

Controlled Natural Language and Opportunities for Standardization

Tobias Kuhn

Chair of Sociology, in particular of Modeling and Simulation, ETH Zurich,
Switzerland

LaRC 2013, Pretoria (South Africa)

8 June 2013

About This Presentation

This presentation is based on the following journal article:

Tobias Kuhn. A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, to appear.

It can be downloaded here: <http://purl.org/tkuhn/cnlsurvey>

Consult the survey article above for references to the languages and tools shown in this presentation.

Outline

This talk consists of the following parts:

- **Introduction:** What are Controlled Natural Languages (CNLs)?
- **Languages:** What are concrete examples of CNLs?
- **Properties:** What are the types and properties of CNLs?
- **Applications:** In what applications are CNLs used?
- **Analysis:** What does the big picture of existing CNLs look like?
- **Evaluation:** Do CNLs actually work?
- **Standardization:** What are the opportunities for Standardization?

Part 1: Introduction

What are Controlled Natural Languages (CNLs)?

AECMA Simplified English AIDA Airbus Warning Language ALCOGRAM ASD Simplified Technical English Atomate Language Attempto Controlled English Avaya Controlled English Basic English BioQuery-CNLI Boeing Technical English Bull Global English CAA Phraseology Caterpillar Fundamental English Caterpillar Technical English Clear And Simple English ClearTalk CLEF Query Language COGRAM Common Logic Controlled English Computer Processable English Computer Processable Language Controlled Automotive Service Language Controlled English at Clark Controlled English at Douglas Controlled English at IBM Controlled English at Rockwell Controlled English to Logic Translation Controlled Language for Crisis Management Controlled Language for Inference Purposes Controlled Language for Ontology Editing Controlled Language Optimized for Uniform Translation Controlled Language of Mathematics Coral's Controlled English Diebold Controlled English DL-English Drafter Language E-Prime E2V IBM's EasyEnglish Wycliffe Associates' EasyEnglish Ericsson English FAA Air Traffic Control Phraseology First Order English Formalized-English ForTheL Gellish English General Motors Global English Gherkin GINO's Guided English Ginseng's Guided English Hyster Easy Language Program ICAO Phraseology ICONOCLAST Language iHelp Controlled English iLastic Controlled English International Language of Service and Maintenance ITA Controlled English KANT Controlled English Kodak International Service Language Lite Natural Language Massachusetts Legislative Drafting Language MILE Query Language Multinational Customized English Nortel Standard English Naproche CNL NCR Fundamental English Océ Controlled English OWL ACE OWLPath's Guided English OWL Simplified English PathOnt CNL PENG PENG-D PENG Light Perkins Approved Clear English PERMIS Controlled Natural Language PILLS Language Plain Language PoliceSpeak PROSPER Controlled English Pseudo Natural Language Quelo Controlled English Rabbit Restricted English for Constructing Ontologies Restricted Natural Language Statements RuleSpeak SBVR Structured English SEASPEAK SMART Controlled English SMART Plain English Sowa's syllogisms Special English SQUALL Standard Language Sun Proof Sydney OWL Syntax Template Based Natural Language Specification ucsCNL Voice Actions

There are a wide variety of CNLs applied to a wide variety of problem domains.

The study of CNLs has so far been somewhat chaotic and disconnected.

Approved Basic Clear Controlled Easy Formal-
ized Fundamental Global Guided International
Light Multinational Plain Processable Pseudo
Restricted Simple Simplified Special Standard
Structured Technical

Introduction

All these languages share important properties and **it makes sense to put them under the same umbrella.**

This presentation should give the **necessary background** to tackle the following questions:

- Is CNL standardization **necessary**?
- Is CNL standardization **possible**?
- Which **aspects** of CNL could and should be standardized?

Part 2: Languages

What are concrete examples of CNLs?

Languages

We restrict ourselves to [English-based](#) languages. First, we look at [twelve selected CNLs](#):

- Syllogisms
- Basic English
- E-Prime
- Caterpillar Fundamental English (CFE)
- FAA Air Traffic Control Phraseology
- ASD Simplified Technical English (ASD-STE)
- Standard Language (SLANG)
- SBVR Structured English
- Attempto Controlled English (ACE)
- Drafter Language
- E2V
- Formalized-English (FE)

Syllogisms

- Simple logic languages originally introduced by [Aristotle](#) ca. 350 BC (in ancient Greek)
- Several versions of syllogisms in English exist, e.g. “Sowa’s Syllogisms”
- Claimed to be the [first reported instance](#) of a CNL

In a simple version, the complete language can be described by just four [simple sentence patterns](#):

Every A is a B.

Some A is a B.

No A is a B.

Some A is not a B.

Syllogisms: Variations and Examples

Slightly more complex versions include patterns like:

Every A is not a B.

No A is not a B.

P is a B.

P is not a B.

Examples

Every man is a human.

Some animal is a cat.

No animal is a plant.

Some animal is not a mammal.

Syllogisms: Reasoning

For certain pairs of syllogisms, a third follows:

No A is a B.

Every C is a A.

⇒

No C is a B.

Example

No reptile is a mammal.

Every snake is a reptile.

⇒

No snake is a mammal.

Basic English

- Presented in 1930
- Should **improve communication** among people around the globe
- **First reported instance** of a controlled version of English
- Influenced Caterpillar Fundamental English, which became itself a very influential language
- Designed as a common basis for communication in **politics, economy, and science**
- Restricts the grammar and makes use of only **850 English root words**
- **Only 18 verbs are allowed:** put, take, give, get, come, go, make, keep, let, do, be, seem, have, may, will, say, see, and send.
- For more specific relations, verbs can be combined with prepositions, such as *put in* to express *insert*, or with nouns, such as *give a move* instead of using *move* as a verb

Basic English: Examples

Examples

The camera man who made an attempt to take a moving picture of the society women, before they got their hats off, did not get off the ship till he was questioned by the police.

It was his view that in another hundred years Britain will be a second-rate power.

- **Many variations** exist that use larger word sets (e.g. the Simple English version of Wikipedia)
- Basic English is still used today and promoted by the **Basic-English Institute**
- **Many texts** have been written in this language, including textbooks, novels, and large parts of the bible

E-Prime (or E')

- **Only restriction:** the verb *to be* is forbidden to use
- Includes **all inflectional forms** such as *are*, *was* and *being* regardless of whether used as auxiliary or main verb
- Presented in 1965 but the idea goes back to the late 1940s
- Motivation is the belief that “**dangers and inadequacies [...] can result from the careless, unthinking, automatic use of the verb ‘to be’**”
- Claimed by its proponents to **enhance clarity**

Instead of “We do this because it is right,” one would write:

Examples

We do this thing because we sincerely desire to minimize the discrepancies between our actions and our stated “ideals.”

Caterpillar Fundamental English (CFE)

- Influential CNL developed at Caterpillar
- Officially introduced in 1971, based on Basic English
- Reported to be the earliest industry-based CNL

Motivation: increasing sophistication of Caterpillar's products and the need to communicate with non-English speaking service personnel in different countries.

“To summarize the problem: There are more than 20,000 publications that must be understood by thousands of people speaking more than 50 different languages.”

Caterpillar Fundamental English: Approach

- The idea of CFE was “to eliminate the need to translate service manuals”
- A trained, non-English speaking mechanic familiar with Caterpillar’s products should be able to understand the language after completing a course consisting of 30 lessons
- Vocabulary is restricted to around 800 to 1,000 words
- Only one meaning defined for each of the words: e.g., *right* only as the opposite of *left*

Caterpillar Fundamental English: Summarizing Rules

- 1 Make positive statements.
- 2 Avoid long and complicated sentences.
- 3 Avoid too many subjects in one sentence.
- 4 Avoid too many successive adjectives and nouns.
- 5 Use uniform sentence structures.
- 6 Avoid complicated past and future tenses.
- 7 Avoid conditional tenses.
- 8 Avoid abbreviations, contractions, and colloquialisms.
- 9 Use punctuation correctly.
- 10 Use consistent nomenclature.

Caterpillar Fundamental English: Examples

Examples

The maximum endplay is .005 inch.

Lift heavy objects with a lifting beam only.

- **Discontinued** by Caterpillar in 1982, because (among other reasons) “the basic guidelines of CFE were not enforceable in the English documents produced”
- As a result, **Caterpillar Technical English (CTE)** was developed
- **Approach of CTE:** Restrictions should be enforceable, and should reduce translation costs (instead of trying to eliminate the need for translations altogether)

FAA Air Traffic Control Phraseology

- Defined by the Federal Aviation Administration (FAA), since at least the early 1980s
- Used for the [communication in air traffic coordination](#)
- [Very similar languages](#): ICAO and CAA phraseologies
- Together they are sometimes called **AirSpeak**
- Vocabulary and meaning are restricted
- [Exemplary restriction](#): “Use the word *gain* and/or *loss* when describing to pilots the effects of wind shear on airspeed.”
- Phraseology statements can be mixed with statements in full English (when no pattern exists to express the desired message)

FAA Air Traffic Control Phraseology: Examples

More than 300 fixed sentence patterns such as “(ACID), IN THE EVENT OF MISSED APPROACH (issue traffic). TAXIING AIRCRAFT/VEHICLE LEFT/RIGHT OF RUNWAY.”

Many more implicit patterns, for example “Issue advisory information on [...] bird activity. Include position, species or size of birds, if known, course of flight, and altitude.”

Examples

United 623, in the event of missed approach, taxiing aircraft right of runway.

Flock of geese, one o'clock, seven miles, northbound, last reported at four thousand.

ASD Simplified Technical English (ASD-STE)

- Often abbreviated to **Simplified Technical English (STE)** or just **Simplified English**
- CNL for the **aerospace industry**
- Had its origins in 1979, but officially presented only in 1986, then under the name **AECMA Simplified English**
- Received its current name in 2004 when AECMA merged with two other associations to form ASD
- **Motivation:** Make texts easier to understand, especially for non-native speakers
- AECMA Simplified English was designed to **make translation into other languages unnecessary**
- ASD-STE's design goals included **improved translation**
- Maintained by the Simplified Technical English Maintenance Group

ASD Simplified Technical English: Restrictions

Restrictions expressed in about 60 general rules:

- **Lexical level** (e.g., “Use approved words from the Dictionary only as the part of speech given”)
- **Syntactic level** (e.g., “Do not make noun clusters of more than three nouns”)
- **Semantic level** (e.g., “Keep to the approved meaning of a word in the Dictionary. Do not use the word with any other meaning.”)
- Fixed vocabulary of terms common to the aerospace domain
- User-defined “**Technical Names**” and “**Technical Verbs**” can be introduced

Example

These safety precautions are the minimum necessary for work in a fuel tank. But the local regulations can make other safety precautions necessary.

Standard Language (SLANG)

- Developed at Ford Motor Company starting from 1990
- Designed for process sheets containing **build instructions** for component and vehicle assembly plants
- **Still used** at Ford and has been **continually extended and updated** to reflect technical and business-related advances
- **Motivation:** engineers can write instructions that are **clear and concise** and at the same time **machine-readable**
- System can, among other things, automatically generate a list of required elements and calculate labor times
- In addition, **machine translation** is applied for the use of such instructions in assembly plants in different countries

Standard Language (SLANG): Examples

- Sentences in **imperative mood** starting with a main verb and followed by a noun phrase
- **Additional restrictions on vocabulary and semantics**
- Parser can check for compliance

Examples

OBTAIN ENGINE BLOCK HEATER ASSEMBLY FROM STOCK
APPLY GREASE TO RUBBER O-RING AND CORE OPENING

SBVR Structured English

- CNL for **business rules** first presented around 2005
- Part of the Semantics of Business Vocabulary and Business Rules (SBVR) standard
- Probably influenced by a similar language called **RuleSpeak** that was first presented in 1994
- The vocabulary is extensible and consists of three types of sentence constituents:
 - **terms** (i.e., concepts)
 - **names** (i.e., individuals)
 - **verbs** (i.e., relations)
 - **keywords** (e.g., fixed phrases, quantifiers, and determiners)

SBVR Structured English: Examples

Examples

A rental **must** *be guaranteed by a* credit card *before a* car *is assigned to the rental.*

Rentals by Booking Mode *contains the* categories *'advance rental'* *and 'walk-in rental.'*

- **Formal semantics:** second-order logic with Henkin semantics
- Some keywords have a **precise meaning**, such as *or* meaning inclusive logical disjunction (unless followed by *but not both*)
- Other keywords are **less precise**, such as the determiner *a* being defined as “universal or existential quantification, depending on context based on English rules”
- Permissible sentence constituents are strictly defined, but not the grammatical rules to form sentences

Attempto Controlled English (ACE)

- CNL with an automatic and unambiguous translation into first-order logic
- First presented in 1996 as a language for software specifications
- Later, the focus shifted towards knowledge representation and the Semantic Web
- Features include complex noun phrases, plurals, anaphoric references, subordinated clauses, modality, and questions

Examples

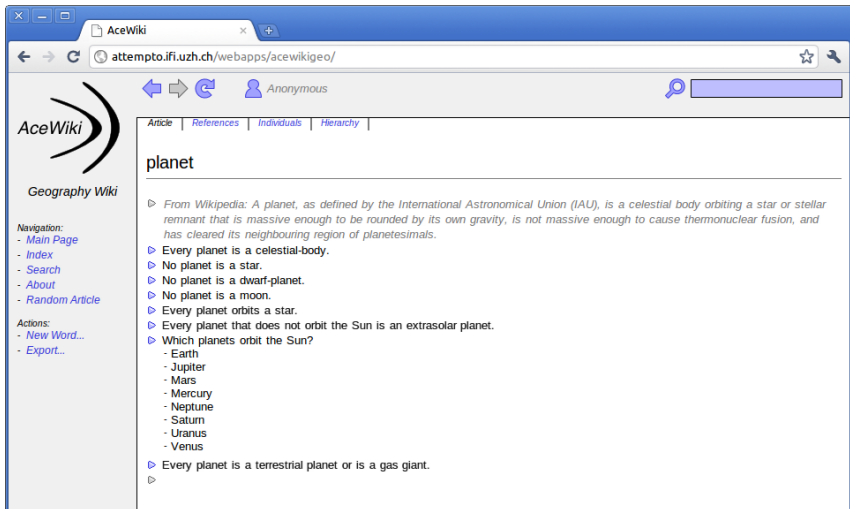
A customer owns a card that is invalid or that is damaged.

Every continent that is not Antarctica contains at least 2 countries.

Attempto Controlled English: Application

- Deterministic mapping to **discourse representation structures** (notational variant of first-order logic)
- **These expressions are underspecified**: many deductions require external background axioms that are not fixed by the ACE definition
- Used in different areas such as **ontology editors, rule systems, and general reasoners**
- Recently, ACE has also been used in the context of **rule-based machine translation**

ACE Application: AceWiki



The screenshot shows a web browser window with the title "AceWiki" and the URL "attempto.ifi.uzh.ch/webapps/acewikigeo/". The browser's address bar includes navigation icons (back, forward, refresh) and a search icon. The page content is displayed in a light blue theme. On the left side, there is a navigation menu for "Geography Wiki" with links for "Main Page", "Index", "Search", "About", "Random Article", "New Word...", and "Export...". The main content area features a breadcrumb trail: "Article | References | Individuals | Hierarchy". The article title is "planet". Below the title, there is a definition from Wikipedia: "A planet, as defined by the International Astronomical Union (IAU), is a celestial body orbiting a star or stellar remnant that is massive enough to be rounded by its own gravity, is not massive enough to cause thermonuclear fusion, and has cleared its neighbouring region of planetesimals." This is followed by a list of properties and questions, each preceded by a blue triangle icon: "Every planet is a celestial-body.", "No planet is a star.", "No planet is a dwarf-planet.", "No planet is a moon.", "Every planet orbits a star.", "Every planet that does not orbit the Sun is an extrasolar planet.", and "Which planets orbit the Sun?". The last item has a sub-list: "Earth", "Jupiter", "Mars", "Mercury", "Neptune", "Saturn", "Uranus", and "Venus".

AceWiki
Geography Wiki

Navigation:
- [Main Page](#)
- [Index](#)
- [Search](#)
- [About](#)
- [Random Article](#)

Actions:
- [New Word...](#)
- [Export...](#)

Article | [References](#) | [Individuals](#) | [Hierarchy](#)

planet

▷ *From Wikipedia: A planet, as defined by the International Astronomical Union (IAU), is a celestial body orbiting a star or stellar remnant that is massive enough to be rounded by its own gravity, is not massive enough to cause thermonuclear fusion, and has cleared its neighbouring region of planetesimals.*

- ▷ Every planet is a celestial-body.
- ▷ No planet is a star.
- ▷ No planet is a dwarf-planet.
- ▷ No planet is a moon.
- ▷ Every planet orbits a star.
- ▷ Every planet that does not orbit the Sun is an extrasolar planet.
- ▷ Which planets orbit the Sun?
 - Earth
 - Jupiter
 - Mars
 - Mercury
 - Neptune
 - Saturn
 - Uranus
 - Venus
- ▷ Every planet is a terrestrial planet or is a gas giant.

ACE: Predictive Editor

Sentence Editor

Every country that is located in Europe is ...

< Delete

text

function word

- a
- an
- every
- everybody
- everything
- no
- nobody

transitive adjective

- new...
- the capital of
- larger than
- located in
- the national-language of
- registered at
- smaller than

new variable

- X
- Y
- Z

proper name

- new...
- Africa
- Aland
- Alberta
- America
- Andorra
- Angela Merkel

passive verb

- new...
- built by
- contained by
- cultivated by
- flown through by
- governed by
- grown by

reference

- the country

OK Cancel

Drafter Language

- CNL used in a system called Drafter-II presented in 1998
- For instructions to word processors and diary managers
- Conceptual authoring approach:
 - Users cannot directly edit the CNL text
 - They can only trigger modification actions
 - Starting from a small stub sentence
 - Incomplete statements are gradually completed by the user

Examples

Schedule **this event** by applying *this method*.

Schedule the appointment by applying *this method*.

- Internally maps to Prolog expressions, which are automatically executed
- Possible structural ambiguity can be resolved based on the given sequence of modification actions

E2V

- CNL introduced in 2001
- Corresponds to the language \mathcal{E}_3 studied in later work
- The ultimate goal is “to provide useable tools for natural language system specification”
- **Deterministic mapping** to 2-variable fragment of first-order logic
- Decidable and computation is NEXPTIME complete
- Defined by **15 simple grammar rules** (plus nine predefined lexical rules)
- Separate, user-defined lexicon contains the content words such as *artist* and *admires*

Examples

Some artist does not despise every beekeeper.

Every artist who employs a carpenter despises every beekeeper who admires him.

Formalized-English (FE)

- CNL for *knowledge representation* introduced in 2002
- Focusing on *expressiveness*
- Based on Conceptual Graphs and the Knowledge Interchange Format (KIF)
- *Covers a wide range of features*: general universal quantification, negation, contexts (statements about statements), lambda abstractions, possibility, collections, intervals, higher-order statements (reducible to first-order logic), and more
- *Quite unnatural* for complex statements

Examples

At least 93% of [bird with chr c a good health] can be agent of a flight.

If 'a binaryRelationType *rt has for chr c the transitivity' then 'if '^x has for *rt ^y that has for *rt ^z' then '^x has for *rt ^z' '.

Part 3: Properties

What are the types and properties of CNLs?

Properties

To get a **more principled view** on CNLs and their properties, we will now look at:

- Definitions of the term CNL
- Related terms
- Properties of CNL environments
- Inherent language properties
- CNL implementation

Existing Definitions

*A **controlled language (CL)** is a restricted version of a natural language which has been engineered to meet a special purpose, most often that of writing technical documentation for non-native speakers of the document language. A typical CL uses a well-defined subset of a language's grammar and lexicon, but adds the terminology needed in a technical domain.*

***Controlled natural language** is a subset of natural language that can be accurately and efficiently processed by a computer, but is expressive enough to allow natural usage by non-specialists.*

Our Definition: Short

A **controlled natural language** is a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics while preserving most of its natural properties.

Our Definition: Long

A language is called a **controlled natural language** if and only if it has all of the following four properties:

- ① It is based on **exactly one natural language** (its “base language”).
- ② The most important difference between it and its base language (but not necessarily the only one) is that it is **more restrictive** concerning lexicon, syntax, and/or semantics.
- ③ It preserves most of the **natural properties** of its base language, so that speakers of the base language can intuitively and correctly understand texts in the controlled natural language, at least to a substantial degree.
- ④ It is a **constructed language**, which means that it is explicitly and consciously defined, and *is not* the product of an implicit and natural process (even though it is based on a natural language that *is* the product of an implicit and natural process).

Related Terms

There are a number of [terms related to CNL](#), and these are easy to confuse:

- Sublanguages
- Fragments of language
- Style guides
- Phraseologies
- Controlled Vocabularies
- Constructed Languages

Sublanguages

Sublanguages are languages that naturally arise when “a community of speakers (i.e. ‘experts’) shares some specialized knowledge about a restricted semantic domain [and] the experts communicate about the restricted domain in a recurrent situation, or set of highly similar situations.”

- Like CNLs, a sublanguage is **based on exactly one natural language** and is **more restricted**
- Crucial difference: **sublanguages emerge naturally**; CNLs are explicitly and consciously defined

Fragments of Language

Fragments of language is a term denoting “a collection of sentences forming a naturally delineated subset of [a natural] language.”

- Closely related to CNL: difference is mainly **methodological**
- Fragments of language are *identified* rather than *defined*
- Kept in the context of the full natural language and related fragments
- Purpose is rather to **theoretically study** them than to directly use them to solve a particular problem

Style Guides

Style guides are documents containing instructions on how to write in a certain natural language.

- Some style guides such as “How to write clearly” by the European Commission provide “**hints, not rules**”
- Such languages do not describe a new language, but only give advice on how to use the given natural language
- Others such as the Plain Language guidelines are stricter and *do* describe a **language not identical to the respective full language**
- Such languages should be considered CNL if they did not emerge naturally

Phraseologies

Phraseology is a term that denotes a “set of expressions used by a particular person or group.”

- **Simpler grammatical structure** than in full natural language
- Not a selection of sentences but a *selection of phrases*
- Can be natural or constructed
- **Constructed phraseologies** are usually considered CNLs

Controlled Vocabularies

Controlled vocabularies are standardized collections of names and expressions, including “lists of controlled terms, synonym rings, taxonomies, and thesauri.”

- Mostly cover a **specific, narrow domain**
- In contrast to CNL, they **do not deal with grammatical issues** (i.e., how to combine the terms to write complete sentences)
- Many CNL approaches, especially domain-specific ones, include controlled vocabularies

Constructed Languages or Artificial Languages or Planned Languages

Constructed languages are languages that did not emerge naturally but have been **explicitly and consciously defined**.

The term includes (but is not limited to) languages such as:

- Esperanto
- Programming languages
- CNLs

Properties of CNLs

CNLs have a variety of different properties:

- Some are **inherent language properties**
- Others are **properties of the environment** in which the language is used

Let's see...

Property: Problem Domain

CNLs can be subdivided according to the **problem they are supposed to solve**:

- To **improve communication** among humans, especially speakers with different native languages (letter code **c**)
- To **improve** manual, computer-aided, semi-automatic, or automatic **translation** (**t**)
- To provide a **natural** and intuitive representation for **formal notations** (**f**); this includes approaches for automatic execution of texts

Type **c** is the oldest, type **t** emerged later, and type **f** is the most recent of the three.

Property: Problem Domain — Alternative Classifications

Alternative binary classifications dominate the existing literature:

- “Human-oriented” ($\sim \mathbf{c}$) vs. “computer-oriented” ($\sim \mathbf{t}$ and \mathbf{f}) languages
- “Naturalistic” ($\sim \mathbf{c}$ and \mathbf{t}) vs. “formalistic” ($\sim \mathbf{f}$) languages

More Properties

CNLs can be intended to be **written (w)** and/or **spoken (s)**.

CNLs can be targeted towards a **specific and narrow domain (d)**.

CNLs can originate from an **academic (a)**, **industrial (i)**, and/or **governmental (g)** environment.

Properties: Overview

Code	Property
c	The goal is comprehensibility
t	The goal is translation
f	The goal is formal representation
w	The language is intended to be written
s	The language is intended to be spoken
d	The language is designed for a specific narrow domain
a	The language originated from academia
i	The language originated from industry
g	The language originated from a government

Inherent Language Properties?

These properties do not seem to be inherent language properties:

- Languages that originated in academia can later be used in industry or a government, and vice versa
- The lexicon can later be declared open or closed
- Written languages can be read aloud
- Spoken languages can be written down

The properties collected so far seem to describe **language environments rather than the languages themselves.**

Inherent Language Properties

There are, of course, inherent language properties. They can be subsumed by the following four dimensions:

- **Precision:** Is there vagueness, ambiguity, context sensitivity, or room for interpretation?
- **Expressiveness:** What is the range of statements that can be expressed?
- **Naturalness:** How much does it resemble natural language?
- **Simplicity/Complexity:** How difficult is it to fully define the language or to implement it in a computer program?

These inherent language properties are, however, **difficult to quantify**.

Languages for Comparison

To get an understanding of the **nature of CNLs**, it is helpful to look at some **languages for comparison**:

- English (or any other natural language)
- Propositional logic
- Manchester OWL Syntax
- First-order logic
- COBOL

Language for Comparison: Propositional Logic

Propositional logic is a *very basic logic language*.

Example

$$A \wedge \neg B \rightarrow C$$

Meaning of this example: *"If A and not B then C."*

A, B, and C could stand for "it is Sunday," "it is raining," and "the park is crowded."

Language for Comparison: Manchester OWL Syntax

The Manchester OWL Syntax is a *user-friendly syntax* for the ontology language OWL.

Example

Pizza **and not** (hasTopping **some** FishTopping) **and not**
(hasTopping **some** MeatTopping)

Instead of logical symbols, natural words such as **not** and **some** are used.

Inherent Language Properties: Classification Scheme for CNLs

Conceptually, CNLs are somewhere in the **gray area between natural languages** on the one end **and formal languages** on the other.

Natural versus formal languages:

- Natural languages such as English are **very expressive, but complex and imprecise**
- Formal languages such as propositional logic are **very simple and precise, but at the same time unnatural and inexpressive**
- CNLs must be somewhere in the middle ...

... but where exactly?

PENS Classification Scheme

Four PENS dimensions:

- **Precision:** from *very imprecise* (e.g., English) to *maximally precise* (e.g., propositional logic)
- **Expressiveness:** from *very inexpressive* (e.g., propositional logic) to *maximally expressive* (e.g. English)
- **Naturalness:** from *very unnatural* (e.g., propositional logic) to *fully natural* (e.g., English)
- **Simplicity:** from *extremely complex* (e.g., English) to *very simple* (e.g., propositional logic)

PENS defines five consecutive non-overlapping classes in each dimension: $P^1 - P^5$, $E^1 - E^5$, $N^1 - N^5$, $S^1 - S^5$

Precision

The precision dimension captures the degree to which the **meaning of a text can be directly retrieved from its textual form**, that is, the sequence of language symbols.

- Imprecise languages (P^1)
- Less imprecise languages (P^2)
- Reliably interpretable languages (P^3)
- Deterministically interpretable languages (P^4)
- Languages with fixed semantics (P^5)

Imprecise languages (P¹)

Examples:

- All natural languages
- E-Prime

Criteria:

- Virtually every sentence of these languages is **vague** to a certain degree
- Without taking context into account, most sentences of a certain complexity are **ambiguous**
- The automatic interpretation of such languages is “**AI-complete**”
- Require a human reader to check syntax and meaning of statements

Less imprecise languages (P^2)

Examples:

- Basic English
- Caterpillar Fundamental English / ASD-STE
- FAA Air Traffic Control Phraseology

Criteria:

- Less ambiguity and vagueness than in natural languages
- Interpretation depends much less on context
- Restrict the use and/or the meaning of a wide range of the ambiguous, vague, or context-dependent constructs
- Restrictions are not sufficient to make automatic interpretation reliable
- No formal (i.e., mathematically precise) underpinning

Reliably interpretable languages (P³)

Examples:

- Standard Language
- SBVR Structured English

Criteria:

- Heavily restricted syntax (not necessarily formally defined)
- Reliable automatic interpretation
- Logical underpinning or formal conceptual scheme to represent semantics
- No fully formalized mapping of sentences to their semantic representations
- External background knowledge, heuristics, or user feedback are required

Deterministically interpretable languages (P⁴)

Examples:

- Attempto Controlled English
- Drafter Language

Criteria:

- **Fully formal** on the syntactic level (can be defined by a formal grammar)
- **Parse deterministically** to a formal logic representation (or a small closed set of all possible representations)
- Representations may be **underspecified**: they may require certain parameters, background axioms, external resources, or heuristics to enable sensible deductions

Languages with fixed semantics (P⁵)

Examples:

- Syllogisms
- E2V
- Formalized-English
- Manchester OWL Syntax
- Propositional logic

Criteria:

- Fully formal and fully specified on syntactic and semantic levels
- Each text has exactly one meaning, which can be automatically derived
- The circumstances in which inferences hold or do not hold are fully defined
- No heuristics or external resources are necessary

Expressiveness

The dimension of expressiveness describes the **range of propositions** that a certain language is able to express.

A language X is more expressive than a language Y if language X can describe everything that language Y can, but not vice versa.

- Inexpressive languages (E^1)
- Languages with low expressiveness (E^2)
- Languages with medium expressiveness (E^3)
- Languages with high expressiveness (E^4)
- Languages with maximal expressiveness (E^5)

Inexpressive languages (E^1)

Examples:

- FAA Air Traffic Control Phraseology
- Standard Language
- Drafter Language
- Syllogisms
- Propositional logic

Criteria:

- No universal quantification, or
- No relations of arity greater than 1 (e.g., binary relations)

Languages with low expressiveness (E^2)

Examples:

- E2V
- Manchester OWL Syntax

Criteria:

- Universal quantification over individuals (possibly limited)
- Relations of arity greater than 1 (e.g., binary relations)
- Are not E^3 -languages

Languages with medium expressiveness (E^3)

Examples:

- Attempto Controlled English

Criteria:

- **General rule structures:** *if-then* statements with multiple universal quantification that can target all argument positions of relations
- **Negation** (strong negation or negation as failure)
- Have all features of E^2
- Are not E^4 -languages

Languages with high expressiveness (E^4)

Examples:

- SBVR Structured English
- Formalized English

Criteria:

- General **second-order universal quantification** over concepts and relations
- Have all features of E^3
- Are not E^5 -languages

Languages with maximal expressiveness (E^5)

Examples:

- Basic English
- E-Prime
- Caterpillar Fundamental English / ASD-STE
- All natural languages

Criteria:

- Can **express anything** that can be communicated between two human beings
- Cover any statement in any type of logic

Naturalness

The dimension of naturalness describes **how close the language is to a natural language** in terms of readability and understandability to speakers of the given natural language.

- Unnatural languages (N^1)
- Languages with dominant unnatural elements (N^2)
- Languages with dominant natural elements (N^3)
- Languages with natural sentences (N^4)
- Languages with natural texts (N^5)

Unnatural languages (N¹)

Examples:

- Propositional logic

Criteria:

- Languages that do not look natural
- Heavy use of symbol characters, brackets, or unnatural keywords
- Use of natural words or phrases as names for certain entities might be possible, but is neither required nor further defined

These are not CNLs according to our definition.

Languages with dominant unnatural elements (N²)

Examples:

- Manchester OWL Syntax

Criteria:

- Natural language words or phrases are an integral part
- Dominated by unnatural elements or unnatural statement structure
- Natural elements do not connect in a natural way to each other
- Untrained readers fail to intuitively understand the statements

These are not CNLs according to our definition.

Languages with dominant natural elements (N³)

Examples:

- FAA Air Traffic Control Phraseology
- Formalized-English

Criteria:

- Natural elements are dominant over unnatural ones
- General structure corresponds to natural language grammar
- Sentences cannot be considered valid natural sentences
- Untrained readers do not recognize the statements as well-formed sentences of their language, but are nevertheless able to intuitively understand them to a substantial degree

Languages with natural sentences (N⁴)

Examples:

- Syllogisms
- Standard Language
- SBVR Structured English
- Attempto Controlled English
- Drafter Language
- E2V

Criteria:

- Valid natural sentences
- If natural flow is maintained, **minor deviations** are permitted, including text color, indentation, hyphenation, and capitalization
- **Untrained readers recognize the statements as sentences of their language** and are able to correctly understand their essence
- Single sentences have a natural flow, but not complete texts

Languages with natural texts (N⁵)

Examples:

- Basic English
- E-Prime
- Caterpillar Fundamental English / ASD-STE
- All natural languages

Criteria:

- **Complete texts and documents** can be written in a natural style and with a natural text flow
- For spoken languages, **complete dialogs** can be produced with a natural flow and a natural combination of speech acts

Simplicity

The fourth dimension is a measure of the **simplicity or complexity of an exact and comprehensive language description** covering syntax and semantics (without presupposing intuitive knowledge about any natural language), if such a complete description is possible at all.

Indicator for simplicity: the **number of pages** needed to describe the language in an exact and comprehensive way.

- Very complex languages (S^1)
- Languages without exhaustive descriptions (S^2)
- Languages with lengthy descriptions (S^3)
- Languages with short descriptions (S^4)
- Languages with very short descriptions (S^5)

Very complex languages (S^1)

Examples:

- Basic English
- E-Prime
- Caterpillar Fundamental English / ASD-STE
- All natural languages

Criteria:

- Have the complexity of natural languages
- Cannot be described in an exact and comprehensive manner

Languages without exhaustive descriptions (S²)

Examples:

- FAA Air Traffic Control Phraseology
- Standard Language
- SBVR Structured English

Criteria:

- Considerably simpler than natural languages
- A significant part of the complex structures are eliminated or heavily restricted
- Too complex to be described in an exact and comprehensive manner
- Usually described by restrictions on a given natural language

Languages with lengthy descriptions (S^3)

Examples:

- Attempto Controlled English
- Drafter Language
- Formalized-English

Criteria:

- Can be defined in an exact and comprehensive manner
- Requires more than ten pages

Languages with short descriptions (S⁴)

Examples:

- E2V
- Manchester OWL Syntax

Criteria:

- Exact and comprehensive description requires more than one page but not more than ten pages

Languages with very short descriptions (S^5)

Examples:

- Syllogisms
- Propositional logic

Criteria:

- Can be described in an exact and comprehensive manner on a single page

CNL Implementations

CNL parsers/checkers can be implemented in a number of programming languages or frameworks:

- **Unification grammars in Prolog**: very powerful and general
- **Parser generator languages** (yacc, GNU Bison, etc.): optimized for programming languages
- **Grammatical Framework (GF)**: optimized for natural languages and translation
- **Codeco**: optimized for CNLs with predictive editors and non-local dependencies like anaphoric references
- **Other general-purpose programming languages** (Java, Python, etc.)

Part 4: Applications

In what applications are CNLs used?

Application Areas

- Semantic Web
- Technical Documentation
- General-Purpose Knowledge Representation
- Personal Rules and Scripts
- Emergency Instructions
- Query Interfaces
- International Communication
- Mathematical Texts
- Software Specifications
- Legislation/Government Documents
- Policies / Business Rules

Application Area: Semantic Web

Languages:

- Ginseng's Guided English
- AIDA
- ClearTalk
- Controlled Language for Ontology Editing (CLOnE)
- Rabbit
- OWL ACE
- OWL Simplified English
- *and several others*

Ginseng's Guided English

P⁵E¹N⁴S³ — f w a

- Query language to access **OWL ontologies**
- Vocabulary is loaded from the respective ontologies
- **120 static grammar rules** plus additional dynamic rules generated from the ontologies
- **Predictive editing approach**

Ask a question: question: question: question: question: question: question: question:

what is th is the h the he height of m t of mount f mount m f mount ma

the there thousand hamilton hammond hampton harney harrisburg hartford harvard hawaii haward height helena magazine maroon massive mauna mckinley mississippi mount curwood davis elbert frissell greylock hood katahdin mansfield marrv mansfield marcy mckinley mitchell mansfield marcy mansfield

Enter text: Enter text of: Enter text of: Enter text of: Enter text of: Enter text of: Enter text of:

or select fr or select fr or select fr or select fr or select fr or select fr or select fr

from pop from pop from pop from pop from pop from pop from pop

up menu.

Ginseng's Guided English: Screenshot

Ginseng v0.86

File

ginseng
guided input natural language search engine

Ask a question:
what are the capitals of the states that border

Clear Go!

Enter text or select from pop-up menu.

SELECT ?what WHERE (<http://www.mooney.net/geo#highElev
"6194"
SELECT ?what WHERE (<http://www.mooney.net/geo#lowElev
SELECT ?what WHERE (<http://www.mooney.net/geo#height>
SELECT ?what WHERE (?instance <http://www.mooney.net/geo#mo
"6194"
SELECT ?what WHERE (?instance <http://www.mooney.net/geo#mo
SELECT ?what WHERE (?instance <http://www.mooney.net/geo#mo

alabama
alaska
america
anyone
anything
anywhere
arizona
arkansas
california
colorado
connecticut
delaware
district
florida
georgia
hawaii
idaho

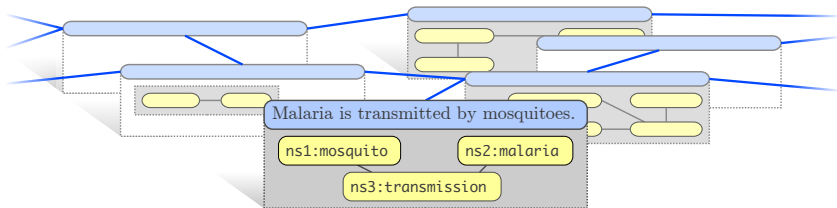
Resources Properties

- http://www.mooney.net/geo#
- river
- lake
- high point
- canton
- state
- mountain
- capital
- low point
- road
- city
- country
- lo point
- hi point

AIDA

P²E⁵N⁴S¹ — f w a

- For informal and underspecified representations of **scientific assertions** in “nanopublications”
- Single English sentences as a scaffold for organizing scientific discoveries and discourse
- **Atomic, Independent, Declarative, and Absolute**



Example

The degree of hepatic reticuloendothelial function impairment does not differ between cirrhotic patients with and without previous history of SBP.

ClearTalk

P³E³N³S³ — f w a

- Documents can be “almost automatically” translated into a formal logic notation and into other natural languages
- It “offers a flexible degree of formality” that lets an author choose to “leave or remove ambiguity”
- *Syntactic restrictions*, e.g. basic sentences have the general form *subject predicate complement modifier-phrases*
- *Semantic restrictions*, e.g. the determiner *a* at subject position represents universal quantification

Examples

Any adverb that modifies a verb must be adjacent to (that verb or another adverb).

Mary hopes that [- Bill loves her -].

Controlled Language for Ontology Editing (CLOnE)

P⁵E²N⁴S⁴ — f w a

- Front-end language for OWL, covering only a small subset
- Defined by ten basic sentence patterns
- Adds procedural semantics on top of OWL for introducing and removing entities and axioms

Examples

Persons are authors of documents.

Carl Pollard and Ivan Sag are the authors of 'Head-Driven Phrase Structure Grammar'.

Forget everything.

Rabbit

P⁵E²N⁴S⁴ — f w g

- Controlled natural language for OWL
- Developed and used by Ordnance Survey, Great Britain's national mapping agency
- Designed for a *specific scenario* for the communication between domain experts and ontology engineers to create ontologies
- *Three types* of statements: declarations, axioms, and import statements

Examples

Sheep is a concept, plural Sheep.

Every River flows into exactly one of River, Lake or Sea.

OWL ACE

P⁵E²N⁴S³ — f w a

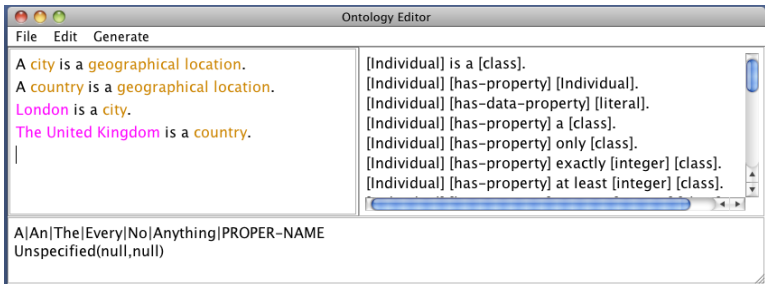
- A subset of ACE that maps to OWL
- Available in the ACE View plugin for Protégé

Snippet	W...	Mes...
Every country that is not bordered by a country is an island-country	4	0
If an island-country is bordered by a country then it is an island-country	4	0
Every territory that is not bordered by a body-of-water is a landlocked-territory	4	0
Every landlocked-territory is a territory that is not bordered by a body-of-water	4	0
Which country is not a NATO-country and borders a NATO-country	4	0
San_Marino is surrounded by Italy.	3	2
Vatican_City is surrounded by Italy.	3	2
Kaliningrad is bordered by Lithuania.	2	0
Kaliningrad is bordered by Poland.	2	0
If there is a NATO-country that is a neutral-country then it is a neutral-NATO-country	3	0
Every neutral-NATO-country is a NATO-country that is a neutral-country.	3	0

OWL Simplified English

P⁵E²N⁴S⁴ — f w a

- No real lexicon, neither built-in nor user-defined
- Only a very small number of predefined function words
- Users have to list the verbs they intend to use
- Other word categories are *inferred* based on syntactic clues such as capitalization and adjacent words



Application Area: Technical Documentation

Languages:

- Caterpillar Fundamental English
- ASD Simplified Technical English
- Standard Language
- Avaya Controlled English
- Caterpillar Technical English
- KANT Controlled English
- NCR Fundamental English
- *and many more*

Avaya Controlled English

P²E⁵N⁵S¹ — c t w d i

- Language for technical publications in the telecommunication and computing industry
- Reduce translation costs and make texts easier to understand
- Lexical restrictions, e.g., “Do not use *abort*”
- Syntactic restrictions, e.g. “Use active voice”
- Semantic restrictions, e.g. “Use *may* only to grant permission”
- Stylistic restrictions, e.g. “Put command names in bold monospaced type”
- Open list of about 250 words defines preferred terminology

Examples

This procedure describes how to connect a dual ACD link to the server.

If the primary server fails, you can use the secondary server.

Caterpillar Technical English

P²E⁵N⁵S¹ — c t w d i

- Second CNL developed at Caterpillar, starting in 1991
- Should **improve consistency and reduce ambiguity** of technical documentation
- Should **improve translation quality** and reduce translation costs with the help of machine translation
- Language checker with **interactive disambiguation**
- About 70,000 terms with a “narrow semantic scope”
- Syntactic restrictions, e.g. concerning the use of conjunctions, pronouns, and subordinate clauses

Example

This category indicates that an alternator is malfunctioning. If the indicator comes on, drive the machine to a convenient stopping place. Investigate the cause and determine the solution.

KANT Controlled English (KCE)

P²E⁵N⁵S¹ — t w a

- CNL for machine translation
- Lexicon, grammar, and semantics are restricted

Example

Secure the gear with twelve rivets.

- Ambiguity can be resolved by augmenting the sentence with SGML tags:

Example with SGML tags

Secure the gear with <attach head='secure' modi='with'> twelve rivets.

NCR Fundamental English

P²E⁵N⁵S¹ — c w d i

- Language for **technical manuals** of the NCR company
- Should make manuals “easier to read and use by NCR employees and customers around the world”
- **Nomenclature**: open set of names of products, tools, routines, modes, conditions, etc.
- **Glossary**: open set of words for technical concepts, e.g. *audit trail*
- **Vocabulary**: fixed set of 1,350 words (verbs, nouns, adverbs, adjectives, pronouns, prepositions, articles, and conjunctions) plus 650 abbreviations

Examples

While repairing the unit, the field engineer also performs normal maintenance if it is needed.

No maintenance can be performed until the maintenance lock has been activated.

Application Area: General-Purpose Knowledge Representation

Languages:

- Syllogisms
- Attempto Controlled English
- E2V
- Formalized-English
- Common Logic Controlled English (CLCE)
- Computer Processable Language (CPL)
- Controlled English to Logic Translation (CELT)
- Gellish English
- PENG

Common Logic Controlled English (CLCE)

P⁵E³N³S³ — f w a

- CNL with mapping to *first-order logic*
- Defined by a grammar in Backus-Naur form
- *Syntactic restrictions*: no plural nouns, only present tense, and variables instead of pronouns, and more
- *Interpretation rules* for unambiguous mapping to logic
- *Parentheses* to determine the structure of deeply nested sentences

Examples

If some person x is the mother of a person y , then the person y is a child of the person x .

Declare give as verb (agent gives recipient theme) (agent gives theme to recipient) (theme is given recipient by agent) (theme is given to recipient by agent) (recipient is given theme by agent).

Computer Processable Language (CPL)

P³E³N⁴S² — f w i

- Basic CPL sentences: *subject + verb + complements + adjuncts*
- Further syntax restrictions, e.g. definite references instead of pronouns
- Seven templates for complex sentences, e.g. “If sentence1 then typically sentence2”
- Parser translates CPL into a formal logic representation
- Parsing involves external tools and resources (e.g. WordNet)
- Paraphrase for verification or correction by the user

Examples

IF a person is carrying an entity that is inside a room THEN (almost) always the person is in the room.

AFTER a person closes a barrier, (almost) always the barrier is shut.

Computer Processable Language (CPL): Screenshot



Controlled English to Logic Translation (CELT)

P⁴E²N⁴S³ — f w i

- CNL inspired by ACE
- Uses existing linguistic and ontological resources: [SUMO](#) and [WordNet](#)
- Deterministic syntax structure
- [Heuristics](#) for mapping the words to SUMO and WordNet
- Implemented as a [unification grammar](#) in Prolog

Examples

Dickens writes Oliver Twist in 1837.

Every boy likes fudge.

Gellish English

P⁴E²N⁴S³ — f w a i

- Common **data language** for industry
- Simple **subject–predicate–object** structures
- **Predefined relations** in the form of fixed phrases, e.g. “is a specialization of” and “is valid in the context of”
- **Fixed upper ontology** with a large number of predefined concepts and relation types
- Texts in Gellish can be transformed into a **formal tabular representation**

Examples

collection C each of which elements is a specialization of animal

the Eiffel tower has aspect h1

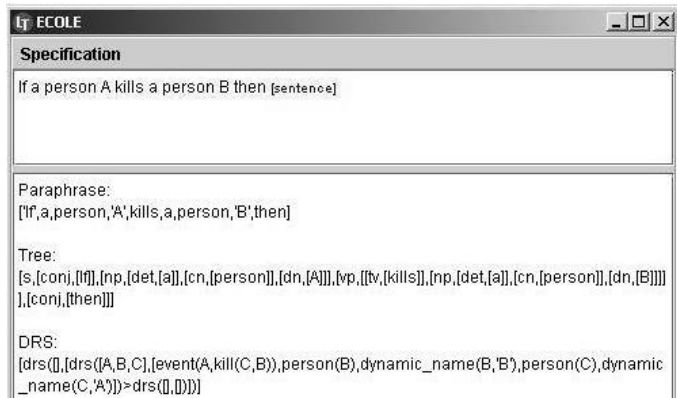
h1 is classified as a height

h1 is qualified as 300 m

PENG

P⁵E³N⁴S³ — f w a

- “Processable English,” inspired by ACE
- Rich but unambiguous language with mapping first-order logic
- Focus on predictive editing



The screenshot shows a window titled "ECOLE" with a "Specification" section containing the sentence "If a person A kills a person B then [sentence]". Below this, the "Paraphrase:" section shows the sentence with placeholders: "[If, a, person, 'A', kills, a, person, 'B', then]". The "Tree:" section shows a complex nested list structure representing the sentence's syntax. The "DRS:" section shows a lambda expression representing the Discourse Representation Structure for the sentence.

```
Specification
If a person A kills a person B then [sentence]

Paraphrase:
[If, a, person, 'A', kills, a, person, 'B', then]

Tree:
[s, [conj, [If], [np, [det, [a]], [cn, [person]], [dn, [A]]], [vp, [[tv, [kills]], [np, [det, [a]], [cn, [person]], [dn, [B]]]], [conj, [then]]]]

DRS:
[drs([], [drs([A, B, C], [event(A, kill(C, B)), person(B), dynamic_name(B, 'B'), person(C), dynamic_name(C, 'A')])>drs([], [])])]
```

Application Area: Personal Rules and Scripts

Languages:

- Drafter Language
- Atomate Language
- Voice Actions
- iLastic Controlled English

Atomate Language

P⁴E²N⁴S³ — f w d a

The screenshot shows a web browser window titled "atomate | personal information engine". The browser address bar shows "atomate | personal information en...". The page has a navigation bar with "my status", "my actions", "my knowledge", and "my sources". There are "0 new notifications" and "[help] [+]" links.

The main content area displays a rule configuration interface. At the top, there are buttons for "email me", "alert me", "text me", "tweet", "update FB status", and "set". The "alert me" button is highlighted with a red box.

Below this, the rule is defined: "whenever" (highlighted with a red box) "the next time" "my" (underlined) "current location" (underlined) "is" (underlined) "Home" (underlined) "and" "on / after" (highlighted with a red box) "5:00 PM every Tue" (underlined). A red question mark is next to "5:00 PM every Tue".

A time selection interface is shown below. It includes a "time" label, a row of numbers from 12 to 11, and a row of AM/PM labels. A yellow highlight is under "5:00 PM". Below this is a row of days: "every: weekday Sun Mon Tue Wed Thu Fri Sat". The "Tue" button is highlighted with a red box.

At the bottom, there is a "with message:" label and a text input field containing "Trash day!".

A yellow bar at the bottom of the configuration area contains the text: "alert me whenever my current location is Home at 5 pm every Tue with the message: 'Trash day!'". To the right of this bar is a "save" button.

Below the yellow bar, there are two buttons: "active" (highlighted with a red box) and "not active". To the right of these buttons is the text: "alert me whenever my current location is Home at 5 pm every Tue with the message: 'Trash day!'". To the right of this text are "[edit]" and "X" buttons.


The bottom left corner of the browser window shows "Done".

Voice Actions

P³E¹N⁴S² — f s d i

- CNL for [spoken action commands](#) for Android phones
- Twelve informally defined command patterns

Voice Action commands

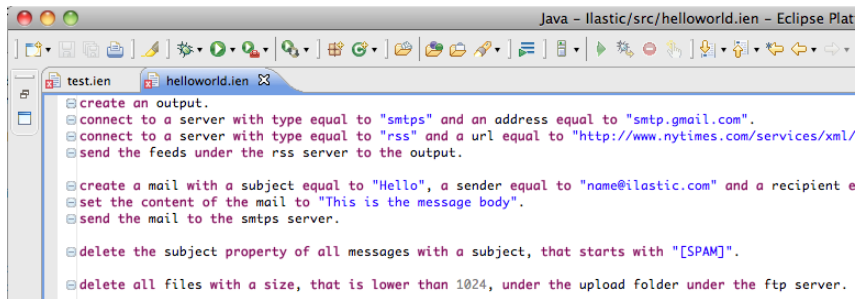
 Voice Actions are only supported in English, French, German, Spanish, and Italian.

Say	Followed by	Examples
"Open"	App name	"Open Gmail"
"Create a calendar event"	"Event description" & "day/date" & "time"	"Create a calendar event: Dinner in San Francisco, Saturday at 7:00PM"
"Map of"	Address, name, business name, type of business, or other location	"Map of Golden Gate Park, San Francisco."
		"Directions to 1200 Colusa Avenue"

iLastic Controlled English

P⁵E³N⁴S³ — f w i

- Language for non-developers to write intuitive and natural scripts
- Automatic retrieval, transformation, and combination of data from the web, databases, files, and other resources



The screenshot shows an Eclipse IDE window titled "Java - ilastic/src/helloworld.iel - Eclipse Plat". The editor displays the following iLastic script code:

```
create an output.
connect to a server with type equal to "smtps" and an address equal to "smtp.gmail.com".
connect to a server with type equal to "rss" and a url equal to "http://www.nytimes.com/services/xml/".
send the feeds under the rss server to the output.

create a mail with a subject equal to "Hello", a sender equal to "name@ilastic.com" and a recipient e
set the content of the mail to "This is the message body".
send the mail to the smtps server.

delete the subject property of all messages with a subject, that starts with "[SPAM]".

delete all files with a size, that is lower than 1024, under the upload folder under the ftp server.
```

Application Area: Emergency Instructions

Language:

- Controlled Language for Crisis Management (CLCM)

Controlled Language for Crisis Management (CLCM)

P²E⁵N⁵S¹ — c t w d a

- CNL for instructions on how to deal with **crisis situations**
- About 80 simplification rules on ...
- **Text structure**, e.g. “Write a title for every specific situation”
- **Formatting**, e.g. “Separate with a new line each block of instructions”
- **Lexicon**, e.g. “Avoid technical terms”
- **Syntax**, e.g. “Avoid passive voice”
- **Semantics**, e.g. “Use only literal meaning”
- **Pragmatics**, e.g. “Remove unimportant information”

Application Area: Query Interfaces

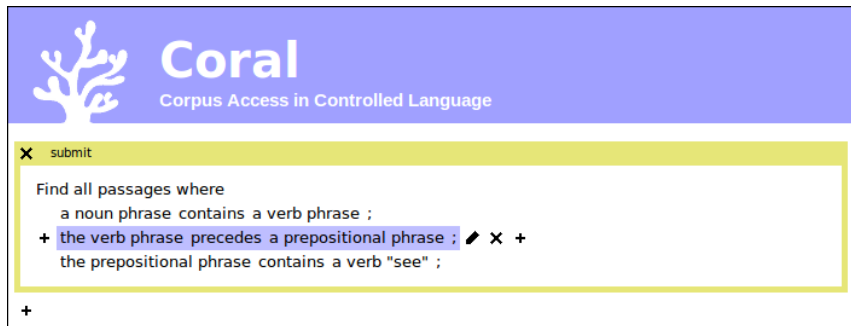
Languages:

- Coral's Controlled English
- Quelo Controlled English
- BioQuery-CNL
- CLEF Query Language

Coral's Controlled English

P⁵E¹N⁴S⁴ — f w d a

- CNL for formal queries to annotated text corpora
- Influenced by ACE, but simpler and much less expressive
- Query interface for users with no computing background



The screenshot shows the Coral interface with a blue header containing a coral logo and the text "Coral Corpus Access in Controlled Language". Below the header is a yellow search bar with a "submit" button. The search query is: "Find all passages where a noun phrase contains a verb phrase ; + the verb phrase precedes a prepositional phrase ; the prepositional phrase contains a verb 'see' ;". The second part of the query is highlighted in blue. At the bottom left of the search area is a "+" sign.

Quelo Controlled English

P⁴E¹N⁴S⁴ — f w a

- CNL used in a query interface called Quelo
- **Conceptual authoring approach:** users cannot directly edit the sentences, but they can trigger modification actions on the underlying formal representation

I am looking for something. It should be equipped with an automatic transmission system and sold by a car dealer. The car dealer should sell a fleet car.

BioQuery-CNL

P⁵E²N⁴S⁴ — f w d a

- CNL for **biomedical queries**
- Query engine based on **answer set programming**
- Initially designed as a subset of ACE with some small modifications handled in a preprocessing step
- Evolved into an **independent language with its own parser**

Example

What are the genes that are targeted by all the drugs that belong to the category Hmg-coa reductase?

CLEF Query Language

P⁵E¹N⁴S³ — f w d a

- CNL used within a system called CLEF (Clinical E-Science Framework)
- Should help clinicians, medical researchers, and hospital administrators to query electronic health records
- Conceptual authoring approach, influenced by Drafter language
- Basic queries are composed of: the set of relevant patients, the received treatments, and the outcomes
- Complex queries with multiple elements of the same type
- Translated to SQL and given to a database engine

Example

For all patients with cancer of the pancreas, what is the percentage alive at five years for those who had a course of gemcitabine?

Application Area: International Communication

Languages:

- Basic English
- FAA Air Traffic Control Phraseology
- Special English
- PoliceSpeak
- SEASPEAK
- EasyEnglish (by Wycliffe Associates)

Special English

P¹E⁵N⁵S¹ — c w s g

- CNL developed and used by the **Voice of America (VOA)**, the official external broadcast institution of the US government
- Used **since 1959 until today** for news on radio, television, and the web
- Second oldest English-based CNL (after Basic English) and the only one that has been in use for such a long period by the same organization
- Vocabulary is restricted to about **1,500 words**, which have changed over time
- Short sentences and should be spoken at a slower speed
- **No explicit restrictions on grammar or semantics**

Special English: Word Book

aid – *v.* to help; to support; *n.* help, assistance

aim – *v.* to point a gun at;
n. a goal or purpose

air – *n.* the mixture of gases
around the earth, mostly
nitrogen and oxygen,
that we breathe

air force – *n.* a military
organization
using airplanes

airplane – *n.* a vehicle with
wings that flies

airport – *n.* a place where
airplanes take off and land

album – *n.* a collection of recorded music



airplane

PoliceSpeak

P²E¹N³S² — c s d g

- CNL to improve police communications of English and French officers at the Channel Tunnel
- Goal: “make police communications more concise, more predictable, more stable and less ambiguous”
- Launched in 1988 and the language was ready in 1992

SEASPEAK

P²E¹N³S² — c s d g

- “International Maritime English”
- For clear communication among ships and harbors
- Development started in 1981
- Similar goal and application area as SEASPEAK and air traffic control phraseologies

EasyEnglish (by Wycliffe Associates)

P²E⁵N⁵S¹ — c t w d

- Not to be confused with IBM's EasyEnglish!
- CNL for transcribing **biblical texts** to improve translation for readers with limited knowledge of English
- Restricted with respect to lexicon, syntax, and semantics
- 1,200 words (level A) and 2,800 words (level B)
- E.g. *fair* can only mean *unbiased*; *to see* cannot be used in the sense *to meet*
- Other words need **explanations in separate EasyEnglish sentences**
- **Strict sentence length limit** of 20 words

Example

The Highlands of Scotland consist of lakes, mountains and moors. The moors are flat empty lands where no trees grow. This land is wonderful and magnificent because it is so empty.

Application Area: Mathematical Texts

Languages:

- ForTheL
- Naproche CNL
- Controlled Language of Mathematics (CLM)

ForTheL

P⁵E³N³S³ — f w d a

- “Formal Theory Language”
- CNL for **mathematical texts**
- Can be automatically translated into **first-order logic**

Example

Definition 4. Let A, B be sets.

A is a subset of B ($A \subseteq B$) IFF all elements of A belong to B .

Lemma 1. Each set has a subset.

Proof. \emptyset is a subset of all sets. QED.

Naproche CNL

P⁵E³N³S³ — f w d a

- Another CNL for mathematical texts with a deterministic mapping to first-order logic
- Automatic checking for logical correctness

Theorem: $\sqrt{2}$ is irrational.

Proof:

Assume that $\sqrt{2}$ is rational. Then there are integers a, b such that $a^2 = 2 \cdot b^2$ and $\gcd(a, b) = 1$. Hence a^2 is even, and therefore a is even. So there is an integer c such that $a = 2 \cdot c$. Then $4 \cdot c^2 = 2 \cdot b^2$, $2 \cdot c^2 = b^2$, and b is even. Contradiction. Qed.

Application Area: Software Specifications

Languages:

- Template Based Natural Language Specification (TBNLS)
- Gherkin

Template Based Natural Language Specification (TBNLS)

P⁵E²N³S⁴ — f w d a i

- CNL for testing control software for passenger vehicles
- Defined by 15 templates
- Mapping to propositional logic with temporal relations

Example

If Button B₄ is down P₁ occurs, then Lamp L₃ is red P₂
hold immediately, until 10 seconds T₁ elapsed.

Gherkin

P⁵E³N⁴S³ — f w d a

- CNL for writing **executable scenarios** for software specifications
- Fixed structuring words such as *Given*, *And*, and *But*
- Restrictions on remaining text in **ordinary programming languages using regular expressions** (“step definitions”)
- Concrete step definitions are not part of Gherkin, but **have to be implemented for the particular task at hand**
- Highly customizable and extensible

Example

Scenario: Unsuccessful registration due to full course

Given I am a student

And a lecture “PA042” with limited capacity of 20 students

But the capacity of this course is full

[...]

Gherkin: Editor

```
1 # This is just a demo of syntax highlighting
2 # And very basic syntax checking (based on gherkin.js)
3 # Try to introduce a syntax error (Replace When with a different
4 # string). This will colour that line red.
5 Feature: Hello World
6   Scenario: Look Ma
7     Given I am in a browser
8     When I make a syntax error
9     Then stuff should be red
10
11 @tags @are @handy
12 Scenario: Look Pa
13   Given I have|
14   When I have (\d+) cukes in my belly
15   | a I have eaten all the cukes
16   | yes | nice |
17   Then stuff should be red
```

Application Area: Legislation/Government Documents

Languages:

- Plain Language (or Plain English)
- Massachusetts Legislative Drafting Language

Plain Language (or Plain English)

P¹E⁵N⁵S¹ — c w g

- Initiative by the **US government** and other organizations
- Origins in the 1970s
- **Goal**: make official documents easier to understand and less bureaucratic
- Examples of rules:
 - “Use pronouns to speak directly to readers”
 - “Avoid double negatives and exceptions to exceptions”
- Many of the guideline rules are **strict**
- With the Plain Writing Act of 2010, US governmental agencies are **obliged to comply** with these restrictions

Massachusetts Legislative Drafting Language

P²E⁵N⁵S¹ — c w d g

- CNL for **legal texts** defined by the Massachusetts Senate
- “to promote uniformity in drafting style, and to make the resulting statutes clear, simple and easy to understand and use”
- Defined by about 100 rules
- **Restricted syntax**, e.g. “Use the present tense and the indicative mood”
- **Restricted semantics**, e.g. “Do not use ‘deem’ for ‘consider’”
- **Restricted document structure**, e.g. “Use short sections or subsections”
- Close to 90 words and phrases that must not be used, with suggested replacements, e.g. *hide* instead of *conceal*, and *rest* instead of *remainder*

Application Area: Policies / Business Rules

Languages:

- SBVR Structured English
- RuleSpeak
- PERMIS Controlled Natural Language

RuleSpeak

P³E⁴N⁴S² — c f w i

- CNL for **business rules**
- Development started in 1985 and was first presented in 1994
- Very similar to **SBVR Structured English**
- Each rule belongs to one of **eleven “functional categories”** such as “computation rule,” “inference rule,” and “process trigger”
- Specific **templates** for each category, e.g. computation rules contain the phrase “must be computed as” or “=”

Example

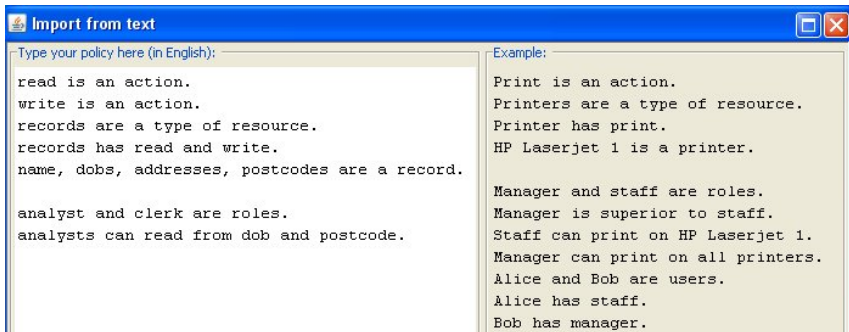
An order may be accepted only if all of the following are true:

- It includes at least one item.
- It indicates the customer who is placing it.

PERMIS Controlled Natural Language

P⁵E²N⁴S⁴ — f w d a

- CNL for [access control policies](#) for grid computing environments
- Based on [CLOnE](#) with extensions for authorization policies
- Mapping to different formal target notations
- [Nine statement patterns](#)



Import from text

Type your policy here (in English):

```
read is an action.  
write is an action.  
records are a type of resource.  
records has read and write.  
name, dobs, addresses, postcodes are a record.  
  
analyst and clerk are roles.  
analysts can read from dob and postcode.
```

Example:

```
Print is an action.  
Printers are a type of resource.  
Printer has print.  
HP Laserjet 1 is a printer.  
  
Manager and staff are roles.  
Manager is superior to staff.  
Staff can print on HP Laserjet 1.  
Manager can print on all printers.  
Alice and Bob are users.  
Alice has staff.  
Bob has manager.
```

Part 5: Analysis

What does the big picture of existing CNLs look like?

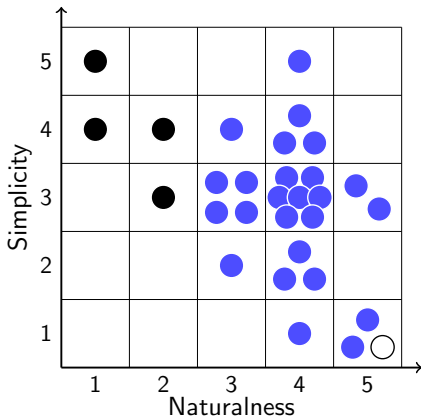
Analysis

We can now analyze the collected data:

- What **inherent properties** do existing CNLs have?
- In what **environments** are existing CNLs used?
- What does the **timeline** of CNL evolution look like?

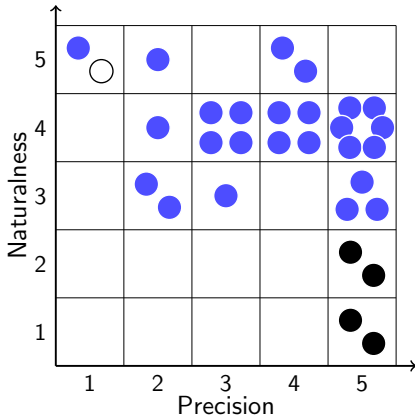
Naturalness vs. Simplicity

PENS classes of CNLs (blue) in comparison to natural (white) and formal (black) languages:



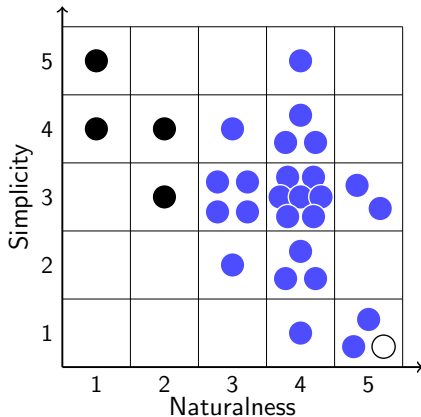
Precision vs. Naturalness

CNLs (blue), natural (white) and formal (black) languages:



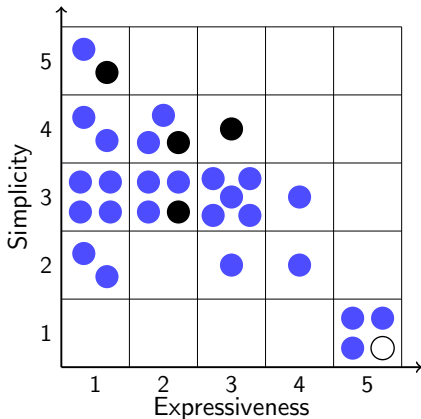
Naturalness vs. Simplicity

CNLs (blue), natural (white) and formal (black) languages:



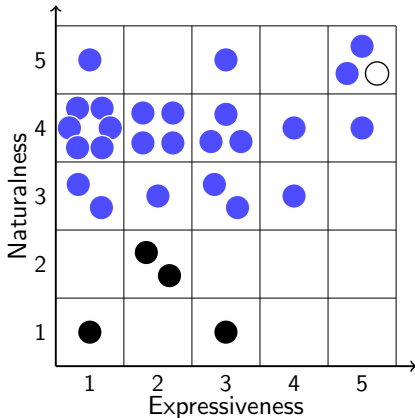
Expressiveness vs. Simplicity

CNLs (blue), natural (white) and formal (black) languages:



Expressiveness vs. Naturalness

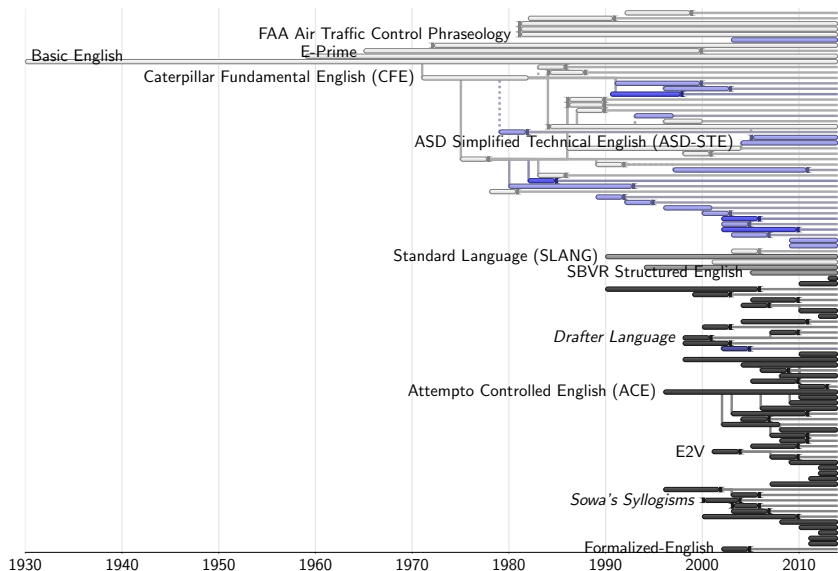
CNLs (blue), natural (white) and formal (black) languages:



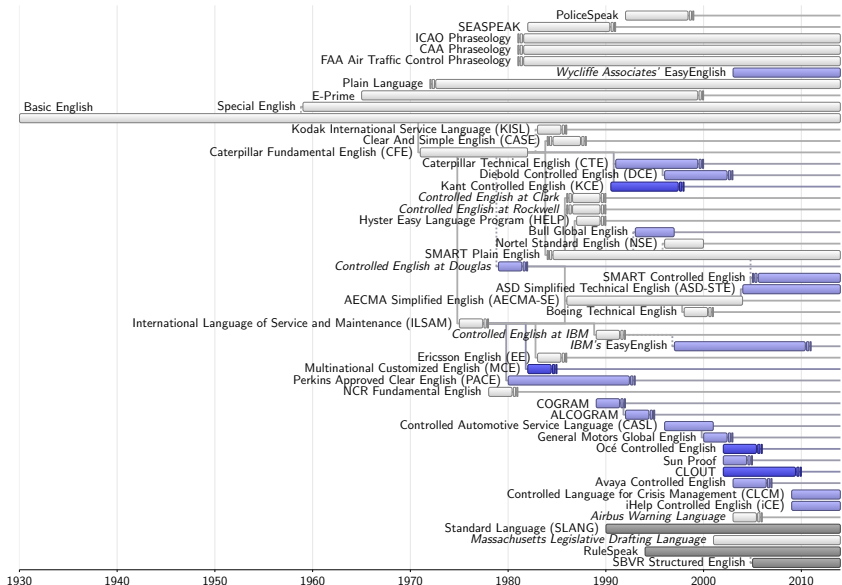
CNL Environment Properties

property	total	combined with property									PENS average			
		c	t	f	w	s	d	a	i	g	P	E	N	S
c comprehensibility	45	-	17	3	40	6	33	4	33	8	2.0	4.3	4.7	1.2
t translation	22	17	-	1	21	0	17	5	18	0	2.0	4.8	5.0	1.1
f formal representation	54	3	1	-	52	1	19	45	10	2	4.4	2.3	3.8	3.2
w written	93	40	21	52	-	1	46	49	42	5	3.3	3.5	4.3	2.3
s spoken	7	6	0	1	1	-	6	0	1	6	2.0	1.6	3.4	1.9
d domain-specific	53	33	17	19	46	6	-	20	29	6	2.8	3.5	4.4	1.9
a academia	50	4	5	45	49	0	20	-	4	1	4.3	2.5	3.9	3.1
i industry	43	33	18	10	42	1	29	4	-	0	2.3	4.3	4.7	1.4
g government	10	8	0	2	5	6	6	1	0	-	2.4	2.5	3.8	2.0

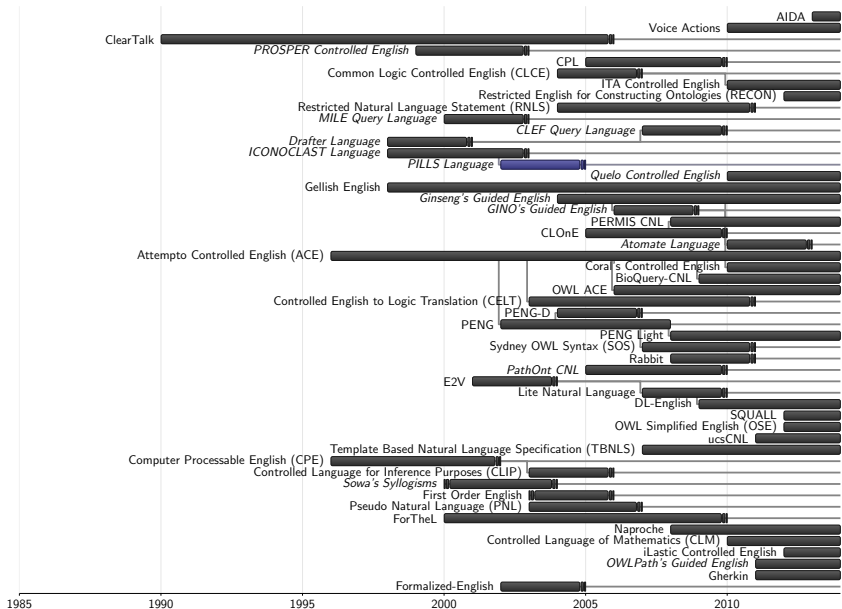
CNL Timeline



Detailed Timeline, Part 1



Detailed Timeline, Part 2



Part 6: Evaluations

Do CNLs actually work?

Evaluations

Research questions for CNL evaluation:

- c** Does a CNL make communication among humans more precise and more effective?
- t** Does a CNL reduce overall translation costs at a given level of quality?
- f** Does a CNL make it easier for people to use and understand logic formalisms?

Evaluations: Type c

- Two studies on **AECMA-SE** showed that the use of controlled English significantly **improves text comprehension**, with a particularly large effect for **complex texts and non-native speakers**
- **CLCM** has been found to have a positive effect on **reading comprehension** for most groups of readers under certain circumstances such as **stress situations**

Evaluations: Type t

- **Multinational Customized English (MCE)** for machine-assisted translation leads to a “five-to-one gain in translation time”
- With **Perkins Approved Clear English (PACE)**, post-editing of machine-assisted translation is “three or four times faster” than without
- Adherence to typical CNL rules improves **post editing productivity and machine translation quality**
- **CLCM** texts are **easier to translate** than uncontrolled ones and the **time needed for post-editing is reduced** on average by 20%

Evaluations: Type f

Two types of studies:

- Studies that test the general **usability** of CNL tools
- Studies that specifically evaluate the **comprehensibility** of the actual languages

Type f Evaluations: Usability

- Study has shown that the **CLOnE** interface is **more usable** than a common ontology editor
- **Coral's controlled English** has been shown to be **easier to use** than a comparable common query interface
- **Positive usability results** have also been reported for: **GINO** (similar to Ginseng), **CLEF**, **CPL**, **PERMIS**, **Rabbit**, and **ACE**

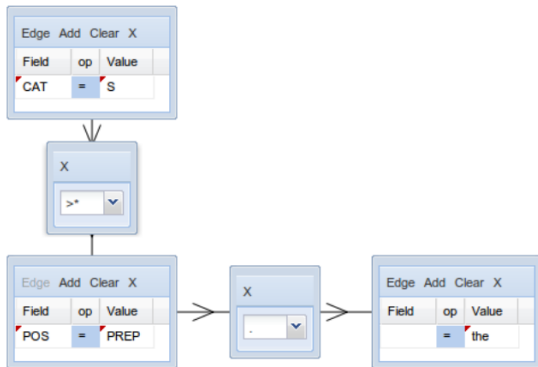
Usability Evaluation: Coral

Coral:

✕

Find all passages where
a sentence contains a preposition that is directly followed by "the" ;

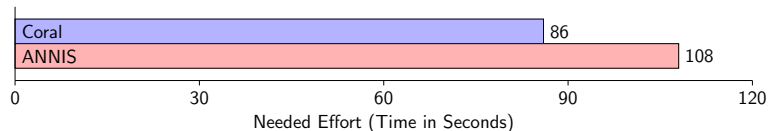
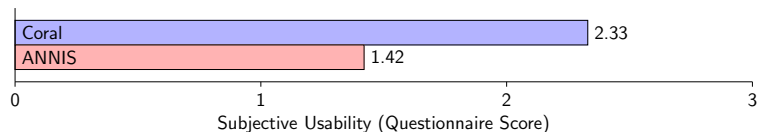
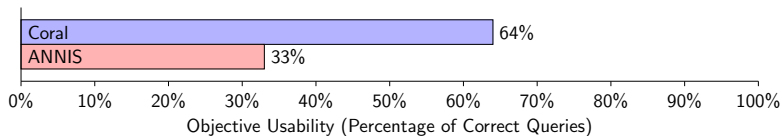
ANNIS (graphical):



ANNIS (AQL code):

CAT="S" & POS="PREP" & "the" & #1 >* #2 & #2 . #3

Coral Usability Evaluation: Results



Type f Evaluations: Comprehensibility

- Has been shown for **CLEF** that common users are able to **correctly interpret** given statements
- **ACE** has been shown to be **easier and faster to understand** than a common ontology notation
- Experiments on the **Rabbit** language gave **mixed results**

Comprehensibility Evaluation: ACE

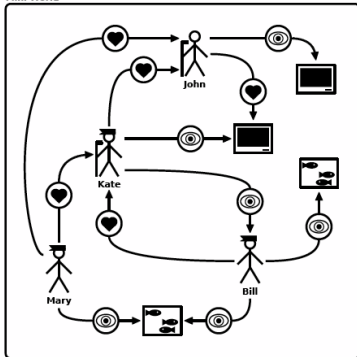
OWL (Manchester syntax)

Bob **HasType** developer
 developer **SubTypeOf** professional
 developer **SubTypeOf** owns **some** cup
 Bob **HasType** owns **some** (**not** cup)
 loves **SubRelationOf** likes

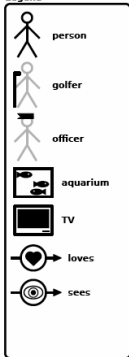
ACE

Bob is a developer.
 Every developer is a professional.
 Every developer owns a cup
 Bob owns something that is not a cup.
 If X loves Y then X likes Y.

Mini World



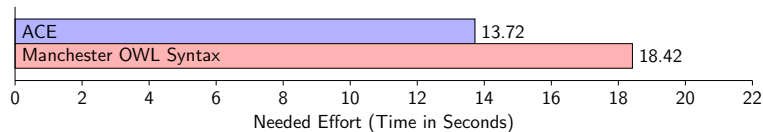
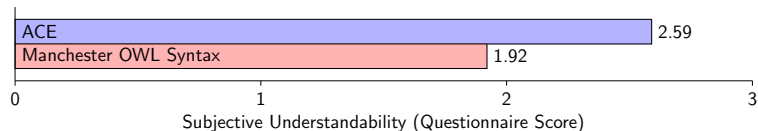
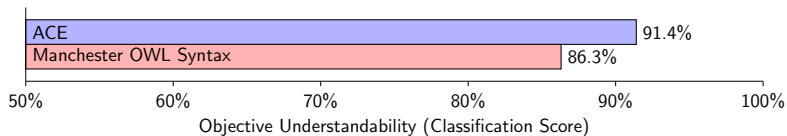
Legend



Which of the statements are true and which are false?

- | true | false | don't know | |
|-----------------------|-----------------------|-----------------------|---|
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Every golfer sees a TV. |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Kate sees no officer. |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Every officer loves nothing but golfers. |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Bill does not love Kate. |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Kate sees nothing but officers. |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | John loves something that is not a TV. |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Everything that sees nothing but aquariums is an officer. |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Kate loves an officer. |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Kate sees Bill. |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Everything that loves a person is a golfer. |

ACE Comprehensibility Evaluation: Results



Part 7: Standardization

What are the opportunities for Standardization?

Standardization: Discussion

Standardization of:

- Particular CNLs
- Properties of CNLs
- CNL classification scheme
- CNL interfaces
- Implementation of CNLs
- Evaluation techniques for CNL



Thank you for your Attention!

Questions?