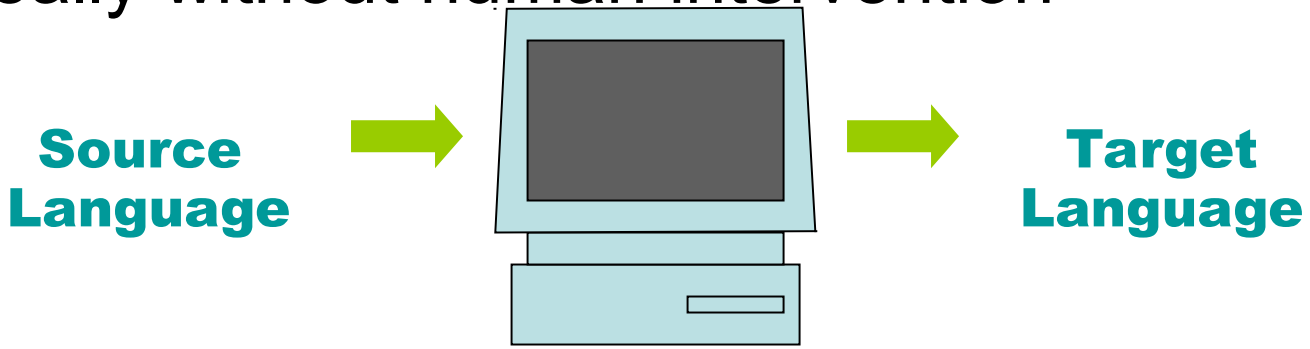
Machine Translation:
Interlingual Methods

Bonnie J. Dorr, Eduard H. Hovy, Lori S. Levin

Thanks to Les Sikos

Overview

- What is Machine Translation (MT)?
  - Automated system
  - Analyzes text from Source Language (SL)
  - Produces “equivalent” text in Target Language (TL)
  - Ideally without human intervention



Overview

- Three main methodologies for Machine Translation
  - Direct
  - Transfer
  - Interlingual

Overview

- Three main methodologies for Machine Translation
  - Direct
  - Transfer
  - Interlingual

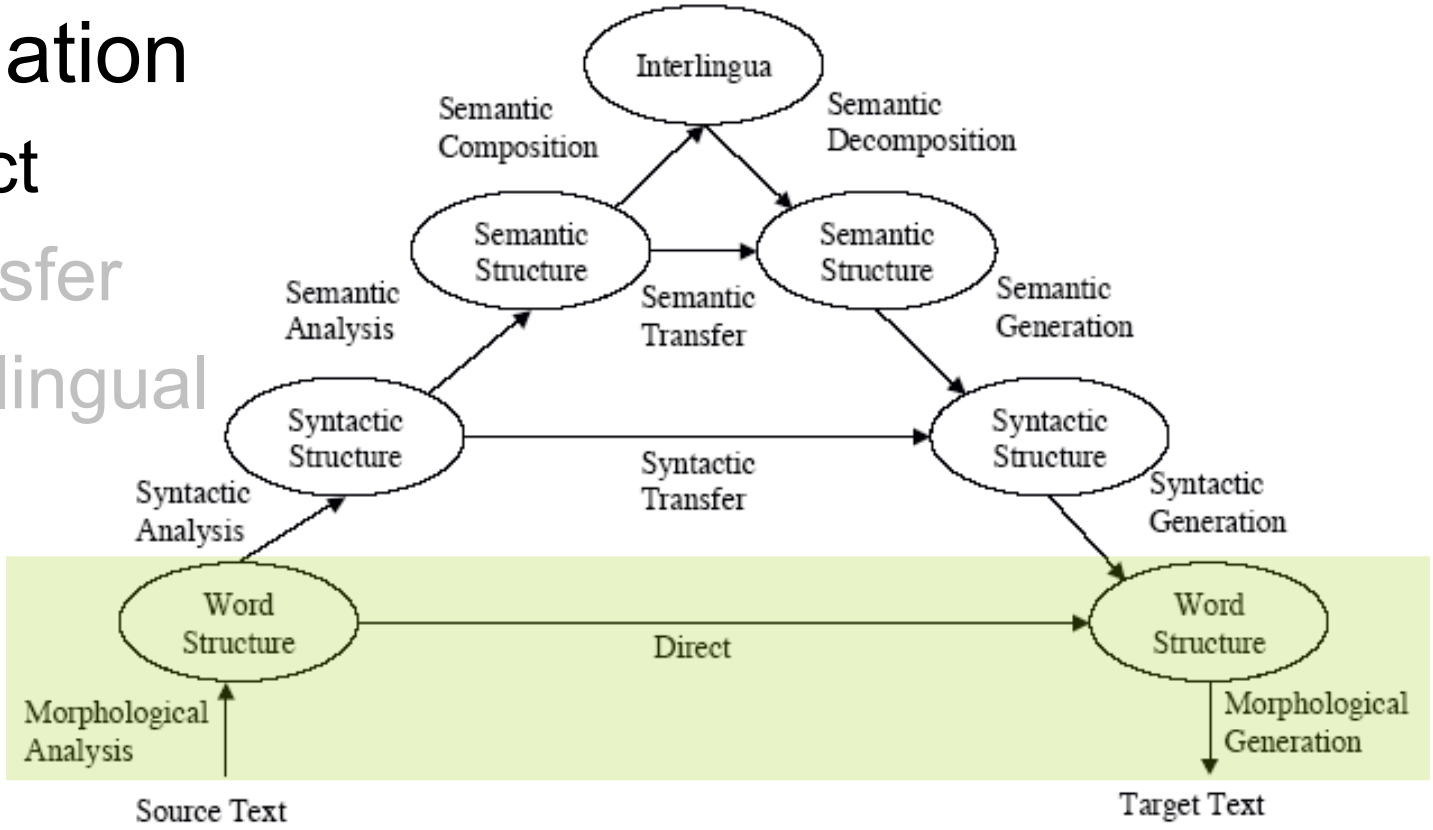


Figure 1: The Vauquois Triangle for MT

Overview

- Three main methodologies for Machine Translation
  - Direct
  - Transfer
  - Interlingual

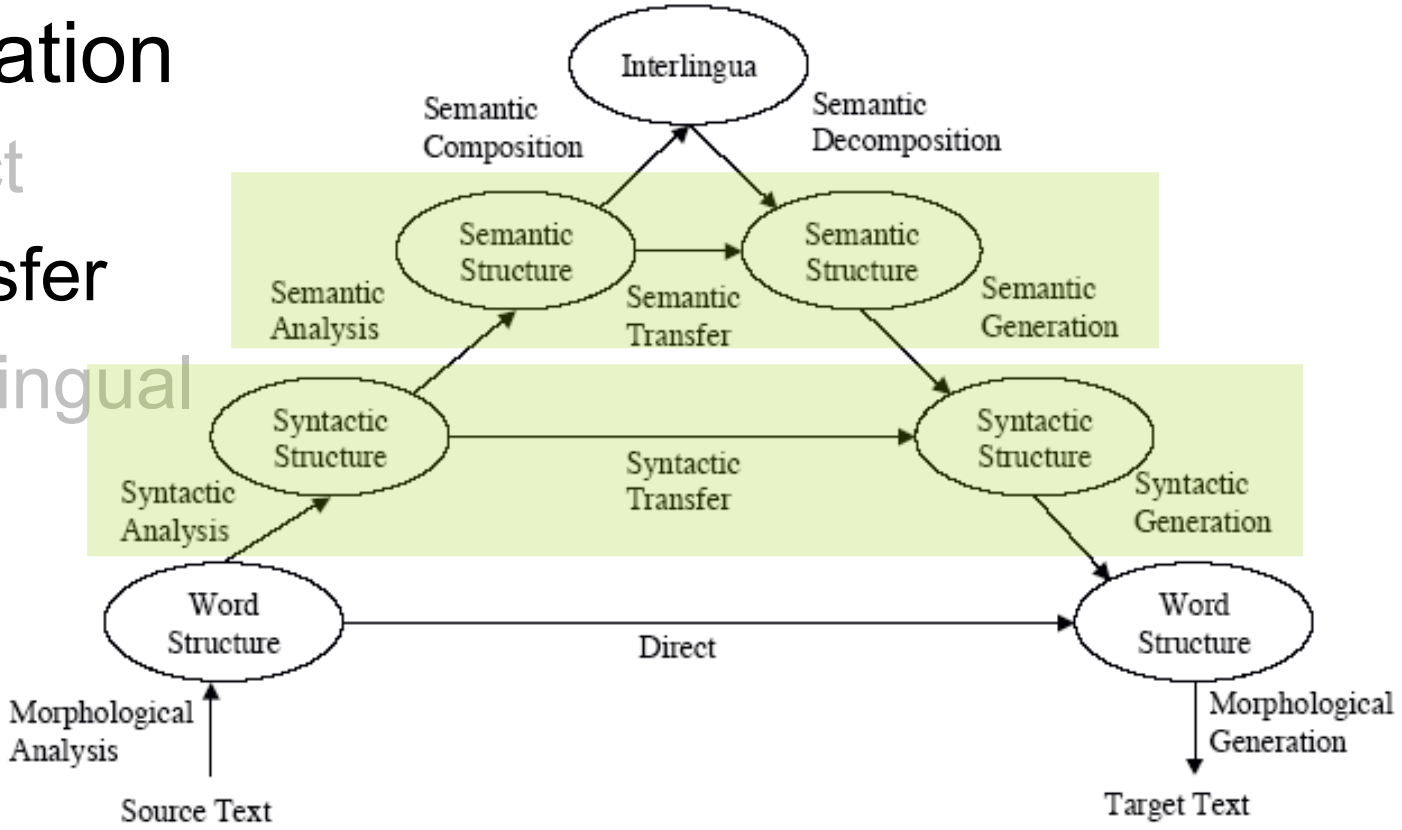


Figure 1: The Vauquois Triangle for MT

Overview

- Three main methodologies for Machine Translation
  - Direct
  - Transfer
  - Interlingual

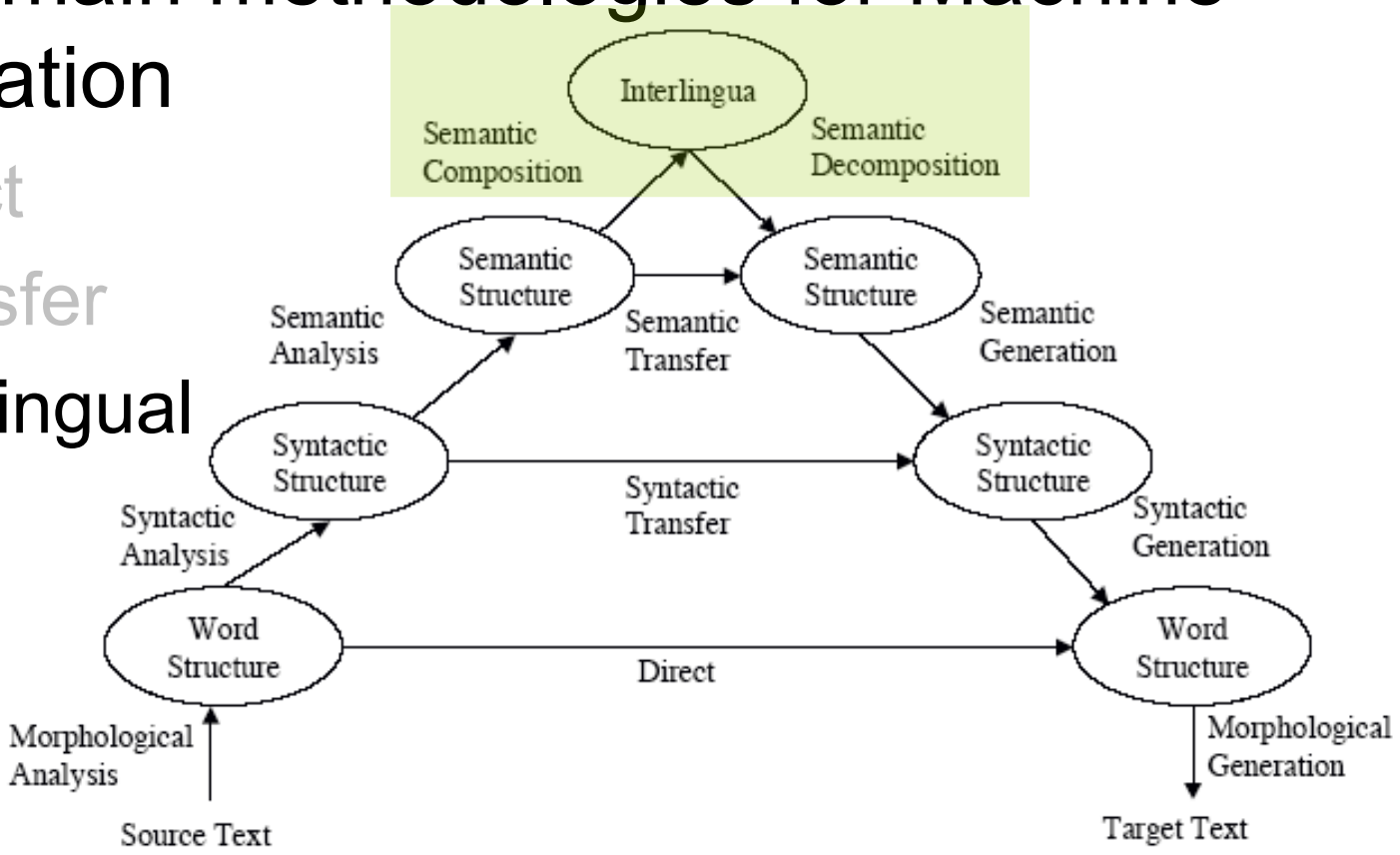


Figure 1: The Vauquois Triangle for MT

Overview

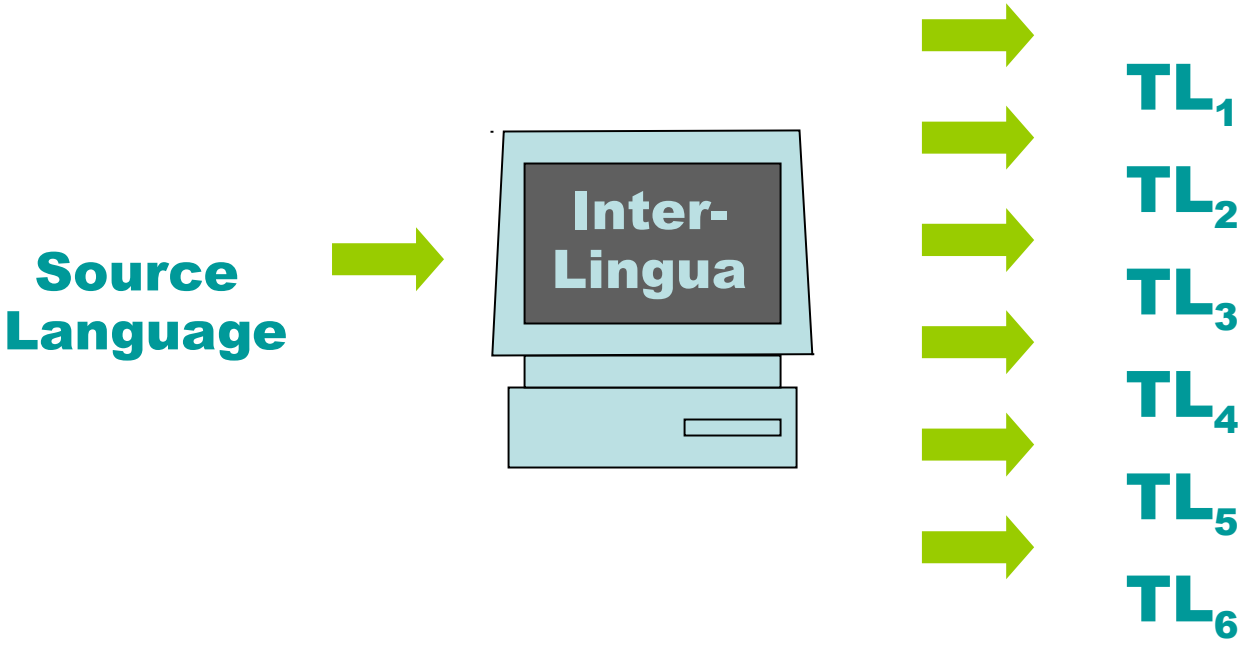
- Interlingua
  - Single underlying representation for both SL and TL
    which ideally
    - Abstracts away from language-specific characteristics
    - Creates a “language-neutral” representation
    - Can be used as a “pivot” representation in the translation

Overview

- Cost/Benefit analysis of moving up the triangle
  - Benefit
    - Reduces the amount of work required to traverse the gap between languages
  - Cost
    - Increases amount of **analysis**
      - Convert the source input into a suitable pre-transfer representation
    - Increases amount of **synthesis**
      - Convert the post-transfer representation into the final target surface form

Overview

- Two major advantages of Interlingua method
  1. The more target languages there are, the more valuable an Interlingua becomes



Overview

- Two major advantages of Interlingua method
  2. Interlingual representations can also be used by NLP systems for other multilingual applications

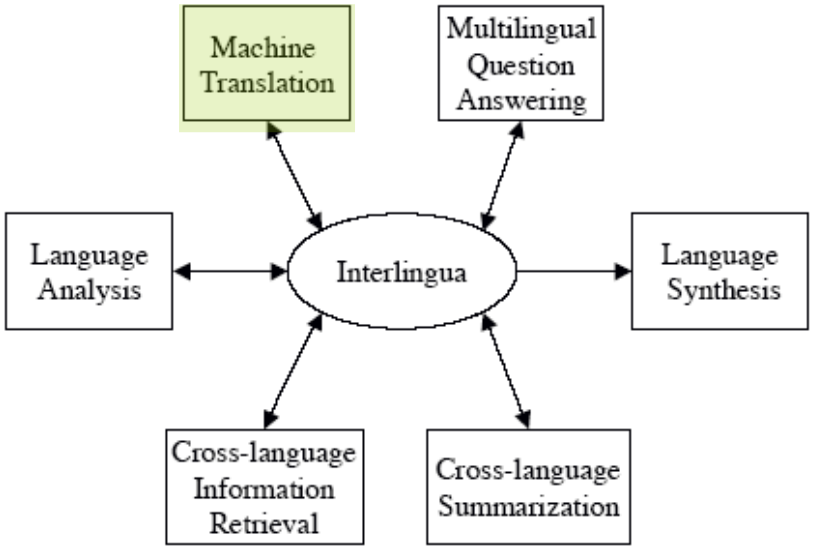


Figure 2: Use of Interlingua in Multiple Applications

Overview

- Sounds great, but…due to many complexities
  - Only one interlingual MT system has ever been made operational in a commercial setting
    - KANT (Nyberg and Mitamura, 1992, 2000; Lonsdale *et al*., 1995)
  - Only a few have been taken beyond research prototype

Current Efforts

- KANT system (Nyberg and Mitamura, 1992)
  - Only interlingual MT system that has ever been made operational in a commercial setting
    - Caterpillar document workflow (mid-90s)
  - Knowledge-based system
  - Designed for translation of technical documents written in Caterpillar Technical English (CTE) to French, Spanish, and German
  - Controlled English – no pronouns, conjunctions,...

Issues

- Loss of Stylistic Elements
  - Because representation is independent of syntax
    - Generated target text reads more like a paraphrase
    - Style and emphasis of the original text are lost
  - Not so much a failure of Interlingua as incompleteness
    - Caused by a lack of understanding of discourse and pragmatic elements required to recognize and appropriately reproduce style and emphasis
    - In some cases it may be an advantage to ignore the author's style
      - Outside the field of artistic texts (poetry and fiction) syntactic form of source text is superfluous

Issues

- Loss of Stylistic Elements
  - Current state of the art
    - It is only possible to produce reliable interlinguas between language groups (e.g., Japanese – Western European) within specialized domains

Issues

- Linguistic Divergences
  - Structural differences between languages
    - Categorical Divergence
      - Translation of words in one language into words that have ***different parts of speech*** in another language
        - *To be jealous*
        - *Tener celos (To have jealousy)*

Issues

- Linguistic Divergences
  - Conflational Divergence
    - Translation of ***two or more words*** in one language ***into one word*** in another language
      - » *To kick*
      - » *Dar una patada (Give a kick)*

Issues

- Linguistic Divergences
  - Structural Divergence
    - Realization of verb arguments in ***different syntactic configurations*** in different languages
      - » *To enter the house*
      - » *Entrar en la casa (Enter in the house)*

Issues

- Linguistic Divergences
  - Head-Swapping Divergence
    - Inversion of a ***structural-dominance*** relation between two semantically equivalent words
      - » *To run in*
      - » *Entrar corriendo (Enter running)*

Issues

- Linguistic Divergences
  - Thematic Divergence
    - Realization of verb arguments that reflect ***different*** thematic to syntactic ***mapping*** orders
      - » *I like grapes*
      - » *Me gustan uvas (To-me please grapes)*

Issues

- Linguistic Divergences may be the norm rather than
  the exception
  - Differences in MT architecture (direct, transfer, interlingual) are crucial for resolution of
    cross-language divergences
    - Interlingua approach takes advantage of the compositionality of basic units of meaning to resolve divergences

Issues

- For example:

  To kick – Dar una patada (Give a kick)

  – Conflational divergence can be resolved by mapping English *kick* into two components before translating
  into in Spanish
    - Motional component (movement of the leg)
    - Manner component (a kicking motion)

Current Efforts

- Pangloss project (Frederking et al., 1994)
  - Ambitious attempt to build rich interlingual expressions
  - Uses humans to augment system analysis
  - Representation includes a set of frames for representing semantic components, each of which
    - Are headed by a unique identifier
    - And have a separate frame with aspectual information (duration, telicity, etc.)
  - Some modifiers are treated as scalars and represented by numerical values

Current Efforts

- Mikrokosmos (Mahesh and Nirenburg, 1995) / OntoSem (Nirenburg and Raskin, 2004)
  - Focus is to produce semantically rich Text-Meaning Representations (TMRs) of text
  - TMRs use a language-independent metalanguage also used for static knowledge resources
  - TMRs aimed at the most difficult problems of NLP; Disambiguation, reference resolution
  - Goal is to populate a fact repository with TMRs as a language-independent search space for question-answering and knowledge-extraction

Current Efforts

- PRINCITRAN (Dorr & Voss, 1996)
  - Approach assumes an interlingua derived from lexical semantics and predicate decomposition
    - Jackendoff 1983, 1990; Levin & Rappaport-Hovav 1995a, 1995b
  - Has not complicated, but rather facilitated, the identification and construction of systematic relations at the interface between each level

Current Efforts

- Motivation for Non-Uniform Approach

  German: *Der Berg liegt im Suden der Stadt*

  – Ambiguous in English:
    - *The mountain lies **in** the south of the city*
    - *The mountain lies **to** the south of the city*
  – In other words, the German phrase maps to two
    distinct representations

Current Efforts

- Using Default knowledge in the KR
  - *Mountains are physical entities, typically distinct and external to cities*
  - System chooses second translation
    - ➔ *The mountain lies **to** the south of the city*
- Using specific facts in the KR
  - *A particular mountain is in the city*
  - System overrides default knowledge and chooses first translation
    - ➔ *The mountain lies **in** the south of the city*

Current Efforts

- The need to translate such sentences accurately is a clear case of where general as well as specific real-world knowledge should assist in eliminating inappropriate translations
  - Knowledge Representational level, not the Interlingual level, provides this capability in this model

Current Efforts

- Lexical Conceptual Structure (LCS)
  - Used as part of many MT language pairs including ChinMT (Habash et al., 2003a)
    - Chinese-English
  - Also been used for other natural language applications
    - Cross-language information retrieval

Current Efforts

- Lexical Conceptual Structure (LCS)
  - Approach focuses on linguistic divergences
  - For example – Conflational divergence

    Arabic: *The reporter caused the email to go to Al-Jazeera in a sending manner.*

    English: *The reporter emailed Al-Jazeera.*

Current Efforts

- LCS representation

```
(event cause
        (thing[agent] reporter+)
        (go loc
                (thing[theme] email+)
                (path to loc
                        (thing email+)
                        (position at loc (thing email+) (thing[goal] aljazeera+)))
                (manner send+ingly)))
```

Current Efforts

- LCS representation

```
(event cause
      (thing[agent] reporter+)
      (go loc
            (thing[theme] email+)
            (path to loc
                  (thing email+)
                  (position at loc (thing email+) (thing[goal]
aljazeera+)))
            (manner send+ingly)))
```

  - Primary components of meaning are the top-level conceptual nodes *cause* and *go*

Current Efforts

- LCS representation

(event cause
(thing[*agent*] reporter+)
(go loc
(thing[*theme*] email+)
(path to loc
(thing email+)
(position at loc (thing email+) (thing[*goal*] aljazeera+)))
(manner send+ingly)))

  - Primary components of meaning are the top-level conceptual nodes *cause* and *go*
  - These are taken together with their arguments
    - Each identified by a semantic role (*agent, theme, goal*)
  - And a modifier (manner) *send+ingly*

LCS as an interlingua?

- Jackendoff wasn't trying to capture all of meaning – just the semantics that corresponds to syntactic generalizations
- Ch-of-loc, causation, states, ... are very fundamental. If we don't get anything else, we should get at least these
- LCS highlights just these relations – not bad for an interlingua, but what about those stylistic things, etc?

Current Efforts

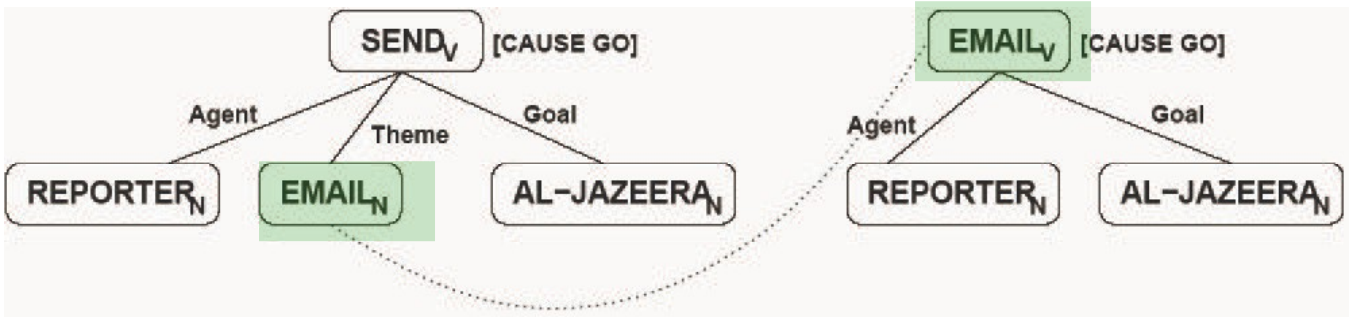
- Approximate Interlingua (Dorr and Habash, 2002)
  - Depth of knowledge-based systems is approximated
  - Taps into the richness of resources in one language (often English)
  - This information is used to map the source-language input to the target-language output

Current Efforts

- Approximate Interlingua (Dorr and Habash, 2002)
  - Focus on linguistic divergences but with fewer knowledge-intensive components than in LCS
  - Key feature
    - Coupling of basic argument-structure information with some, but not all, components the LCS representation
    - Only the top-level primitives and semantic roles are retained
  - This new representation provides the basis for generation of multiple sentences that are statistically pared down – ranked by TL constraints

Current Efforts

- Approximate Interlingua representation:



  - Check top-level conceptual nodes for matches
  - Check unmatched thematic roles for 'conflatability'
    - Cases where semantic roles are absorbed into other predicate positions
  - Here there is a relation between the conflated argument EMAIL_N and EMAIL_V