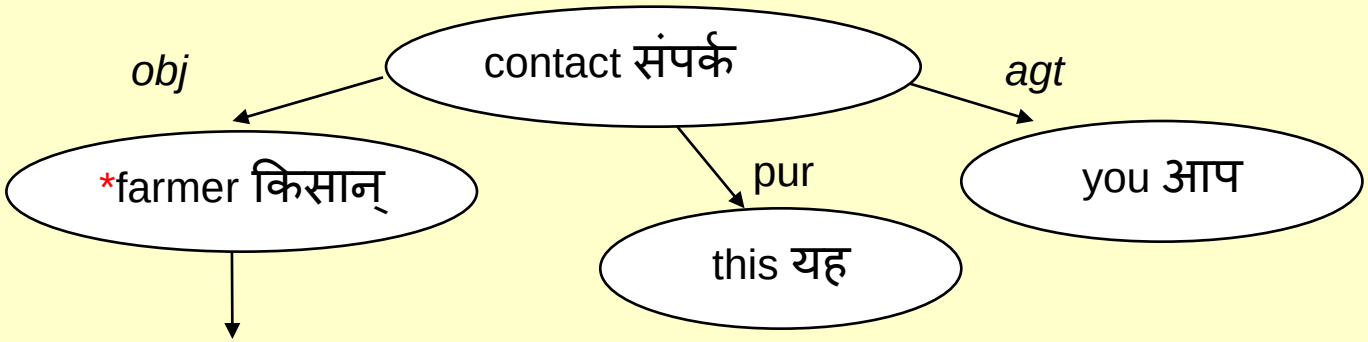
Lexical Choice is unambiguous



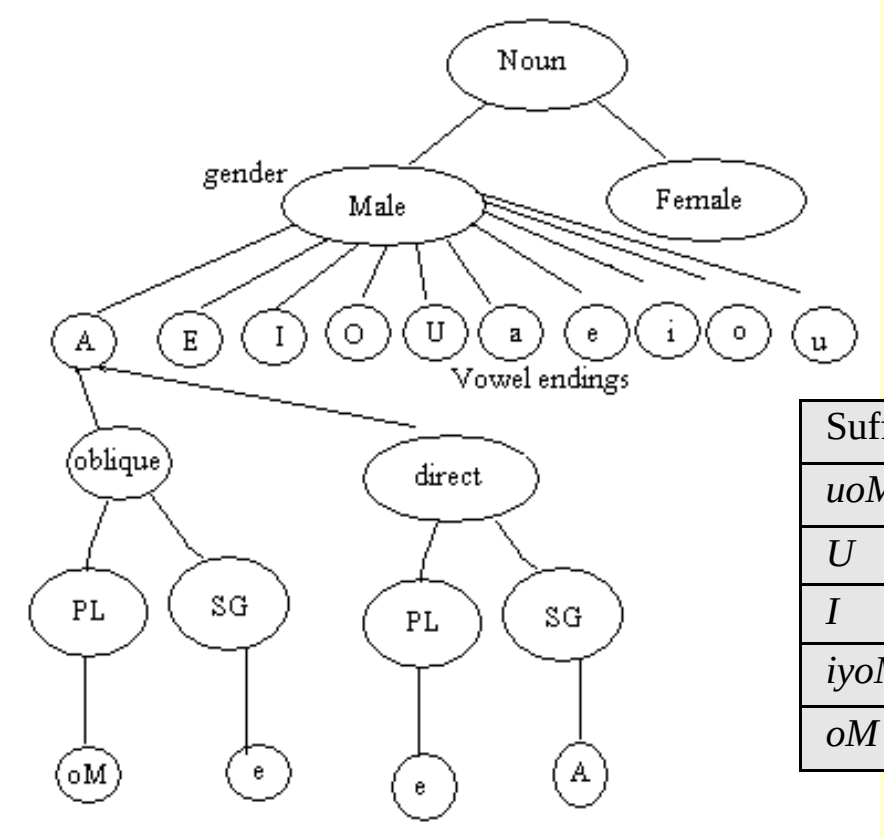
Case Marking

Relation	Parent+	Parent-	Child+	Child-
Obj	V	VINT	N	
Agt	V	@past	N	

- Depends on UNL Relation and the properties of the nodes
- Case get transferred from head to modifiers



Morphology Generation: Nouns



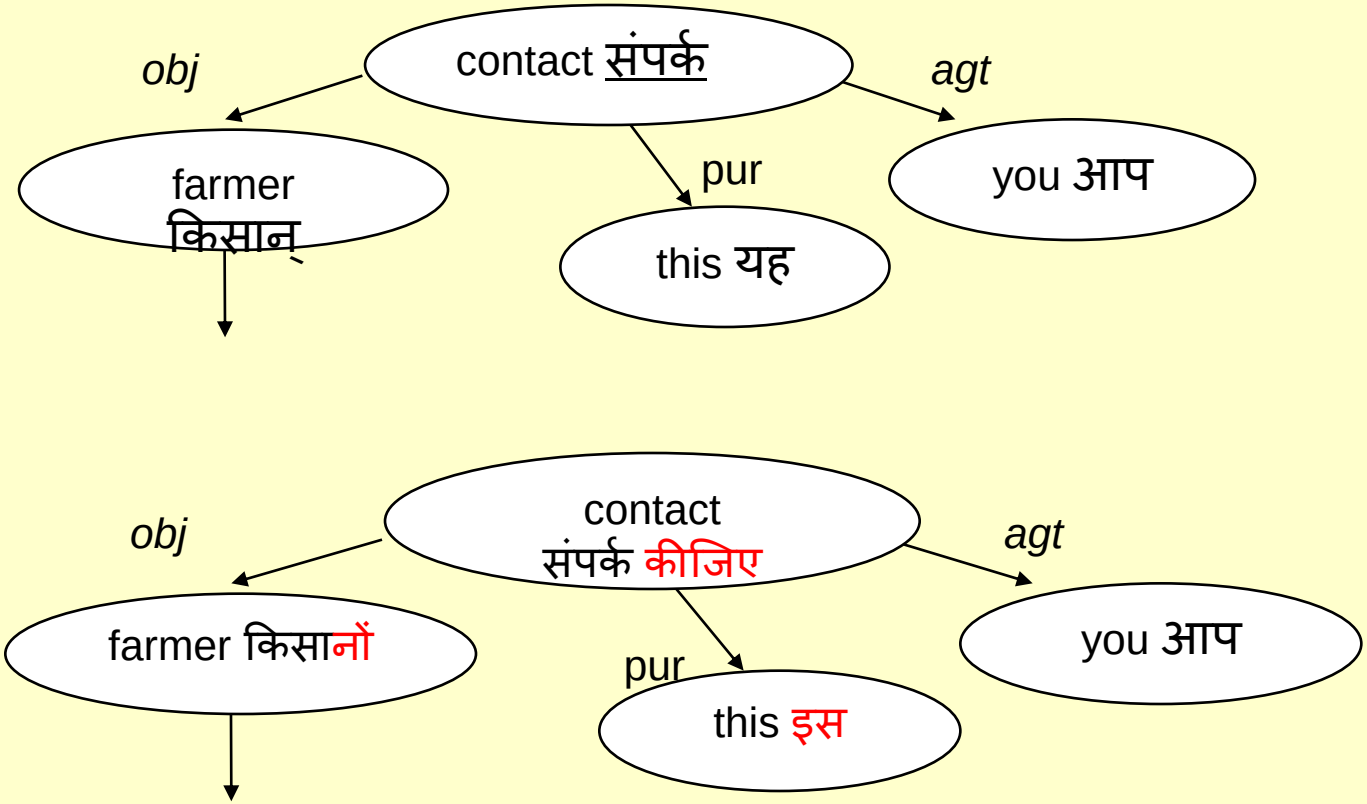
The boy saw me.
लड़के ने मुझे देखा ।
Boys saw me.
लड़कों ने मुझे देखा ।
The King saw me.
राजा ने मुझे देखा ।
Kings saw me.
राजाओं ने मुझे देखा ।

Suffix	Attribute values
uoM	*@N,@NU,@M,@pl,@oblique*
U	*@N,@NU,@M,@sg,@oblique*
I	*@N,@NI,@F,@sg,@oblique*
iyoM	*@N,@NI,@F,@pl,@oblique*
oM	*@N,@NA,@NOTCH,@F,@pl,@oblique*

Verb Morphology

Suffix	*Tense*	*Aspect*	*Mood*	*N*	*Gen*	*P*	*V E*
-e rahaa thaa	@past	@progress	-	@sg	@male	3rd	e
-taa hai	@present	@custom	-	@sg	@male	3rd	-
-iyaa thaa	@past	@complete	-	@sg	@male	3rd	I
saktii hain	@present	-	@ability	@pl	@female	3rd	A

After Morphology Generation

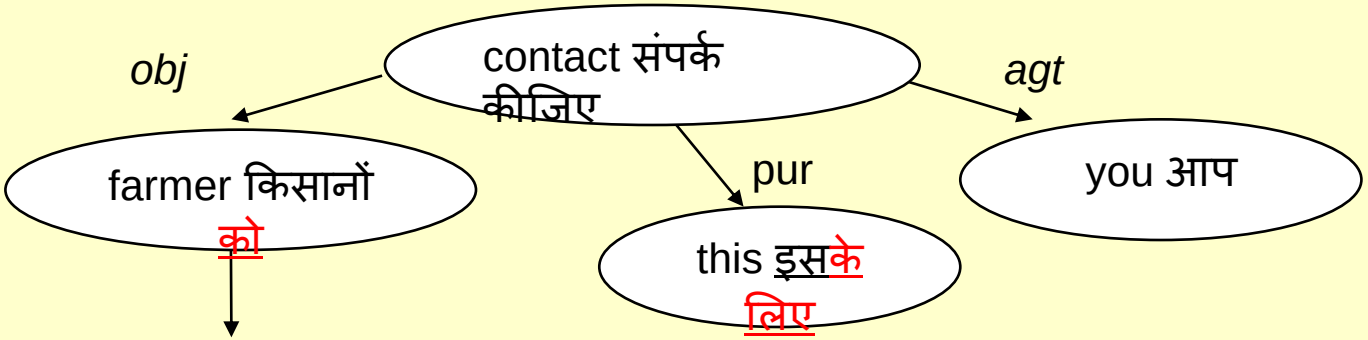


Function Word Insertion

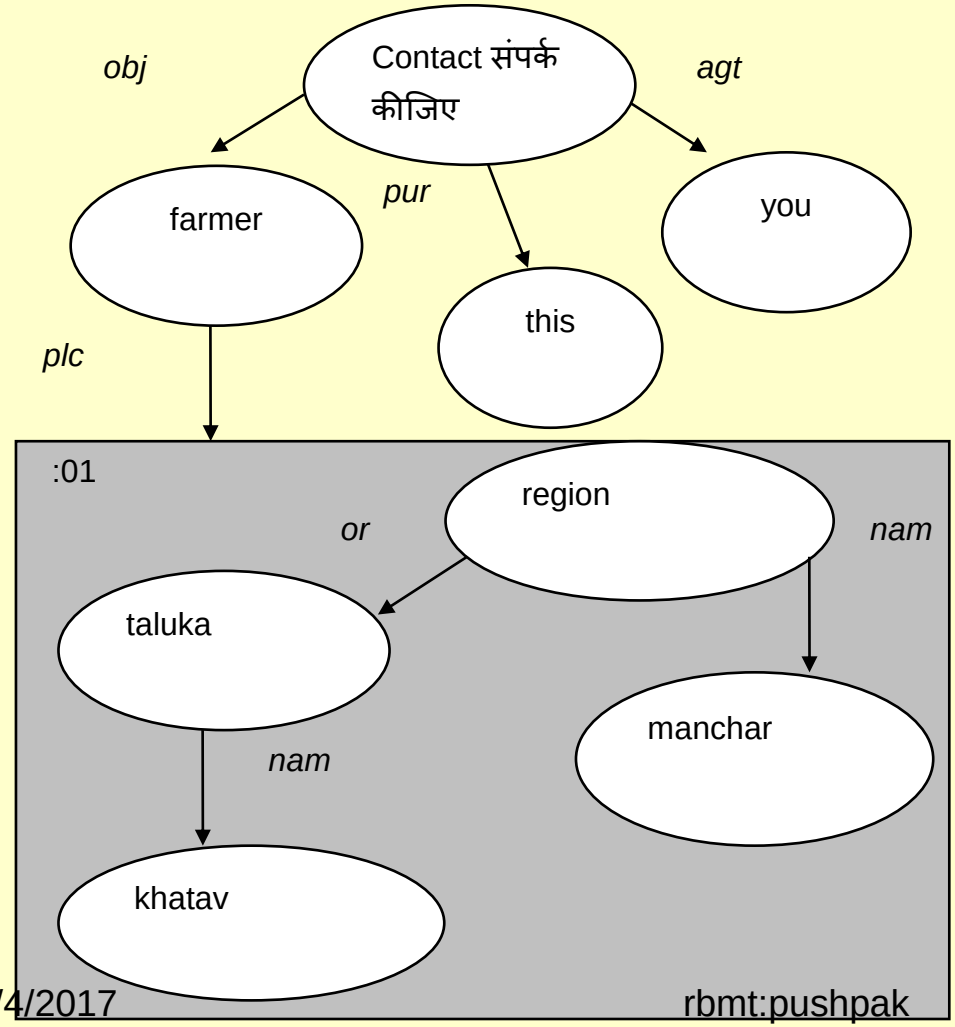
संपर्क कीजिए किसानों यह आप क्षेत्र तालुके मंचर खटाव

संपर्क कीजिए किसानों को इसके लिए आप क्षेत्र या तालुके के मंचर खटाव

Rel	Par+	Par-	Chi+	Chi-	Ch/FW
Obj	V	VINT	N#ANIMT	@topic	को



Linearization



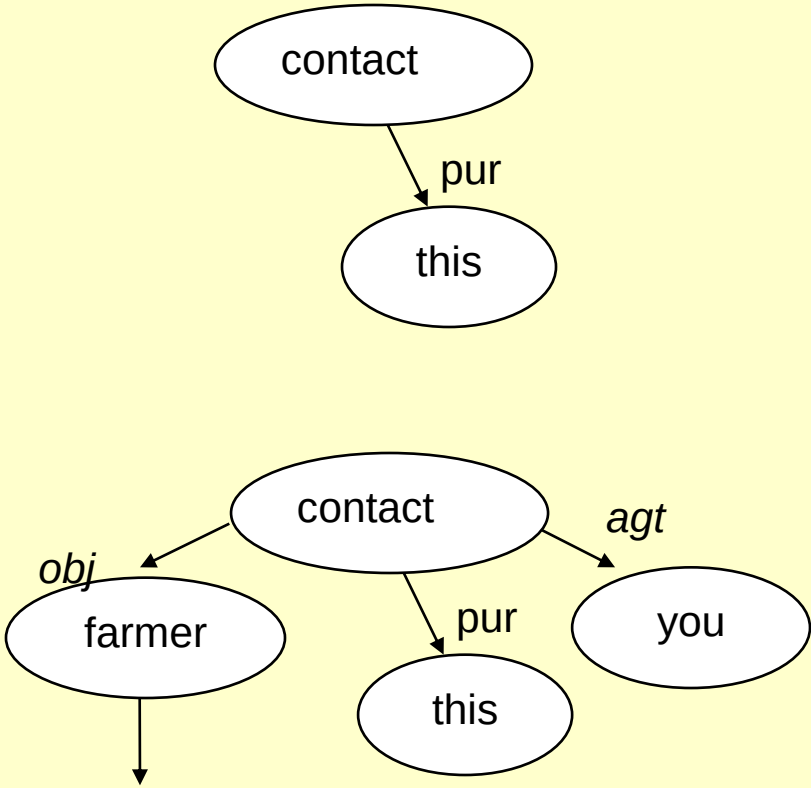
इस आप मंचर क्षेत्र
This you manchar region

खटाव तालुके किसानों
khatav taluka farmers

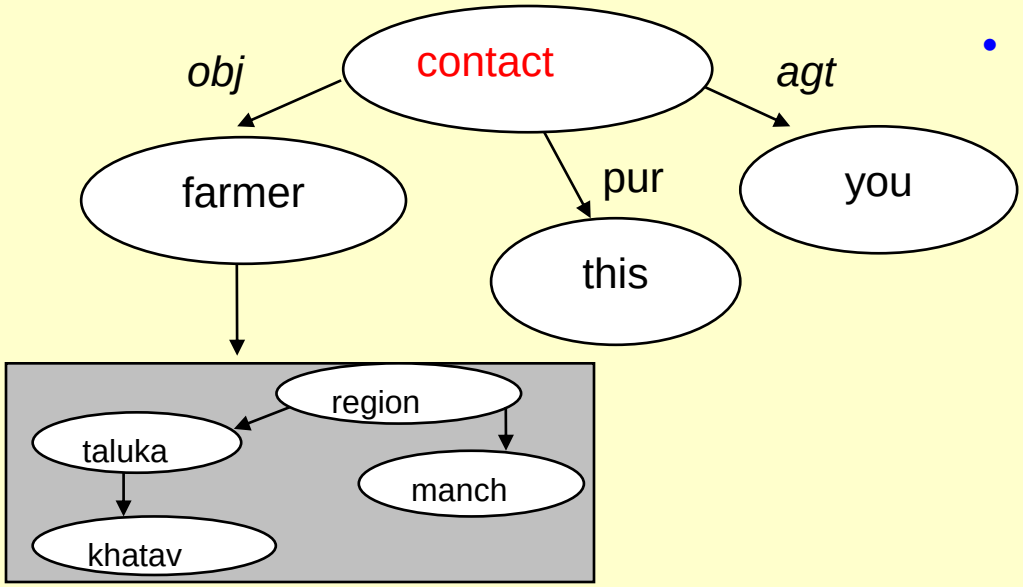
संपर्क
contact

Syntax Planning: Assumptions

- The relative word order of a UNL relation's relata does not depend on:
  - Semantic Independence: the semantic properties of the relata.
  - Context Independence: the rest of the expression.
- The relative word order of various relations sharing a relatum does not depend on
  - Local Ordering: the rest of the expression.

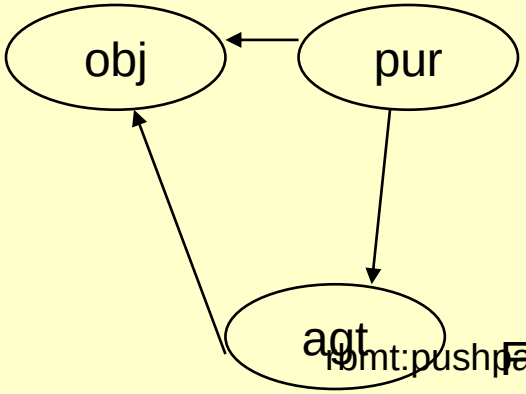


Syntax Planning: Strategy



- Divide a nodes relations in
  Before_set = {obj,pur,agt}
  After_set = {}

	agt	aoj	obj	ins
Agt		**L**	**L**	**L**
Aoj			**L**	**L**
Obj				**R**
Ins				

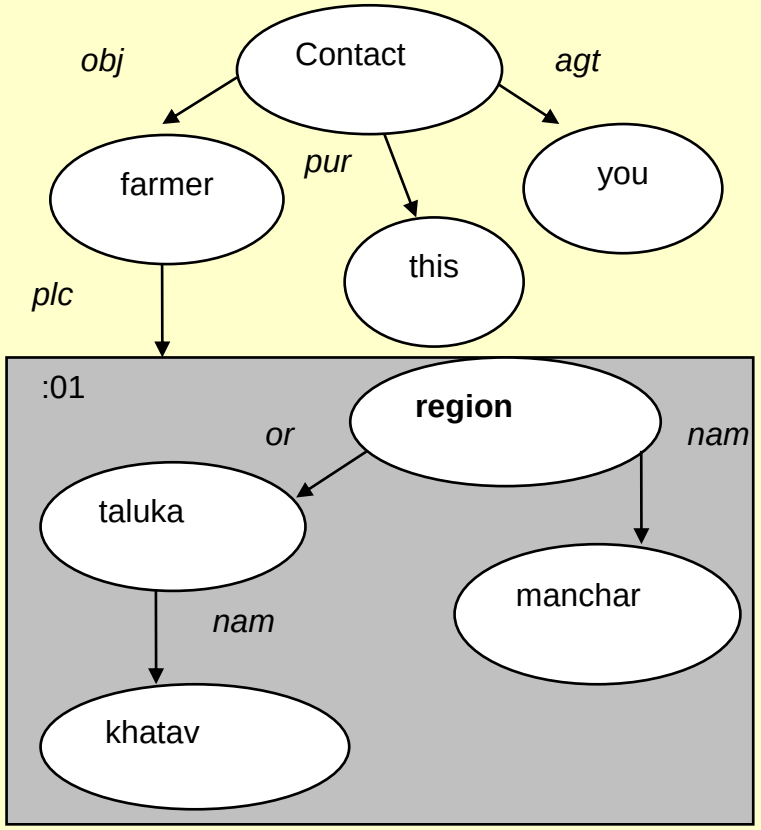


Topo Sort each group:
pur agt obj
this you farmer

Final order: this you farmer contact

Syntax Planning Algo

Stack	Before Current	After Current	Output
region farmer contact			this you
	manchar	taluka	
manchar **region** taluka farmer contact			
region taluka farmer Contact			this you manchar
taluka farmer contact			this you manchar **region**



All Together (UNL -> Hindi)

Module	Output
UNL Expression	obj(contact(..
Lexeme Selection	संपर्क किसान् यह आप क्षेत्र् तालुक् मंचर खटाव contact farmer this you region taluka manchar khatav
Case Identification	संपर्क किसान्* यह आप क्षेत्र्* तालुक्* मंचर खटाव contact farmer* this you region* taluka* manchar khatav
Morphology Generation	संपर्क कीजिए किसानों यह आप क्षेत्र contact .@imperative farmer.@pl this you region तालुके मंचर खटाव taluka manchar Khatav
Function Word Insertion	संपर्क कीजिए किसानों को इसके लिए आप क्षेत्र contact farmers this for you region या तालुके के मंचर खटाव or taluka of Manchar Khatav
Syntax Linearization	इसके लिए आप मंचर क्षेत्र या खटाव This for you manchar region or khatav तालुके के किसानों को संपर्क कीजिए । taluka of farmers to contact

How to Evaluate UNL Deconversion

- UNL -> Hindi
- Reference Translation needs to be generated from UNL
  - Needs expertise
- Compromise: Generate reference translation from original English sentence from which UNL was generated
  - Works if you assume that UNL generation was perfect
- Note that fidelity is not an issue

Manual Evaluation Guidelines

Fluency of the given translation is:
(4) Perfect: Good grammar
(3) Fair: Easy-to-understand but flawed grammar
(2) Acceptable: Broken - understandable with effort
(1) Nonsense: Incomprehensible

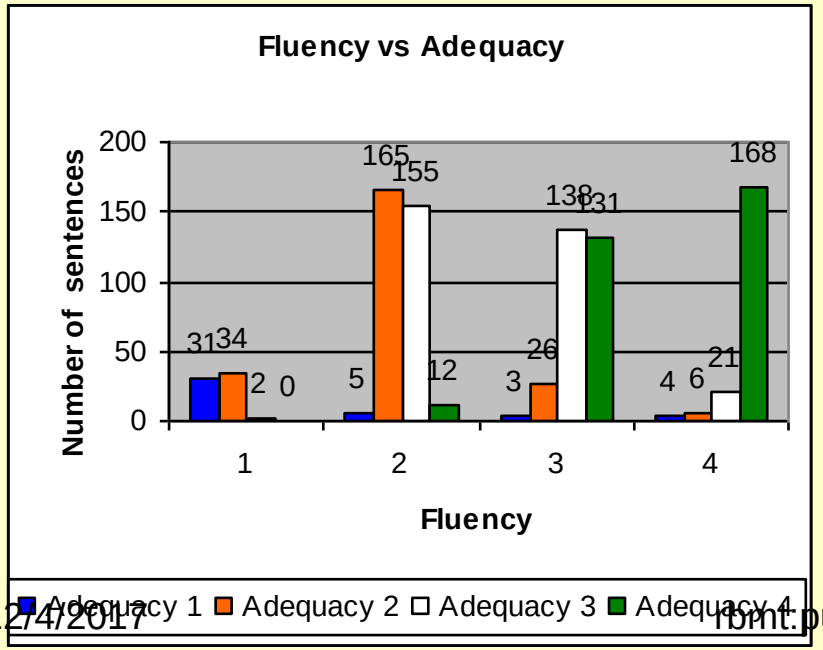
Adequacy: How much meaning of the reference sentence is conveyed in the translation:
(4) All: No loss of meaning
(3) Most: Most of the meaning is conveyed
(2) Some: Some of the meaning is conveyed
(1) None: Hardly any meaning is conveyed

Sample Translations

Hindi Output	Fluency	Adequacy
कवक के कारण आम की नाजुक पत्तियां यदि झुलसे रहे हैं 0.5 प्रतिशत का बोर्डो मिश्रण 10 लीटर पानी के साथ तो छिड़का जाना चाहिए	2	3
परीक्षण के अनुरुप नियमित रूप से फलों की अच्छी वृद्धि के लिए खादों की खुराकें दी जानी चाहिए	3	4
हमें मेथी का फसल के बाद बोता है मेथी और धनिया फसल को अच्छ् बढ़ने या अच्छ् बढ़ने बता	1	1
जीवाण्विक संक्रमण से इसकी जड़ें प्रभावित होती हैं	4	4
इमु का पक्षी रेटाइट का परिवार को संबंधित होता है और थोड़ा यह शुतुरमुर्ग के साथ समान दिखती हैं	2	3

Results

	BLEU	Fluency	Adequacy
Geometric Average	0.34	2.54	2.84
Arithmetic Average	0.41	2.71	3.00
Standard Deviation	0.25	0.89	0.89
Pearson Cor. *BLEU*	1.00	0.59	0.50
Pearson Cor. Fluency	0.59	1.00	0.68



- Good Correlation between Fluency and BLUE
- Strong Correlation between Fluency and Adequacy
- Can do large scale evaluation using Fluency alone

Caution: Domain diversity,
Speaker diversity

Summary: interlingua based MT

- English to Hindi
- Rule governed
- High level of linguistic expertise needed
- Takes a long time to build (since 1996)
- But produces great insight, resources and tools

Future work

- Machine Learning of UNL from text
  - Relations
  - attributes
- Naturalness of NLG from UNL
- Register
- Combine RBMT with SMT and NMT

Thank you

Underlying formlisms

- The knowledge representation system FrameKit
- A language for representing domain models (a semantic extension of FRAMEKIT)
- Specialized grammar formalisms, based on Lexical-Functional Grammar
- A specially constructed language for representing text meanings (the interlingua)
- The languages of analysis and generation lexicon entries, and of the structural mapping rules

Procedural components

- A syntactic parser with a semantic constraint interpreter
- A semantic mapper for treating additional types of semantic constraints
- An interactive augmentor for treating residual ambiguities;