Interpreting Plurals in the Naproche CNL

Marcos Cramer and Bernhard Schröder

University of Bonn and University of Duisburg-Essen
cramer@math.uni-bonn.de, bernhard.schroeder@uni-due.de
http://www.naproche.net

Abstract. The Naproche CNL is a controlled natural language for mathematical texts. A recent addition to the Naproche CNL are plural statements. We discuss the collective-distributive ambiguity in the context of mathematical language, as well as pairwise interpretations of collective plurals. Additionally, we present a special scope ambiguity conjunctions give rise to. Finally, we describe an innovative plural interpretation algorithm implemented in Naproche for disambiguating plurals in DRT and giving them the interpretation that would normally be preferred in a mathematical context.

Key words: Naproche, CNL, plurals, DRT, distributive reading, collective reading

1 Introduction

The Naproche CNL [2] is a controlled natural language for mathematical texts, i.e. a controlled subset of the semi-formal language of mathematics (SFLM) as used in mathematical journals and textbooks. The Naproche system translates Naproche CNL texts first into Proof Representation Structures (PRSs, [2]), an adapted version of Discourse Representation Structures, which are further translated into lists of first-order formulae which are used for checking the logical correctness of a Naproche text using automated theorem provers.

The two main applications that we have in mind for Naproche are to make formal mathematics more readable to the average mathematician, and to use it as a tool that supports undergraduate students in writing formally correct proofs and thus get used to (a subset of) SFML.

A recent addition to the Naproche CNL are plural statements. By this we mean not only statements involving nouns in the plural (e.g. "numbers") and verbs conjugated in plural forms (e.g. "are"), but also conjunctive coordinations of noun phrases (e.g. "$x + y$ and $x \cdot y$ are even"). We discuss two kinds of ambiguities that originate from plural statements: the ambiguity between collective and distributive readings of plurals, and a special scope ambiguity conjunctions give rise to. Both ambiguities are resolved by an innovative *plural interpretation algorithm* that is geared towards the use of plurals in mathematical texts, and described in detail in this paper. Plural definite noun phrases (e.g. "the real numbers") are not yet implemented in Naproche and are left out of the discussion in this paper.

2 Proof Representation Structures

Proof Representation Structures (*PRSs*) are Discourse Representation Structures, which are enriched in such a way as to represent the distinguishing characteristics of the mathematical language. For the purpose of this paper, we present a simplified definition of PRSs:

A PRS is a pair consisting of a list of discourse referents and an ordered list of conditions,[1] usually depicted as a box, similarly to a DRS:

$d_1, \ldots, d_m$
c_1
$\vdots$
c_n

Just as in the case of DRSs, PRSs and PRS conditions are defined recursively: Let A, B be PRSs and $d, d_1, \ldots, d_n$ discourse referents. Then

- for any n-ary predicate p (e.g. expressed by adjectives and noun phrases in predicative use and verbs in SFLM), $p(d_1, \ldots, d_n)$ is a PRS condition.
- A mathematical formula is a PRS condition.
- $\neg A$ is a PRS condition, representing a negation.
- $B \Rightarrow A$ is a PRS condition, representing an assumption (B) and the set of claims made inside the scope of this assumption (A).
- *static*(A) is a PRS condition.

Accessibility in PRSs is defined analogously to accessibility in DRSs: Thus discourse referents introduced in conditions of the form $\neg A$ or $B \Rightarrow A$ are not accessible from outside these conditions. We have introduced an additional condition of the form *static*(A), which allows us to represent existential claims with a static rather than a dynamic existential quantification: Thus discourse referent introduced in a condition of the form *static*(A) are also not accessible from outside this condition.

3 Collective vs. distributive readings of plurals

The following sentence is ambiguous:[2]

(1) Three men lifted a piano.

It can mean either that three men lifted a piano together (in a single lifting act), or that there were three lifting acts, each of which involved a different man lifting a piano. The first is called the *collective* reading, the second the

[1] The use of ordered lists rather than sets in the definition of PRSs was motivated in [2]

[2] A comprehensive overview over plural readings is given by [6].

distributive reading.[3] The ambiguity arises because the agent of a lifting event can either be a collection of individuals or a single individual.

In SFLM, both the collective and the distributive reading exist:

(2) 12 and 25 are coprime.

(3) 2 and 3 are prime numbers.

Instead of (2), one could also say "12 is coprime to 25." So the adjective "coprime" can be used in two grammatically distinct ways, but in both cases refers to the same mathematical binary relation: either it is (predicatively or attributively) attached to a plural NP that gets a collective reading, or it has as a complement a prepositional phrase with "to". When used in the first way, we call "coprime" a *collective adjective*, when used in the second way, a *transitive adjective*. We say that the two logical arguments of "coprime" can be *grouped* into one collective linguistic argument, a plural NP with a collective reading. In general, mathematical adjectives expressing a symmetric binary relation have these two uses (cf. "parallel", "equivalent", "distinct", "disjoint"; in the case of "distinct" and "disjoint", the preposition used for the transitive case is "from" rather than "to"). Other cases of grouped arguments are "x and y commute" (cf. "x commutes with y") and "x connects y and z" (cf. "x connects y to z"). "x is between y and z" is an example of an expression with a grouped argument for which there is no corresponding expression without grouped arguments.

Since "prime number" expresses a unary relation, it is not possible to group two of its logical arguments into a single linguistic argument; this explains why (3) can't have a collective reading of the sort that (2) has. Which expressions can have grouped arguments is coded into the lexicon of the Naproche CNL.

An ambiguity like that of (1) can only arise when an expression (here the verb "to lift") has a linguistic argument that can be either a collectively interpreted plural NP or a singular NP (and can hence also be a distributively interpreted plural NP). Such expressions are extremely rare in SFLM. One example that we are aware of is the adjective "inconsistent":

(4) φ and ψ are inconsistent.

(4) can be mean either that the set of formulae $\{\varphi, \psi\}$ is an inconsistent set of formulae, or that φ is inconsistent and ψ is inconsistent. This ambiguity is avoided in Naproche by not marking "inconsistent" as an expression with grouped arguments in our lexicon, so that (4) only has the distributive reading; the collective reading can only be expressed with explicit set notation in Naproche.

4 Scope ambiguity

Another kind of ambiguity of special interest for our treatment of plurals and noun phrase conjunctions is a scope ambiguity that arises in certain sentences containing a noun phrase conjunction and a quantifier:

[3] We ignore cumulative readings here, because they play a negligible role in the mathematical contexts we have in mind.

(5) A and B contain some prime.

(5) can mean either that A contains a prime and B contains a (possibly different) prime, or that there is a prime that is contained in both A and B. In the first case we say that the scope of the noun phrase conjunction "A and B" contains the quantifier "some", whereas in the second case we say that the scope of "some" contains the noun phrase conjunction. We call the first reading the *wide-conjunction-scope* reading and the second the *narrow-conjunction-scope* reading.

Sometimes certain considerations of reference or variable range force one of the two readings, as in (6) and (7).

(6) x and y are integers such that some odd prime number divides $x + y$.

(7) x and y are prime numbers p such that some odd prime number q divides $p + 1$.[4]

(6) only has a narrow-conjunction-scope reading, because the existentially introduced entity is linked via a predicate ("divides") to a term ("$x + y$") that refers to the coordinated noun phrases individually. (7) on the other hand only has a wide-conjunction-scope reading, because the variable p must range over the values of both x and y, and q depends on p.

In general, there is, like in common language use, a strong tendency in SFLM texts to resolve scope ambiguities by giving wider scope to a quantifier that is introduced earlier in a sentence than to a quantifier introduced later in the sentence. This is a principle that we have already long ago adopted into Naproche in order to avoid scope ambiguities in the Naproche CNL. With the addition of coordinated NPs, we extended this principle to their scopes, with the exception of the cases like (6) where another reading is forced by certain syntactical considerations. Section 6 contains an account of how cases like (6) are identified.

5 Pairwise interpretations of collective plurals

In SFLM texts, one often sees sentences like (8) and (9), which are interpreted in a pairwise way as in (10) and (11):

(8) 7, 12 and 25 are coprime.

(9) All lines in A are parallel.

(10) $\mathit{coprime}(7, 12) \wedge \mathit{coprime}(12, 25) \wedge \mathit{coprime}(7, 25)$

(11) $\forall x, y \in A\ (x \neq y \rightarrow \mathit{parallel}(x, y))$[5]

[4] Given that this example is made up, one might ask whether it really occurs in SFLM texts that a plural noun followed by a variable is predicatively linked to a conjunction of terms as in this example. One real example that we found comes from page 4 of [1]: "Notice that 13, 37, 61, ..., are primes p such that $p^3 + 2$ and $p^3 + 1$ are squarefree."

[5] The distinctness condition here can be ignored in the case of reflexive relations like "parallel", but is certainly needed for non-reflexive relations like "coprime" or "disjoint".

Sometimes, especially in connection with the negative collective adjectives "distinct" and "disjoint", this interpretation is reinforced through the use of the word "pairwise", in order to ensure that one applies the predicate to all pairs of objects collectively referred to by the plural NP. But given that this pairwise interpretation is at any rate the standard interpretation of such sentences even in the absence of the adverb "pairwise", we decided not to require the use of the word "pairwise" in the Naproche CNL.

The Naproche CNL allows only this pairwise interpretation for a plural NP that is used as a grouped argument of such a collective adjective. (12) is a sentence where another reading (14) might naturally be preferred to the pairwise interpretation (13) that Naproche assigns to it:

(12) Some numbers in A and B are coprime.

(13) $\exists n, m\ (\mathit{number}(n) \wedge n \in A \wedge n \in B \wedge \mathit{number}(m) \wedge m \in A \wedge m \in B \wedge \mathit{coprime}(n, m))$

(14) $\exists n, m\ (\mathit{number}(n) \wedge n \in A \wedge \mathit{number}(m) \wedge m \in B \wedge \mathit{coprime}(n, m))$

However, it seems to us that such sentences hardly appear in real mathematical texts.

6 The plural interpretation algorithm

In the Naproche system, the PRS construction algorithm for the representation of single sentences has been added to the standard threading algorithm for DRS construction (see [4]), and is implemented in Prolog. The algorithm can cope with plurals, plural ambiguity resolution and pairwise interpretations as explained in the previous sections. We illustrate how the algorithm treats plurals by considering the following example sentence:

(15) x and y are distinct primes p such that $2p+1$ is a square number and some odd prime divides $x + y$.

This example has only one natural reading, and illustrates all the natural disambiguation methods mentioned in the previous sections: The plural construction "x and y" is modified by one predicate ("distinct") that needs to be interpreted collectively and by one predicate ("prime") that needs to be interpreted distributively. One of the existential NPs in the such-that clause ("a square number") has to be given a narrow scope, while the other ("some odd prime") has to be given a wide scope. The algorithm specifies a formal procedure to attain this natural reading.

The algorithm works by first producing a preliminary representation (Fig.1):

Here the NP conjunction gets a plural discourse referent (p in Fig. 1), which is linked to the discourse referents of the conjuncts by a *plural_dref condition*. We give the NP conjunction wide scope over all quantifiers introduced later, and

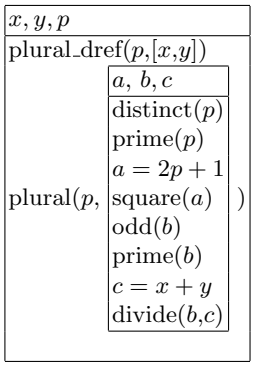


Fig. 1. Preliminary PRS

all assertions made in the scope of the plural NP are inserted in a special *plural sub-PRS*. The *plural-dref* and *plural* conditions used in such preliminary PRSs are book-keeping devices and not part of the official PRS language.[6]

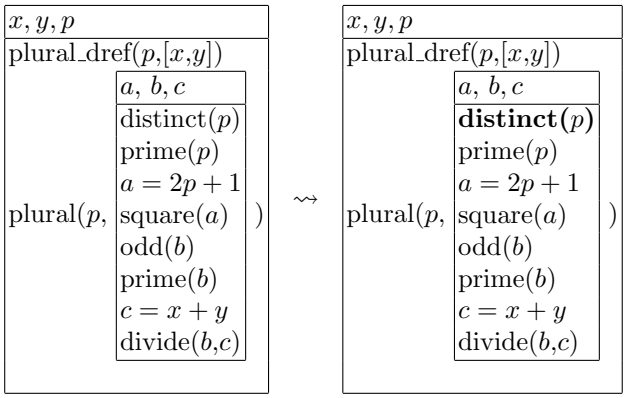
The goal of the algorithm is to eliminate the plural discourse referents in favour of the singular discourse referents they subordinate. This has to be done separately for the distributively and collectively interpreted parts. The distributive interpretations opens a scopus, in which there may occur dependent variables. The algorithm consists of five steps, which can be summarized as follows: For each plural referent:

1. Mark the collective uses of the plural referent.
2. Mark the distributive uses of the plural referent and dependent variables.
3. Separate the scopus of distributive uses of the plural referent from the rest.
4. Replace collective variable occurrences.
5. Replace distributive variable occurrences.

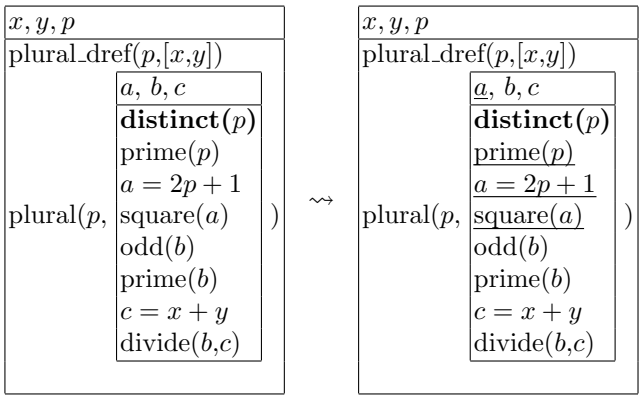
Now we describe each of the steps more formally:

1. Marking the collective uses of the plural referent: In the plural sub-PRS, we mark every PRS condition which consists of a predicate that has the plural discourse referent as grouped argument ("distinct(p)" in the example PRS, marked by boldface). That the plural discourse referent is a grouped argument is derived from the fact that the number of arguments, with which the predicate appears in the plural sub-PRS, is one less than its logical number of arguments fixed in the lexicon, and from the fact that the lexicon specifies the possibility of grouping two of its arguments into one.

[6] Alternatively, one may consider these conditions as extensions of the PRS language, in which case the semantics of the *plural* condition has to be an underspecified semantics in the sense of [3], which represents the different scopal interpretations of the plural NP.

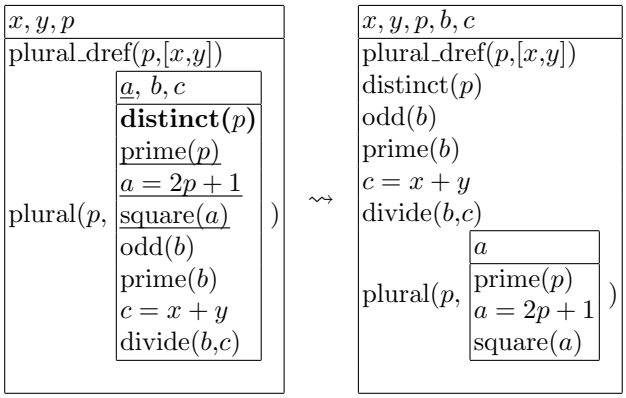


2. Marking the distributive uses of the plural referent and dependent variables: In the plural sub-PRS, we recursively mark (in the figure by underlining) all PRS conditions that were not marked in step 1 and contain the plural discourse referent or a marked discourse referent, and all discourse referents contained in a PRS condition marked in this way, until no more conditions and discourse referents can be marked by this process:

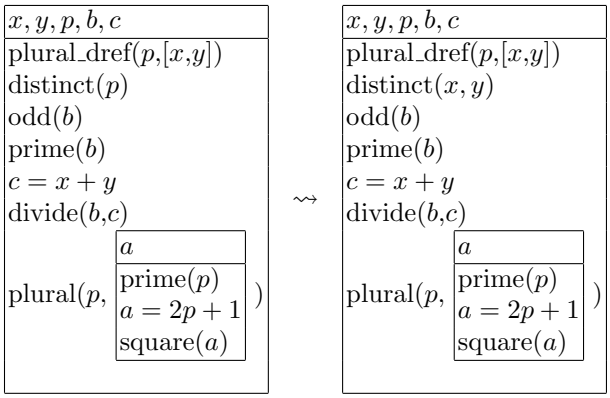


3. Separating the scopus of distributive uses of the plural referent from the rest: All discourse referents and PRS conditions in the plural sub-PRS not marked in step 2 get pulled out of the plural sub-PRS and inserted into its super-PRS:[7]

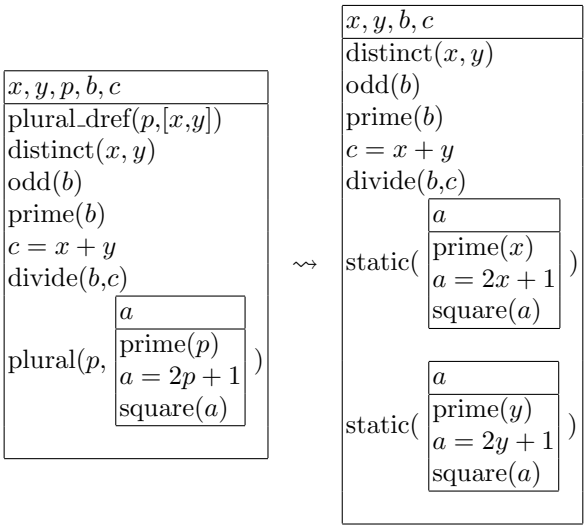
[7] Since this step moves discourse referents and conditions around, one might wonder whether it can cause formally bound variables to become free. This, however, is impeded by the recursive procedure in step 2: If a certain discourse referent stays in the plural sub-PRS, no condition containing this discourse referent can be pulled out of the plural sub-PRS.



4: Replacing collective variable occurrences: For every PRS condition $p(d)$ with grouped argument d, and every pair d_1, d_2 of distinct discourse referents linked to d via a plural_dref condition, we create a PRS condition of the form $p(d_1, d_2)$ and remove the original PRS condition $p(d)$ (in our example this amounts to replacing "distinct(p)" by "distinct(x, y)"):



5. Replacing distributive variable occurrences: For every discourse referent d linked to the plural discourse referent p, we make a static copy of the plural sub-PRS in which every instance of p is replaced by d, removing the original plural sub-PRS:



The final PRS corresponds to the natural reading of sentence (15) that we described at the beginning of this section.

7 Related and Future Work

The syntax of Attempto Controlled English (ACE) allows plurals, which are interpreted in ACE in an unambiguous way [8]. The disambiguation used by ACE is very distinct from Naproche's: while Naproche gives preference to distributive and wide-conjunction-scope readings, ACE allows only collective and narrow-conjunction-scope readings, unless the word "each" is used. This difference is due to the fact that for Naproche we focused on the interpretations common in SFLM, whereas ACE took the English language as a whole into account. Our focus on mathematical language also made it important for us to treat "x and y are coprime" and "x is coprime to y" as logically equivalent, which ACE does not do.

ForTheL, the controlled natural language of the System for Automated Deduction (SAD), a project with similar goals to Naproche, already included the two uses of words like "parallel" and "to commute" and produced the same representation no matter in which way they were used [7].

At the moment, Naproche does not yet allow anaphoric pronouns like "it" and "they". When Naproche is extended to allow them, some rules specifying how to control the many ways in which an anaphoric antecedent for "they" can be chosen (see [5]) will have to be specified and implemented, again with special attention to existing usage in SFLM.

8 Conclusion

We have implemented a plural intepretation algorithm that can handle a number of constructs related to plurals in a way that seems desirable for a mathematical

CNL: While a distributive reading of plurals is preferred, a collective reading is chosen for predicates with grouped arguments and the pairwise interpretation of predicates with grouped arguments is chosen when feasible. Additionally, the scope ambiguity that noun phrase conjunctions give rise to is disambiguated with respect to the syntactic-semantic context.

References

1. Cohen, G. L.: Derived Sequences. Journal of integer sequences, Vol. 6 (2003)
2. Cramer, M., Fisseni, B., Koepke, P., Kühlwein, D., Schröder, B., Veldman, J.: The Naproche Project – Controlled Natural Language Proof Checking of Mathematical Texts. CNL 2009 Workshop, LNAI 5972 proceedings (2010)
3. Egg, M.: Semantic underspecification. In: Semantics (ed. by C. Maienborn, K. von Heusinger, P. Portner) (HSK vol. 33). De Gruyter Mouton (2011)
4. Johnson, M., Klein, E.: Discourse, anaphora and parsing, Proceedings of the 11th coference on Computational linguistics (1986)
5. Kamp, H., Reyle, U.: From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Kluwer Academic Publisher (1993)
6. Link, G.: Plural. In: Semantics (ed. by A. von Stechow and D. Wunderlich) (HSK vol. 6). de Gruyter (1991).
7. Paskevich, A.: The syntax and semantics of the ForTheL language (2007)
8. Schwertel, U.: Plural Semantics for Natural Language Understanding – A Computational Proof-Theoretic Approach. PhD thesis. University of Zurich (2005)