Semantic Annotation and Lexico-Syntactic Paraphrase

Bonnie J. Dorr†, Rebecca Green†, Lori Levin‡, Owen Rambow∗, David Farwell§, Nizar Habash†∗, Stephen Helmreich§, Eduard Hovy¶, Keith J. Miller^, Teruko Mitamura‡, Florence Reeder^, Advaith Siddharthan∗

†Institute for Advanced Computer Studies
University of Maryland
{bonnie, rgreen, habash}@umiacs.umd.edu

‡Language Technologies Institute
Carnegie Mellon University
{lsl, teruko}@cs.cmu.edu

∗Department of Computer Science
Columbia University
{rambow, as372}@cs.columbia.edu

§Computing Research Laboratory
New Mexico State University
{david, shelmrei}@crl.nmsu.edu

¶Information Sciences Institute
University of Southern California
hovy@isi.edu

^MITRE Corporation
{freeder, keith}@mitre.org

Abstract

The IAMTC project (Interlingual Annotation of Multilingual Translation Corpora) is developing an interlingual representation framework for annotation of parallel corpora (English paired with Arabic, French, Hindi, Japanese, Korean, and Spanish) with deep-semantic representations. In particular, we are investigating meaning equivalent paraphrases involving conversives and non-literal language use, as well as extended paraphrases involving syntax, lexicon, and grammatical features. The interlingua representation has three levels of depth. Each level is characterized by the types of meaning equivalent paraphrases that receive identical representations at that level.

1. Introduction

An important issue for computational linguists and lexicographers is the question of meaning-equivalent paraphrases, including lexical synonymy, conversives (*buy*/*sell*), idioms (*kick the bucket*/*die*), and more extended paraphrases, such as *Its network of eighteen organizations has lent a billion dollars to microenterprises* and *The network comprises eighteen organizations which have disbursed a billion dollars to microenterprises.*

Semantic annotation projects such as PropBank (Kingsbury and Palmer, 2002) and FrameNet (Baker et al., 1998) do not cover extended paraphrases. (Barzilay and McKeown, 2001) have proposed automatic methods for extraction of extended paraphrases. However such methods are subject to the usual pitfalls of machine learning systems, requiring large amounts of training data and having imperfect precision and recall.

Our approach in the IAMTC project (Interlingual Annotation of Multilingual Translation Corpora) is complementary to other semantic annotation projects and to projects that automatically label semantic paraphrases. Firstly, we are annotating texts in seven languages (Arabic, English, French, Hindi, Japanese, Korean, and Spanish). Secondly, we are investigating meaning-equivalent paraphrases by annotating multiple versions of the same text, usually one non-English source language and two independently produced English translations. The annotation scheme includes three levels of depth. Each level is characterized by the types of paraphrase that are resolved at that level.

2. Project Overview

The IAMTC project has four goals — development of an interlingual representation framework, annotation of bilingual corpora, development of semantic annotation tools, and design of evaluation metrics for assessing interlingual representations.

The methodology for developing the interlingual representation involves careful study of text corpora in seven languages. Each corpus is bilingual and multi-parallel. Each text in a corpus has at least three versions, a non-English original and at least two English translations. The comparison of annotated multi-parallel corpora sets us apart from other semantic annotation projects. The multiple parallel texts allow us to document naturally occurring paraphrases of the same meaning. The interlingua framework will include a formal definition of the three levels of representation, characterization of paraphrases that are resolved at each level, and coding manuals for each level.

The corpus annotation effort is aimed at making a publically available semantically annotated corpus that can be useful for any natural language processing application that seeks to include more semantic depth. We expect the corpus to be useful for improvement of machine translation, summarization, and information extraction.

Our efforts in tool development have led us not only to create tools for annotation, but also to build interfaces for comparing and reconciling annotations within and across annotators. Because our annotation tools facilitate the building of trees with feature structures at each node, we expect that the tools will be useful for other types of linguistic annotation beyond the IAMTC project.

Finally, an important part of the IAMTC project is to design and test a variety of new evaluation metrics for assessing the interlingual representations and choosing an appropriate granularity of meaning representation. Evaluation is generally important in language technologies, but is particularly important in an area like interlingua design that deals with meanings that are deep, multi-faceted, and not well-defined. Our evaluation metrics include inter-annotator agreement, intra-annotator consistency, and success in NLP applications.

3. The Interlingua Representation

Recognizing the complexity of interlinguas, we adopt an incrementally deepening approach, which allows us to produce relatively stable annotations while exploring alternatives at the next level down. We currently identify three levels of representation, referred to as IL0, IL1, and IL2. Each level of representation incorporates additional semantic features and removes existing syntactic ones.

IL0 is a deep syntactic dependency representation, constructed by hand-correcting the output of a dependency parser based on Connexor (`www.connexor.com`). Though this representation is purely syntactic, it abstracts as much as possible from surface-syntactic phenomena. For example, auxiliary verbs and other function words are removed from IL0. In addition, corresponding active and passive voice sentences receive the same representation in IL0. Thus it is more abstract than the Praguian Analytical level, but more syntactic than the Tectogrammatical level (Bohmova et al., 2003). IL0 is a useful starting point for IL1 in that syntactic dependencies are often indicative of semantic dependencies. Figure 1, which appears on the last page of the paper, shows the IL0 representation for the sentence *Sheikh Mohamed, who is also the Defense Minister of the United Arab Emirates, announced at the inauguration ceremony that "we want to make Dubai a new trading center."*

IL1 is an intermediate semantic representation. Open class lexical items (nouns, verbs, adjectives, and adverbs) are associated with concepts drawn from the Omega ontology (Hovy et al., 2003). Also at this stage, syntactic relations are replaced by semantic roles such as AGENT, THEME, and GOAL. However, IL1 is not an interlingua; it does not normalize over all linguistic realizations of the same semantics. Figure 2, which appears on the last page of the paper, shows the IL1 corresponding to the IL0 in Figure 1. Concept names and thematic role names added by the annotators are in upper case; some nodes are associated with more than one concept.

IL0 and IL1 have been documented with coding manuals and have been used by annotators to tag several texts. (See Section 4.) However, IL2, the deepest meaning representation, is still under development. The methodology for designing IL2 involves comparison of IL1's in the multi-parallel corpus in order to see how meaning equivalent IL1's can be reconciled or merged. IL2 is expected to normalize over:

- Conversives (e.g., *X bought a book from Y* vs. *Y sold a book to X*), as does FrameNet (Baker et al., 1998) at the more general level of Commercial_transaction.
- Non-literal language usage (e.g., *X started its business* vs. *X opened its doors to customers*).
- Extended paraphrases involving syntax, lexicon, and grammatical features (see example in introduction).

Figures 3-6 illustrate the relationship between IL1 and IL2. The examples are tentative at this point, since IL2 has not yet been formalized. Figure 3 shows the expected representation of *Mary bought a book from John* and *John sold a book to Mary* at IL1 and IL2. The IL1's for the two sentences are different because the verbs *buy* and *sell* use different participants as agents. However, the IL2 representation captures the common meaning of the buying and selling events, as has been suggested by many theories of meaning representation.

```
IL2:
  TRANSFER-POSSESSION
    [JOHN, source]
    [MARY, goal]
    [BOOK, theme]
    [PURCHASE, manner]

IL1 candidate #1:
   BUY
     [MARY, agent]
     [JOHN, source]
     [BOOK, theme]

IL1 candidate #2:
   SELL
      [JOHN, agent]
      [MARY, goal]
      [BOOK, theme]
```

Figure 3: **IL1 and IL2 for Conversives**

Figures 4-6 show an extended paraphrase in French and English. The English and French sentences are from parallel texts in the January 1997 edition of the UNESCO Courier, which is available in 29 languages and Braille. Figure 4 shows an English sentence and its IL1. The head of the English IL1 is the concept LEND. Figure 5 shows a French sentence and its IL1. The head of the French IL1 is the concept COMPRISE.

Figure 6 sketches some proposed mappings from IL1 to IL2, which would be needed in order to reconcile the

English:

Its network of eighteen independent organizations in Latin America has lent one billion dollars to microenterprises.

English IL1:

```
LEND
  [NETWORK, agent]
     [COMPRISE, mod]
         [ORGANIZATIONS, part]
  [DOLLARS, theme]
  [MICROENTERPRISE, goal]
```

Figure 4: **English Sentence and IL1**

French:

Le réseau regroupe dix-huit organisations indépendantes qui ont déboursé un milliard de dollars.

'The network comprises eighteen independent organizations which have disbursed a billion dollars'

French IL1:

```
COMPRISE
  [NETWORK, whole]
  [ORGANIZATIONS, part]
    [DISBURSE, mod]
       [NETWORK, agent]
       [DOLLARS, theme].
```

Figure 5: **French Sentence and IL1**

IL1's from Figures 4 and 5. The words *of* and *regrouper* are found to express the concept COMPRISE. The argument, ORGANIZATION of both words help to confirm that *of* and *regrouper* describe the same relation. Similarly, the concept TRANSFER-MONEY is identified as a common concept for *lend* and *debourser*, which share two arguments, NETWORK and DOLLARS.

The range of paraphrase phenomena being addressed by the different representation levels is summarized in Table 1, which is based on examples from (Hirst, 2003), (Kozlowski et al., 2003), and (Rinaldi et al., 2003). The table indicates for which types we expect to produce normalized representations reflecting the similarity in meaning between paraphrases of that type and at which level the normalization will take place.

4. Work to Date

The IAMTC project has trained approximately ten annotators, each of whom has annotated twelve texts. The twelve texts consist of two English translations of each of six foreign language articles. All annotators have worked on the same texts. This section describes the annotation procedures.

IL1-IL2 Mappings:

```
of/regroupe   <-> COMPRISE
lend/debourse <-> TRANSFER-MONEY
```

IL2:

```
COMPRISE:
  [NETWORK, whole]
  [ORGANIZATIONS, part]

TRANSFER-MONEY
  [NETWORK, agent]
  [DOLLARS, theme]
  [MIRCROENTERPRISE, goal]
```

Figure 6: **Conversion from IL1 to IL2**

In order to prepare IL0, the texts are first passed through the Connexor dependency parser. Project experts then edit the dependency trees using TrEd (Pajas, 1998). During the editing process, parsing errors are corrected and the project's conventions for dependency trees are enforced. The project's conventions concern the treatment of closed-class function words and copular sentences.

The IL0's are then passed to the annotators. The IL1 for a sentence is the result of assigning concepts and semantic roles to the nodes of the IL0. The IAMTC has developed an interface called TIAMAT for producing IL1's. Through TIAMAT, annotators can access the Omega ontology (Hovy et al., 2003), which contains concept names from WordNet (Fellbaum, 1998) and Mikrokosmos (O'Hara et al., 1998), and thematic role names from the Lexical Conceptual Structure (LCS) verb database (Dorr et al., 2001). Annotators are currently instructed to choose at least one WordNet concept and one Mikrokosmos concept for each content word in the text, or to choose a dummy concept if no suitable concept is found. Concepts that are covered in the LCS database are accompanied by a list of semantic roles. The annotators can assign a role to each of the verb's dependents, choosing either from the LCS frame or from a list of about fifteen roles that are defined in the coding manual. Annotators can also consult the IL0 in TrEd while they are using TIAMAT.

5. Evaluation

In our initial experiments, we have measured inter-annotator agreement of our semantic annotations. Many metrics have been proposed for measuring intercoder agreement in a coding task, including Kappa (Carletta, 1996) and a "Wood Standard" based on comparison of peers (Habash and Dorr, 2002). Since each text might be tagged by a different number of annotators and annotators may pick more than one sense per word, we are currently experimenting with metrics that take into account for each word the number of annotators and the number of senses that the ontology makes available for that word.

The evaluation process also involves consistency checking and reconciliation. Consistency checking involves comparison of meaning equivalent (synonymous or nearly syn-

Relationship Type	Example	Where Normalized
Syntactic variation	The gangster killed at least 3 innocent bystanders. *vs.* At least 3 innocent bystanders were killed by the gangster.	IL0
Lexical synonymy	The toddler sobbed, and he attempted to console her. *vs.* The baby wailed, and he tried to comfort her.	IL1
Morphological derivation	I was surprised that he destroyed the old house. *vs.* I was surprised by his destruction of the old house.	IL2
Clause subordination vs. anaphorically linked sentences	This is Joe's new car, which he bought in New York. *vs.* This is Joe's new car. He bought it in New York.	IL2
Different argument realizations	Bob enjoys playing with his kids. *vs.* Playing with his kids pleases Bob.	IL2
Noun-noun phrases	She loves velvet dresses. *vs.* She loves dresses made of velvet.	IL2
Head switching	Mike Mussina excels at pitching. *vs.* Mike Mussina pitches well. *vs.* Mike Mussina is a good pitcher.	IL2
Overlapping meanings	Lindbergh flew across the Atlantic Ocean. *vs.* Lindbergh crossed the Atlantic Ocean by plane.	IL2
Comparatives vs. superlatives	He's smarter than everybody else. *vs.* He's the smartest one.	Not normalized
Different sentence types	Who composed the Brandenburg Concertos? *vs.* Tell me who composed the Brandenburg Concertos.	Not normalized
Inverse relationship	Only 20% of the participants arrived on time. *vs.* Most of the participants arrived late.	Not normalized
Inference	The tight end caught the ball in the end zone. *vs.* The tight end scored a touchdown.	Not normalized
Viewpoint variation	The U.S.-led invasion/liberation/occupation of Iraq . . . You're getting in the way. *vs.* I'm only trying to help.	Not normalized

Table 1: **Relationship Types Underlying Paraphrase**

onymous) words in parallel texts. This a preliminary step toward identifying meaning equivalent sentences whose IL1's can and cannot be merged into a single IL2. It also helps us to evaluate the ontology and TIAMAT. In order to evaluate the ontology, we are interested in the extent to which there exist nodes that can express the common meaning of near synonyms. In assessing TIAMAT we would like to know how easy it is to navigate through the ontology in order to find the nodes that express the common meaning of near synonyms.

Reconciliation is the process of comparing two or more annotations produced by different people. There is a tool available for displaying multiple annotations with color coding for agreement or disagreement. The reconciliation process is conducted partly by each annotator separately and partly by interaction between the annotators. We are interested in finding out whether agreement on subsequent annotations increases as a result of reconciliation.

Another criterion to evaluate is the usefulness of the interlingual representations in NLP tasks. Since the ultimate goal is to generate a representation that is useful for MT (among other NLP tasks), we plan to measure the ability to generate accurate surface texts from the representation. We plan to use an available generator, Halogen (Langkilde and Knight, 1998). Sentences will be generated from interlinguas and then compared with the originals through a variety of standard MT metrics (ISLE, 2003). This will serve to determine whether the elements of the representation language are sufficiently well defined and whether they can serve as a basis for inferring interpretations from semantic representations or (target) semantic representations from interpretations.

6. Identification of Paraphrases

The ability to discern paraphrases is beneficial to virtually all linguistic applications, including information retrieval, information extraction, question-answering, text summarization, and machine translation. In the IAMTC framework, two sentences with the same IL2 are considered paraphrases even if they have different IL1's. The IL2 annotation on the corpus will allow us to easily study the different surface realizations of a given meaning pattern. Our intention is that these corpora will be used to improve the accuracy and robustness of semantic analysis in many NLP applications.

Acknowledgments

This work was supported by the National Science Foundation under grant number IIS-0325695. We would also like to thank our hard working annotators and graduate students.

7. References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe, 1998. The Berkeley FrameNet Project. In Christian

Boitet and Pete Whitelock (eds.), *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*. San Francisco, California: Morgan Kaufmann Publishers.

Barzilay, R. and K.R. McKeown, 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France.

Bohmova, A., J. Hajic, E. Hajicova, and B. Hladka, 2003. *The Prague Dependency Treebank: Three-Level Annotation Scenario*. Kluwer Academic Publishers.

Carletta, J.C., 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Dorr, Bonnie J., Mari Olsen, Nizar Habash, and Scott Thomas, 2001. LCS Verb Database. Technical Report Online Software Database, University of Maryland, College Park, MD. Http://www.umiacs.umd.edu/ ˜bonnie/ LCS_Database_Docmentation.html.

Fellbaum, Christiane, 1998. *WordNet: An Electronic Lexical Database*. MIT Press. Http://www.cogsci.princeton.edu/˜wn [2000, September 7].

Habash, Nizar and Bonnie J. Dorr, 2002. Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. In *Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*. Tiburon, CA.

Hirst, Graeme, 2003. Paraphrasing paraphrased. In *Keynote address for The Second International Workshop on Paraphrasing: Paraphrase acquisition and Applications*. Association for Computational Linguistics.

Hovy, Eduard, Andrew Philpot, Judith L. Klavans, Ulrich Germann, and Peter T. Davis, 2003. Extending metadata definitions by automatically extracting and organizing glossary definitions. In *Proceedings of the National Conference on Digital Government Research*. Boston, MA.

ISLE, 2003. Towards Systematizing MT Evaluation. In *Proceedings of MT Summit IX Evaluation Workshop*. New Orleans, Louisiana.

Kingsbury, P. and M. Palmer, 2002. From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Spain.

Kozlowski, Raymond, Kathleen F. McCoy, and K. Vijay-Shanker, 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*. ACL 2003.

Langkilde, I. and K. Knight, 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for computational Linguistics joint with the 17th International Conference on Computational Linguistics, ACL/COLING*. Montreal, Canada.

O'Hara, Tom, Kavi Mahesh, and Sergei Nirenburg, 1998. Lexical acquisition with wordnet and the mikrokosmos ontology. In *Proceedings of the ACL Workshop on the Use of WordNet in NLP*. Montreal, Canada.

Pajas, Petr, 1998. Tree editor manual. Technical report, CLSP Summer Sorkshop, Johns Hopkins University, Baltimore, MD.

Rinaldi, Fabio, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Moll, 2003. Exploiting paraphrases in a question-answering system. In *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*. ACL 2003.

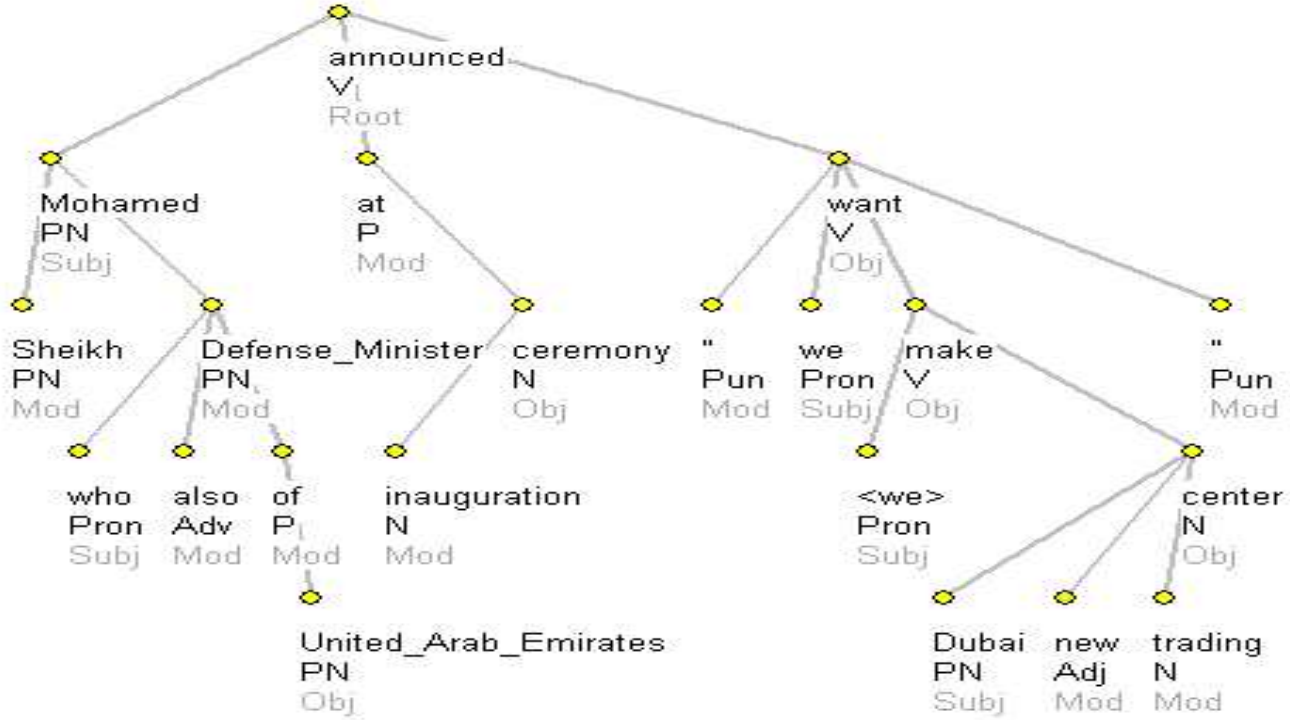

Figure 1: **An IL0 Dependency Tree**

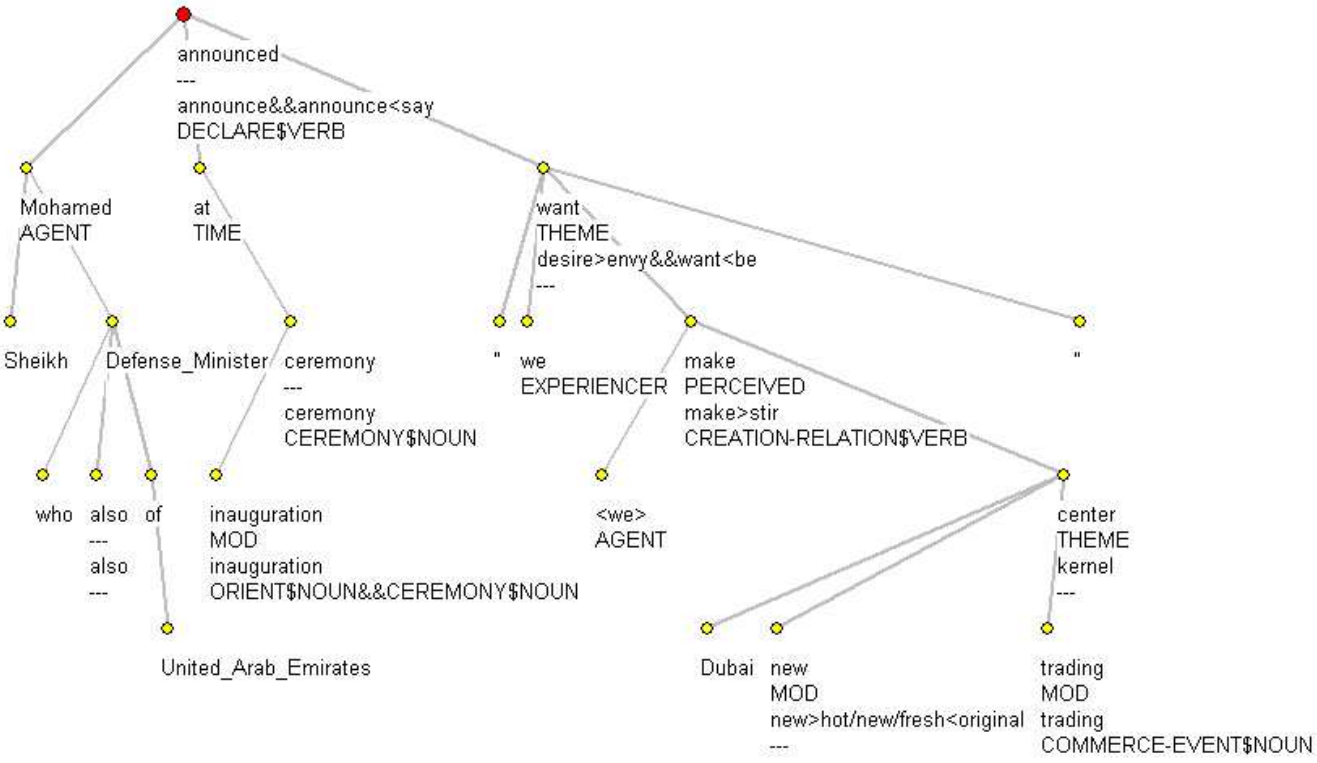

Figure 2: **An IL1 Representation**