# LG-Soar

Clint Tustison

Brigham Young University

Research funded in part by the NSF

# Introduction

- Challenge
  - Extract predicates from various natural language texts
    - Newspaper headlines
    - Eligibility criteria for medical clinical trials
    - GEDCOM files (format for encoding genealogical information
- Issues
  - Variability in linguistic structure of utterances
  - Final (extracted) representation must be usable
  - Soar integration: Flexible multipurpose platform
    - Goal-directed problem solving
    - Agent-based architecture
    - Proven in other applications

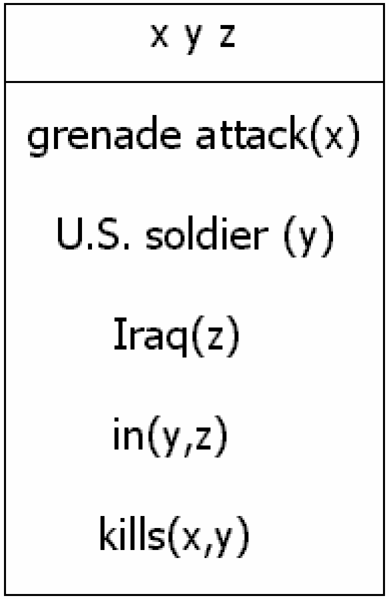
# Approach

- Tools
  - Link-Grammar Parser
    - Sleator, Lafferty, Temperley
    - Characteristics
      - Syntactic dependency parse
      - Constraints for determining grammaticality
    - Benefits
      - written in C → very fast
      - Robust - ability to process (un)grammaticality / spelling errors
      - Free - http://www.link.cs.cmu.edu/link
      - Easily integrated

# Tools: representation

- Discourse Representation Theory
  - Hans Kamp & Uwe Reyle (University of Stuttgart)
  - Theory to represent relations that exist within and across utterances
  - Ability to account for semantic and pragmatic information
  - Easily translatable into predicate logic
- Predicate Logic
  - Formal properties, allow for wide range of applications, usable crosslinguistically
  - Vocabulary, syntax, semantics
    - First-order: quantification over individuals (FOPC)
    - Higher-order: quantification over relations, etc.

# News headlines extraction



x y z

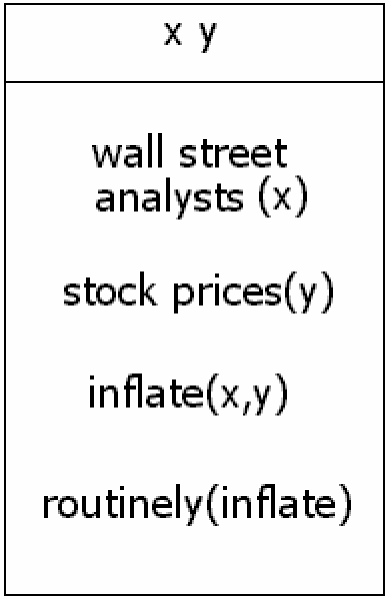
grenade attack(x)

U.S. soldier (y)

Iraq(z)

in(y,z)

kills(x,y)



x y

wall street analysts (x)

stock prices(y)

inflate(x,y)

routinely(inflate)

grenade attack(x) & u.s. soldier(y) & iraq(z) & in(y,z) & kills(x,y).

wall street analysts(x) & stock prices(y) & inflate(x,y) & routinely(inflate).

# Clinical trials extraction

- Novel Adjuvants for Peptide-Based Melanoma Vaccines

- INCLUSION CRITERIA:
  - Ages Eligible for Study: 18 Years and above
  - Genders Eligible for Study: Both
  - Diagnosis of stage III or IV cutaneous, mucosal, or ocular melanoma
  - . . .

- EXCLUSION CRITERIA:
  - Steroid therapy
  - Allergic reaction to Montanide ISA 51
  - Positive for hepatitis B, hepatitis C, or HIV
  - . . .

# Predicates: inclusion criteria

- Ages Eligible for Study:  18 Years  and above
  - age(Person,X) & X >= 18.
- Genders Eligible for Study:  Both
  - gender(Person,X) & (female == X || male == X).
- Diagnosis of stage III or IV cutaneous, mucosal, or ocular melanoma
  - diagnosis(Person,X) & melanoma(X) & type(X,Y) & (cutaneous(Y) || mucosal(Y) || ocular(Y)) & stage(X,Z) & (Z == 3 || Z == 4).

# Predicates: exclusion criteria

- Allergic reaction to Montanide ISA 51
    - ¬(allergy(Person,X) & montanide(X)).
- Steroid therapy
    - ¬(therapy(Person,X) & steroid(X)).
- Positive for hepatitis B, hepatitis C, or HIV
    - ¬(condition(Person,X) & hepatitis_B(X) || hepatitis_C(X) || hiv(X)).

# GEDCOM Extraction

- individual(i1,name('Dovie MELLISSIA /STEVENSON/'),sex(f),parentin(f1),childin(f2),birthdate('18 Sep 1908'),baptismdate('10 Apr 1919'),endowdate('9 Mar 1976'),deathdate(''),birthplace('OKTAHA, MUSKOGEE, OK, USA'),deathplace(''),burialplace('')).

- individual(i2,name('WILLIAM JAMES /STEVENSON/'),sex(m),parentin(f4),childin(f5),birthdate('5 Sep 1880'),baptismdate('13 Sep 1903'),endowdate('9 May 1969'),deathdate('22 Nov 1964'),birthplace('PENDLETON, WARREN, PA'),deathplace('TULARE, TULARE, CA'),burialplace('VISALIA, TULARE, CA')).

# Inferencing

```
/********************************************************************************
Which husband/wife combination was born on the same day in the same place?
*********************************************************************************/
husband_wife(HusbandName,HBirthdate,WifeName,WBirthdate,X) :-
    individual(Husband,name(HusbandName),_,_,_,birthdate(HBirthdate),_,_,_,
    birthplace(X),_,_),family(_,husband(Husband),_,_),
    parse_date(HBirthdate,HDay,HMonth,HYear),individual(Wife,name(WifeName),_,
    _,_,birthdate(WBirthdate),_,_,_,birthplace(X),_,_),family(_,_,wife(Wife),_),
    parse_date(WBirthdate,WDay,WMonth,WYear),HYear == WYear,HMonth == WMonth,
    HDay == WDay.
```

```
HusbandName = Garland /Bailey/          HusbandName = Charles Arthur /Goodpasture/
HBirthdate = 16 Apr 1912                HBirthdate = 25 Dec 1894
WifeName = Carolyn /Warren/             WifeName = Betty Lucille /Rittga/
WBirthdate = 16 Apr 1912                WBirthdate = 25 Dec 1894
Place = Gracemont, Caddo, Oklahoma      Place = Gracemont, Caddo, Oklahoma
```

# LG-Soar Progress

- Ability to handle various grammatical structures
  - Transitives
  - Intransitives
  - Imperatives
  - Negation
  - Definiteness/indefiniteness
  - Modals
  - Certain anaphoric constructions
  - Nominal compounds
  - Modification
  - Prepositional phrase attachment to NPs
  - Relative clauses

# Contributions/Future Work

- Contributions
  - Robustly extract predicates from natural language
    - Multiple domains
    - Various natural language syntactic constructions
  - Use applications to access predicates
    - Inferencing and querying
- Future Work
  - Additional domains
  - Integrate with external knowledge sources
    - Wordnet
    - UMLS
  - Upgrade to higher-order predicate calculus to allow predication over relations and events, not just individuals
  - Senseval 3

# Conclusion

- Coals
  - Vocabulary is difficult to write
  - Only one parsed output per utterance
  - Coverage and correctness need improvement

- Nuggets
  - Fast
  - Robust
  - Implementation in other languages
  - Can be easily integrated with other applications/corpora