On Controlled Natural Languages: Properties and Prospects

Adam Wyner, Krasimir Angelov, Guntis Barzdins, Danica Damljanovic, Brian Davis, Norbert Fuchs, Stefan Hoefler, Ken Jones, Kaarel Kaljurand, Tobias Kuhn, Martin Luts, Jonathan Pool, Mike Rosner, Rolf Schwitter, and John Sowa[**]

Multiple Institutes

Abstract. This collaborative report highlights the properties and prospects of Controlled Natural Languages (CNLs). The report poses a range of questions concerning the goals of the CNL, the design, the linguistic aspects, the relationships and evaluation of CNLs, and the application tools. In posing the questions, the report attempts to structure the field of CNLs and to encourage further systematic discussion by researchers and developers.

1 Introduction

Controlled Natural Languages (CNLs) are engineered languages which use a selection of the vocabulary, morphological forms, grammatical constructions, semantic interpretations, and pragmatics which are found in a natural language such as English. To date, more than 40 CNLs have been defined, covering English, Esperanto, French, German, Greek, Japanese, Mandarin, Spanish, and Swedish [1]. They facilitate human-human communication (e.g. translation [2] or technical documentation [3]) and human-machine communication (e.g. interfaces with databases [4] or automated inference engines [5]). For example, in the case of human-machine communication, CNLs enable casual users to use formal query languages such as SQL or SPARQL. Given that CNLs are derived from a natural language, a wider range of speakers of the language find it easier to learn, use, read, and write in comparison to other engineered languages such as computer programming languages, formal languages, or international auxiliary languages [6]. It is an exciting time to work on CNLs, which are becoming more sophisticated, useful, and widespread in both academic research and business applications in a range of areas such as aerospace, manufacturing, oil exploration, business rules, public administration, medicine, and biology. CNLs appear to be particularly significant with respect to information extraction of and reasoning with the content of documents on the internet. Indeed, we are ever closer to prototype systems which realise Leibniz's ambition:

[**] Adam Wyner is the main author, editor, and contact person. Correspondence: Adam Wyner, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom, adam@wyner.info. This report is the outcome of a collaboration amongst the listed contributors; see Section 4 for further contact information. It is based on an original collaborative document:
http://docs.google.com/Doc?id=dd3zb82w_03976bbfm.

> The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate, without further ado, to see who is right. [7, p. 51]

In this short report, which follows from the Workshop on Controlled Natural Languages CNL 2009 (8-10 June 2009, Marettimo Island, Italy)[1] and subsequent community discussion, we outline some of the key *properties* that motivate and guide the design of particular CNLs as well as relate one CNL to another. We highlight key areas for future development and deployment. In this respect, our report draws attention to the interrelations and cohesion among CNLs in contrast to [8], which emphasises the divergences. This report is intended to stimulte further discussion among developers of CNLs and development of the languages.

Our intention is not to define CNLs in general (necessary and sufficient conditions), or to distinguish them from sublanguages or language fragments, or to classify current CNLs. The report is also not a review of the extensive literature (see the links and references on Controlled Natural Languages).[2] By the same token, we have not strictly distinguished "natural" from "engineered" languages, a distinction which relies on assumptions: reading and writing are not "natural" since they are highly engineered and require explicit instruction; first language acquisition cannot occur in the absence of linguistic interaction with speakers of the language; and there is a spectrum of natural languages which are related to a "dominant" language such as creoles, dialects, sets of search terms, early child language, sign language, and so on. Furthermore, within linguistics, there are significant debates concerning, for example, the definition of "word" [9], syntactic grammaticality judgements, and the semantic or pragmatic interpretations of a given sentence [10]. Rather than engage in these issues, our view is that the relations among CNLs arise from a range of properties (some of which have degrees) which can be used to identify a particular CNL in a multidimensional conceptual grid, when compared with other CNLs; in this, one might see specification of the properties and relations as a move towards an *ontology* of engineered languages.

In the following, we first discuss a range of properties of CNLs, then several topics for future development, followed by a link to a wiki which provides a forum for further collaborative discussions on the state of the art and for future developments.

2 The Properties

In this section, we present questions about properties of CNLs. To answer a question is to ascribe a property to the CNL under discussion; we can then enumerate the properties of the language. With these properties, we can relate one CNL to another or relate the CNL to the natural language from which it is derived.

We have distinguished three broad sections from more generic, to design guidelines, and then to more specifically linguistic properties. We do not presume to have fully determined all the properties and the ways they vary; it is for future work to spell this out in detail.

[1] `http://attempto.ifi.uzh.ch/site/cnl2009`

[2] `http://sites.google.com/site/controllednaturallanguage/`

2.1 Generic Properties

The generic properties relate to high level aspects of the language. While there are overlaps, they are different ways to draw out the specification of the language.

- Who are the intended users? For example, there might be a narrow user base with a specialised application (say cancer researchers) or a broad user base with a generalised application (newspaper comment support).
- What are the purposes? Among other purposes, we may have simplification, standardisation, accessibility, input support, specification of procedural information, information extraction, translation, generation, and inference. Purposes might be tied to specific tasks such as knowledge acquisition, ontology authoring, querying. Depending on the purpose, the language may or may not be translated into a logic formalism or programming language.
- Is the language domain dependent or independent?

The users, purposes, and domain of a CNL determine the requirements which are the design properties that the CNL needs to meet. We discuss requirements in the next subsection.

2.2 Design Properties

The design properties present more specific indicators of the requirements. For several of the properties, usability tests may be conducted as the language is iteratively developed; the results of the tests may be used to guide each subsequent iteration towards the goal of the design of the language. Answers to several of the questions are contingent on the users, purposes, and requirements as well as answers to other design questions. Several of the questions relate to a notion of *habitability* [11]; according to [12] and [13], a language is habitable if: 1) users are able to construct expressions of the language that they have not previously encountered, without significant conscious effort; and 2) users are able to avoid easily constructing expressions that fall outside the bounds of the language.

- Is the language easy to describe, teach, and learn? For example, a CNL is very easy to describe (and thus to teach and learn) if its restrictions, with respect to the full language, can be defined by just a few sentences.
- Is the language easy to read and scan? For example, perhaps a construct such as “not at least” should be disallowed because it is hard to read and better alternatives exist (“less than”). On the other hand, a construct such as “a man that a woman that a dog sees likes waits.” might be allowed, even if it is hard to read because to bar it would violate the principle of compositionality. On the other hand, one might use psycholinguistic heuristics to guide the selection of constructions.
- Is the language easy to write? Is there “syntactic sugar”, which are expressions that are easier to read and write. It might be easier to be able to write: “All different: Paris, London, Tallinn, ...” to express that Paris, London, etc. must be different instances in the model that the sentence describes. Yet it introduces a new keyword “All different” which does not add any semantic expressivity, because the meaning

of the sentence could also be expressed as “Paris is not London. Paris is not Tallinn, ...”. Furthermore, it is debatable whether “All different: Paris, London, Tallinn, ...” is a construction often found in corpora of natural language.

- Is the language easy to understand? A CNL should look and feel like a natural language. The meanings of CNL statements should fit the meaning that readers would naturally assign to them. The selection of words and grammatical patterns might follow some guidelines such as found in the international Plain Language movement[3]. For example, certain complex constructions could be ruled out.
- Is the language predictable and unambiguous, that is, do the constructions have fixed and constant interpretations in an application domain?
- Is the language formally or informally defined? How accessible is the definition, meaning how much expert knowledge is required to understand the definition?
- How are semantic restrictions handled? Are there sortal restrictions with dependent types, database schemas of relational databases, or a logic in an ontology?
- Are statements translated into a logic? If so, how expressive is the logic? Descriptions logics provide subsets of first order logic, some of which are decidable [14]. In natural language semantics, there are claims that second order logics, which would support quantification over properties and relations, are required [15]. Are intensional expressions (temporal, belief, modal operators) translated into a modal logic? If statements are translated into a logic, what is the relationship – one to one (one form has one meaning and vice versa), one to many (ambiguity), many to one (simplification), or many to many. Where a range of interpretations are available, how is a selection made?
- Are discourses translated into a logic? If so, what sorts of discourse phenomena are addressed such as donkey anaphora, temporal anaphora, and reference to abstract objects (e.g. events, facts, or propositions)?
- What are the formal properties of the CNL in terms of expressivity, tractability, and decidability? What are the consequences, in practice and for the user, of the formal properties?
- How is the CNL evaluated? The CNL might be evaluated according the properties it has and how well it satisfies the requirements. For example, speed, coverage, accuracy of translation, user satisfaction, and so on may be properties used to evaluate the language.
- Is there a mapping to some graphical representation, e.g. conceptual graphs?
- Is the design of the language psycholinguistically motivated? For example, if the definition of the language allows two synonymous sentences, but one is shown by psycholinguistic research to take longer to read and to more often give rise to errors, should the language rule out the problematic sentence?
- Is there an explicit statement of the syntactic and semantic theory which underwrites the language?
- Is the CNL easily and systematically extensible (adding lexical, morphological, syntactic, and semantic elements or components)?
- What is the relationship between the parsing process and semantic representation? Is it a pipeline or parallel structure?

In the next subsection, we outline several of the linguistic properties of interest.

[3] http://www.plainlanguagenetwork.org/

2.3 Linguistic Properties: Lexicon, Morphology, Syntax, Semantics, and Pragmatics

In this section, we consider a range of linguistic properties that are relevant to the definition of a CNL. As there is great variety, we only consider general points. The choices here may depend, to a great extent, on answers to the design properties. The range of sources for "benchmark" linguistic properties varies greatly: one might take the advanced Oxford Learner's vocabulary and grammar as the basis; alternatively, the Oxford English Dictionary and a comprehensive grammar [16] might suit; or finally, there is extensive research literature such as that on diathesis [17] or pragmatics [18], among many others. Where and on what basis to make the "cut" in the language may follow from generic or design properties as well as linguistically informed choices.

- What corpus (if any) is one using to judge which linguistic forms to include in the language?
- What linguistic literature or theory (if any) is one using to justify the linguistic properties of the language?
- What classes of nouns, verbs, adjectives, adverbs, quantifiers, etc. are supported?
- Does the lexicon support polysemy or only monosemy? How is polysemy resolved relative to context?
- Is the language mono-lingual, or does it support multi-linguality?
- What morphological word formation rules are supported? Generally, we might expect rules for singular and plural nouns, agreement on verbs, tense and aspect. Some languages require gender and case agreement as well. Are there rules for nominalisation and compounding?
- Are interrogative and imperative forms supported? Are they generated from assertions or must they be explicitly written?
- Are idioms and metaphors allowed? For example, *Bill kicked the bucket*.
- Diathesis alternations (passive-actives, middles, ditransitives, causatives, inchoatives which signal the beginning of an action, and others). What inferential patterns are supported? For example, what are the implications between the following pairs of sentences?
  ia. John loaded the truck with hay.
  ib. John loaded hay on the truck.
  iia. Ann threw the ball to Beth.
  iib. Ann threw Beth the ball.
  iiia. John opened the door.
  iiib. The door opened.
- Where we have synonymous syntactic forms (outside the scope of diathesis), which should we adopt? How should relationships between them be defined?
  ia. Every employee that owns a car is rich.
  ib. If an employee owns a car then the employee is rich.
  iia. John's car is fast.
  iib. There is a car of John. The car is fast.
- Is there syntactic sugar, i.e. redundant expressions that make some expressions easier to state.

- Can there be discontinuous constituent structures, interruptions, or higher-level speech acts? For example, *Bill, so far as anyone knows, isn't in Africa any longer*.
- What sorts of query, relative clause, and sentence subordination markers are supported?
- What sorts of subordinate clauses are supported?
- Is the semantics compositional? If so, what notion of compositionality is at work?
- What logico-syntactic issues are addressed in the semantics such as opacity, quantifier scope, entailment patterns for different sorts of adjectives and adverbs, inference problems in modal logic, and the syntax-semantics interface, where syntactic structures place constraints on semantic interpretation.
- Is discourse supported? What aspects of anaphora are considered: times, locations, facts, propositions, and definite descriptions?
- Is there support for dialogical argumentation, where parties can make contradictory statements?
- Are the syntax and semantics of the language modular in the sense that there are components that can be added or removed to suit particular purposes while still preserving the functionalities of the remaining components? The modules might be (among others):
  - declarative (FOL) statements: *Some carpenter is an artist*.
  - modal operators such as alethic or deontic statements: *Bill may leave*; *Parties shall not breach terms of contract, otherwise penalties will be applicable*.
  - procedural statements: *She baked bread and then brought it to granny*.
  - propositional attitudes: *She believes that Bill is happy*.
  - mathematical proof statements: *Let's assume N is even. Then....*
  - rhetorical statements: *Rather than waste time on learning, Charles pursued a lucrative career.*
  - temporal, locative operators: *Bill left yesterday. He had been staying at Jill's house.*
- What sorts of presuppositions are supported?

2.4 Relationships and evaluation

To this point, we have outlined various properties that CNLs may have. In this section, we consider relationships among CNLs, including how they may be evaluated.

- What are the dependencies between the properties?
- What are the subsumption and similarity relations of existing CNLs?
- Is one CNL (or some of its subcomponents) interoperable with another CNL? For example, can the lexicon of a CNL be transformed automatically into a format so that the lexicon can be used by another CNL? Similarly, can the syntactic parse of one be input to the semantic interpreter of another? Is there a CNL "interchange" language, framework, or reference architecture?
- How modular is the CNL? In other words, is it possible to have a "plug and play" architecture for CNLs?

- Should a CNL be developed for only one target language or is it useful to develop a CNL simultaneously in multiple typologically distinct natural languages, so that incompatibilities between languages are discovered and remedied early. This is particularly crucial for translation.
- How can the relative performance of different CNLs be measured? What are the best practices guidelines? What questionaires and statistical analyses are used? Is performance measured against a "reference" CNL?
- Is there a common "pool" of use cases that we can use to evaluate a CNL?
- What is the measure of "naturalness" or habitabiliy of a lexicon or grammar? Is it linguistic judgement or, for example, statistical occurrence in a corpus of texts such as the British National Corpus or similar?

2.5 Application properties

Finally, we have several additional considerations that relate to applications of CNLs.

- Are there automatic consistency checks? Consistency is only applicable to CNLs which do semantic interpretation and inference. Is the input statement a contradiction of a statement already in the knowledge base? Is the input statement a contradiction of a statement which is implied by the knowledge base?
- Are there automatic redundancy checks? In other words, can we check whether the input statement is already in the knowledge base of statements? If the CNL includes a semantics, is the input statement implied by statements already in the knowledge base?
- Where the language is associated with an inference engine, are explanations of the inferences given in the CNL and are they context dependent?
- Is there guidance on style? Is the guidance separate from or built into the editing tools? That is, in the first approach, users are given instructions on the style such as "write short and grammatically simple sentences", "use nouns instead of pronouns", "use determiners", and "use active instead of passive". In the second approach, we have "predictive writing" where the editing tools suggest constructions in keeping with the style.
- What support tools are provided by the CNL? Among the tools, there can be a morphological analyser, parser (syntactic), lineariser (mapping from abstract representation to text), (predictive) editing tools which support the formulation of well-formed sentences and give feedback, editing loops, paraphraser, disambiguator/clarificatory dialogue, semantic interpretator, reasoner, and theorem prover.
- Is there a spoken language interface (for either input or output)?
- How is the language maintained and developed? Is the CNL proprietary or open?

3 A CNL Wiki

This paper is but a short overview of key considerations in the design and development of CNLs. For further collaborative discussions and future developments, see the wiki, where researchers discuss better principles, practices, and design patterns that can be of use to CNL developers/designers.[4].

[4] cnl.wikia.com

4 Contributor Information

- Krasimir Angelov
  Chalmers University of Technology and Göteborg University
  krasimir@chalmers.se
- Guntis Barzdins
  University of Latvia
  Guntis.Barzdins@mii.lu.lv
- Danica Damljanovic
  University of Sheffield
  d.damljanovic@dcs.shef.ac.uk
- Brian Davis
  Digital Enterprise Research Institute
  brian.davis@deri.org
- Norbert E. Fuchs
  University of Zurich
  fuchs@ifi.uzh.ch
- Stefan Hoefler
  University of Zurich
  hoefler@cl.uzh.ch
- Ken Jones
  kennethjone@gmail.com
- Kaarel Kaljurand
  University of Zurich
  kaljurand@gmail.com
- Tobias Kuhn
  University of Zurich
  kuhntobias@gmail.com
- Martin Luts
  ELIKO Competence Centre
  martin.luts@eesti.ee
- Jonathan Pool
  Utilika Foundation
  pool@utilika.org
- Mike Rosner
  University of Malta
  mike.rosner@um.edu.mt
- Rolf Schwitter
  Macquarie University
  Rolf.Schwitter@mq.edu.au
- John Sowa
  sowa@bestweb.net
- Adam Wyner
  University College London
  adam@wyner.info

References

1. Pool, J.: Can controlled languages scale to the web? In: Proceedings of the 5th International Workshop on Controlled Language Applications. (2006)
2. Mitamura, T., Nyberg, E.: Controlled english for knowledge-based mt: Experience with the kant system. In: Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine ranslation (TMI95), Belgium, Centre for Computational Linguistics, Katholieke Universiteit Leuven (1995) 158–172
3. Adriaens, G., Schreors, D.: From cogram to alcogram: Toward a controlled english grammar checker. In: Proceedings of the 14th Conference on Computational Linguistics, volume 2, Association for Computational Linguistics (1992) 595–601
4. Mueckstein, E.M.: Controlled natural language interfaces: the best of three worlds. In: CSC ï£¡85: Proceedings of the 1985 ACM thirteenth annual conference on Computer Science, Association for Computing Machinery (1985) 176–178
5. Bringsjord, S., Arkoudas, K., Clark, M., Shilliday, A., Taylor, J., Schimanski, B., Yang, Y.: Reporting on some logic-based machine reading research. In Etzioni, O., ed.: Machine Reading - Papers from the 2007 AAAI Spring Symposium. Number SS-07-06, Association for the Advancemend of Artificial Intelligence, AAAI Press (2007) 23–28
6. Kuhn, T.: Controlled English for Knowledge Representation. PhD thesis, Faculty of Economics, Business Administration and Information Technology of the University of Zurich (2010)
7. Leibniz, G.: Leibniz: Selections. Charles Scribner's Sons (1951)
8. O'Brien, S.: Controlling controlled english - an analysis of several controlled language rule sets. In: Controlled Translation - Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Application Workshop (EAMT-CLAW03), Dublin City University, Ireland (2003) 105–114
9. di Sciullo, A.M., Williams, E.: On the definition of Word. MIT Press (1987)
10. Schütze, C.: The Empirical Base of Linguistics. Chicago University Press (1996)
11. Watt, W.C.: Habitability. Amercian Documentation **19**(3) (1968) 338–351
12. Epstein, S.: Transportable natural language processing through simplicity - the pre system. ACM Transactions on Information Systems **3**(2) (1985) 107–120
13. Ogden, W., Bernick, P.: Using Natural Language Interfaces. In: Handbook of Human-Computer Interaction. Elsevier Science Publishers B.V. (1996)
14. Antoniou, G., Assmann, U., Baroglio, C., Decker, S., Henze, N., Patranjan, P.L., Tolksdorf, R., eds.: Reasoning Web, Third International Summer School 2007, Dresden, Germany, September 3-7, 2007, Tutorial Lectures. Volume 4636 of Lecture Notes in Computer Science. Springer (2007)
15. Barwise, J., Cooper, R.: Generalized quantifiers and natural language. Linguistics and Philosophy **4** (1981) 159–219
16. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.: A Comprehensive Grammar of the English Language. Pearson Longman (1985)
17. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press (1993)
18. Kadmon, N.: Formal Pragmatics: Semantics, Pragmatics, Presupposition, and Focus. Wiley-Blackwell (2001)