MathLang
- Gradual computerisation of mathematical texts -

ULTRA Group
Heriot-Watt University Edinburgh
Prof. Fairouz Kamareddine
Dr. Joe Wells
MathLang Members:
M. Maarek (10/2002 – 6/2007)
K. Retel (11/2004 –)
R. Lamar (10/2006 –)
C. Zengler (1/2008 –)
Numerous undergraduate and MSc students since 2000

Mathematicians speak their own language

CML - Common Mathematical Language

+ CML is expressive: it has linguistic categories like proofs and theorems.
+ CML has been refined by intensive use and is rooted in long traditions.
+ CML is approved by most mathematicians as a communications medium

Mathematicians speak their own language

CML - Common Mathematical Language

+ CML is expressive: it has linguistic categories like proofs and theorems.
+ CML has been refined by intensive use and is rooted in long traditions.
+ CML is approved by most mathematicians as a communications medium
- Since CML is based on natural language, it is informal and ambiguous.
- CML is incomplete: Much is left implicit, appealing to the reader's intuition.
- CML is automation-unfriendly: A CML text is a plain text and cannot be easily automated

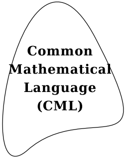
What are the options for computerisation?

Computers can handle mathematical text at various levels:

- Images of pages may be stored. While useful, this is not a good representation of language or knowledge.
- Typesetting systems like LaTeX, TeXmacs, can be used.
- Document representations like OpenMath, OMDoc, MathML, can be used.
- Formal logics used by theorem provers (Coq, Isabelle, Mizar, Isar, etc.) can be used.

We are gradually developing a system named MathLang which we hope will eventually allow building a bridge between the latter 3 levels.

The starting point - the orginal CML text



Lemma 1. For $m, n \in \mathbb{N}$ one has: $m^2 = 2n^2 \implies m = n = 0$.

Proof Define on $\mathbb{N}$ the predicate:

$$P(m) \iff \exists n. m^2 = 2n^2 \,\&\, m > 0.$$

Claim. $P(m) \implies \exists m' < m.P(m')$. Indeed suppose $m^2 = 2n^2$ and $m > 0$. It follows that m^2 is even, but then m must be even, as odds square to odds. So $m = 2k$ and we have

$$2n^2 = m^2 = 4k^2 \implies n^2 = 2k^2$$

Since $m > 0$, if follows that $m^2 > 0$, $n^2 > 0$ and $n > 0$. Therefore $P(n)$. Moreover, $m^2 = n^2 + n^2 > n^2$, so $m^2 > n^2$ and hence $m > n$. So we can take $m' = n$.

By the claim $\forall m \in \mathbb{N}.\neg P(m)$, since there are no infinite descending sequences of natural numbers.

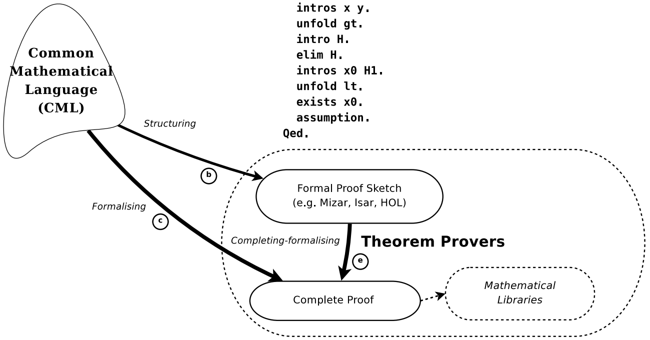
Now suppose $m^2 = 2n^2$ with $m \neq 0$. Then $m > 0$ and hence $P(m)$. Contradiction. Therefore $m = 0$. But then also $n = 0$.

Corollary 1. $\sqrt{2} \notin \mathbb{Q}$.

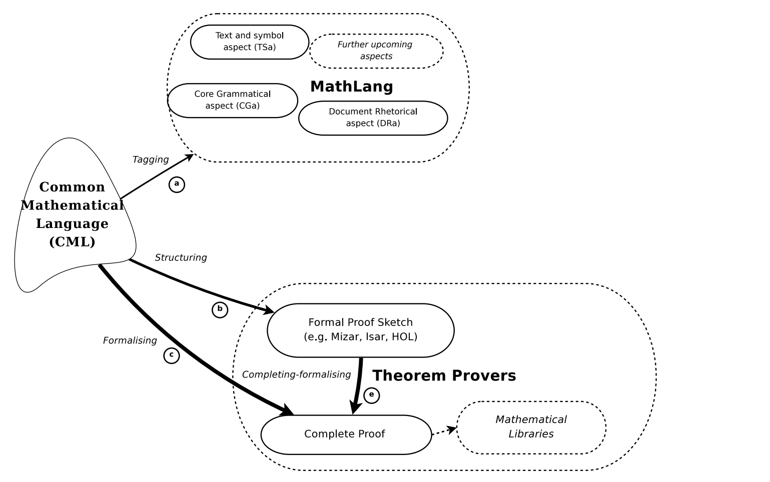
Proof Suppose $\sqrt{2} \in \mathbb{Q}$, i.e. $\sqrt{2} = p/q$ with $p \in \mathbb{Z}, q \in \mathbb{Z} - \{0\}$. Then $\sqrt{2} = m/n$ with $m = |p|, n = |q| \neq 0$. It follows that $m^2 = 2n^2$. But then $n = 0$ by the lemma. Contradiction shows that $\sqrt{2} \notin \mathbb{Q}$.

And what's the aim?

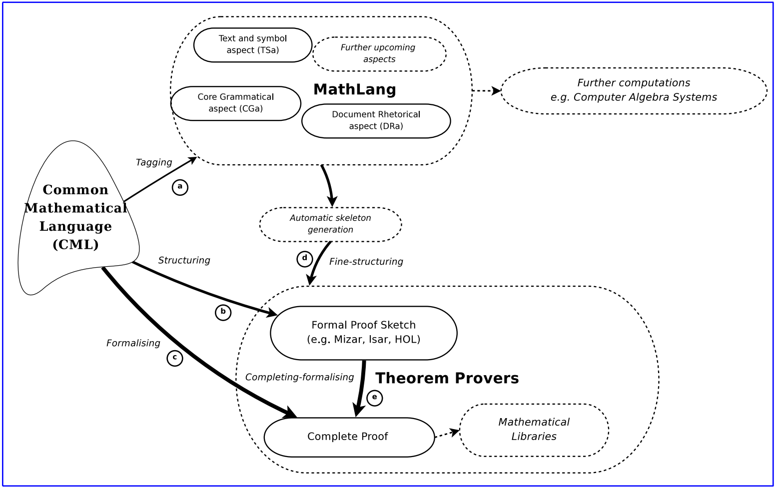
```
Theorem th111 (x y : nats) : gt x y -> lt y x.
Proof.
  intros x y.
  unfold gt.
  intro H.
  elim H.
  intros x0 H1.
  unfold lt.
  exists x0.
  assumption.
Qed.
```



MathLang divides and simplifies the path



MathLang divides and simplifies the path



The Starting Point: CGa
Core Grammatical Aspect

- CGa is a formal language derived from the Mathematical Vernacular MV (de Bruijn, 1987) and Weak Type Theory WTT (Kamareddine and Nederpelt, 2004)
- aims at expliciting the grammatical role played by the elements of a CML text
- We have 9 different categories to represent grammatical/linguistic/syntactic categories of the original CML
- Each of these categories has a corresponding weak type.
- This weak type system allows checking of the document from a syntactical point of view.

The categories of CGa

- Term: *the triangle ABC, the center of ABC,* $\sqrt{2}$,...
- Set: $\mathbb{N}$, *the natural numbers,...*
- Noun: *number, circle, a group,...*
- Adjective: *equilateral, prime, Abelian,...*
- Statement: *a* $=$ *b, P lies between Q and R, p is prime,...*
- Declaration: *Let A be a set, let p be prime,...*
- Definition: *p is prime whenever... , A triangle is equilateral when..., ...*
- Context: *Assume a is even, Let a=b,...*
- Step: *a is odd, hence a* $\neq$ *0*

But no one wants to rewrite everything: TSa Text & Symbol Aspect

- term *the triangle ABC, the center of ABC,* $\sqrt{2}$*,...*
- set $\mathbb{N}$*, the natural numbers,...*
- noun *number, circle, a group,...*
- adjective *equilateral, prime, Abelian,...*
- statement *a* $=$ *b, P lies between Q and R, p is prime,...*
- declaration *Let A be a set, let p be prime,...*
- definition *p is prime whenever... , A triangle is equilateral when..., ...*
- context *Assume a is even, Let a=b,...*
- step *a is odd, hence a* $\neq$ *0*

What does TSa look like?

The original Text

$$A \subset B \iff \forall x (\, x \in A \implies x \in B \,)$$

What does TSa look like?

The original Text

$A \subset B \Longleftrightarrow \forall x (\, x \in A \Longrightarrow x \in B \,)$

A and B are sets

$\boxed{\mathrm{A}} \subset \boxed{\mathrm{B}} \Longleftrightarrow \forall x (\, x \in \boxed{\mathrm{A}} \Longrightarrow x \in \boxed{\mathrm{B}})$

What does TSa look like?

The original Text

$A \subset B \iff \forall x(\, x \in A \implies x \in B\,)$

A and B are sets

$\boxed{A} \subset \boxed{B} \iff \forall x(\, x \in \boxed{A} \implies x \in \boxed{B})$

x is locally declared as a term

$\boxed{A} \subset \boxed{B} \iff \forall \boxed{\boxed{x}}(\, \boxed{x} \in \boxed{A} \implies \boxed{x} \in \boxed{B})$

What does TSa look like?

The original Text

$$A \subset B \iff \forall x(\, x \in A \implies x \in B\,)$$

A and B are sets

$$\boxed{\mathrm{A}} \subset \boxed{\mathrm{B}} \iff \forall x(\, x \in \boxed{\mathrm{A}} \implies x \in \boxed{\mathrm{B}})$$

x is locally declared as a term

$$\boxed{\mathrm{A}} \subset \boxed{\mathrm{B}} \iff \forall \boxed{\mathrm{x}}(\, \boxed{\mathrm{x}} \in \boxed{\mathrm{A}} \implies \boxed{\mathrm{x}} \in \boxed{\mathrm{B}})$$

$\subset$, $\in$, $\forall$, $\implies$ are statements

$$\boxed{\boxed{A} \subset \boxed{B}} \iff \boxed{\forall \boxed{x}(\boxed{\boxed{\boxed{x} \in \boxed{A}} \implies \boxed{\boxed{x} \in \boxed{B}}})}$$

And finally...

The whole construct is a definition

$$A \subset B \Longleftrightarrow \forall x (x \in A \longrightarrow x \in B)$$

And finally...

The whole construct is a definition

$A \subset B \iff \forall x (x \in A \Longrightarrow x \in B)$

The same with interpretations:

<subset><A>A $\subset$ <B>B $\iff$

<forall>$\forall$ <x>x (<impl><in><x>x $\in$ <A>A $\Longrightarrow$ <in><x>x $\in$ <B>B)

And finally...

The whole construct is a definition

$A \subset B \iff \forall x (x \in A \implies x \in B)$

The same with interpretations:

<subset> <A> A $\subset$ <B> B $\iff$

<forall> $\forall$ <x> x (<impl> <in> <x> x $\in$ <A> A $\implies$ <in> <x> x $\in$ <B> B)

or it could be:

<subset> <A> A is a subset of <B> B iff

<forall> for all <x> x it holds <impl> <in> <x> x $\in$ <A> A $\implies$ <in> <x> x $\in$ <B> B

The larger Context: DRa
Document Rhetorical Aspect

- **Structural components of a document** like *book, chapter, section, subsection,...*
- **Mathematical components of a document** like *theorem, corollary, definition, proof,...*
- **Relations between these components** like *A justifies B, B uses C,...*
- Enhance readability and ease navigation of documents

What does it look like? - The original text

Lemma 1. For $m, n \in \mathbb{N}$ one has: $m^2 = 2n^2 \implies m = n = 0$.

Proof Define on $\mathbb{N}$ the predicate:

$$P(m) \iff \exists n. m^2 = 2n^2 \;\&\; m > 0.$$

Claim. $P(m) \implies \exists m' < m. P(m')$. Indeed suppose $m^2 = 2n^2$ and $m > 0$. It follows that m^2 is even, but then m must be even, as odds square to odds. So $m = 2k$ and we have

$$2n^2 = m^2 = 4k^2 \implies n^2 = 2k^2$$

Since $m > 0$, if follows that $m^2 > 0, n^2 > 0$ and $n > 0$. Therefore $P(n)$. Moreover, $m^2 = n^2 + n^2 > n^2$, so $m^2 > n^2$ and hence $m > n$. So we can take $m' = n$.

By the claim $\forall m \in \mathbb{N}. \neg P(m)$, since there are no infinite descending sequences of natural numbers.

Now suppose $m^2 = 2n^2$ with $m \neq 0$. Then $m > 0$ and hence $P(m)$. Contradiction. Therefore $m = 0$. But then also $n = 0$.

Corollary 1. $\sqrt{2} \notin \mathbb{Q}$.

Proof Suppose $\sqrt{2} \in \mathbb{Q}$, i.e. $\sqrt{2} = p/q$ with $p \in \mathbb{Z}, q \in \mathbb{Z} - \{0\}$. Then $\sqrt{2} = m/n$ with $m = |p|, n = |q| \neq 0$. It follows that $m^2 = 2n^2$. But then $n = 0$ by the lemma. Contradiction shows that $\sqrt{2} \notin \mathbb{Q}$.

What does it look like? - A first level annotation

Lemma 1. For $m, n \in \mathbb{N}$ one has: $m^2 = 2n^2 \implies m = n = 0$. `Lemma`

Proof Define on $\mathbb{N}$ the predicate: `Proof`

$$P(m) \iff \exists n. m^2 = 2n^2 \;\&\; m > 0.$$

Claim. $P(m) \implies \exists m' < m. P(m')$. Indeed suppose $m^2 = 2n^2$ and $m > 0$. It follows that m^2 is even, but then m must be even, as odds square to odds. So $m = 2k$ and we have

$$2n^2 = m^2 = 4k^2 \implies n^2 = 2k^2$$

Since $m > 0$, if follows that $m^2 > 0, n^2 > 0$ and $n > 0$. Therefore $P(n)$. Moreover, $m^2 = n^2 + n^2 > n^2$, so $m^2 > n^2$ and hence $m > n$. So we can take $m' = n$.

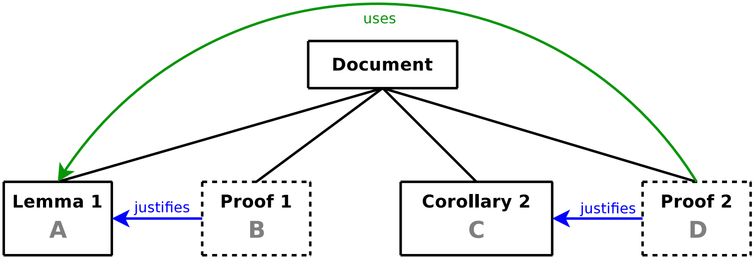
By the claim $\forall m \in \mathbb{N}. \neg P(m)$, since there are no infinite descending sequences of natural numbers.

Now suppose $m^2 = 2n^2$ with $m \neq 0$. Then $m > 0$ and hence $P(m)$. Contradiction. Therefore $m = 0$. But then also $n = 0$.

Corollary 1. $\sqrt{2} \notin \mathbb{Q}$. `Corollary`

Proof Suppose $\sqrt{2} \in \mathbb{Q}$, i.e. $\sqrt{2} = p/q$ with $p \in \mathbb{Z}, q \in \mathbb{Z} - \{0\}$. Then $\sqrt{2} = m/n$ with $m = |p|, n = |q| \neq 0$. It follows that $m^2 = 2n^2$. But then $n = 0$ by the lemma. Contradiction shows that $\sqrt{2} \notin \mathbb{Q}$. `Proof`

What does it look like? - The corresponding DRa tree



What does it look like? - Going one level deeper

Lemma 1. For $m, n \in \mathbb{N}$ one has: $m^2 = 2n^2 \implies m = n = 0$. Lemma

Proof Define on $\mathbb{N}$ the predicate: Definition Proof

$$P(m) \iff \exists n. m^2 = 2n^2 \;\&\; m > 0.$$

Claim Proof

Claim. $P(m) \implies \exists m' < m. P(m')$. Indeed suppose $m^2 = 2n^2$ and $m > 0$. It follows that m^2 is even, but then m must be even, as odds square to odds. So $m = 2k$ and we have

$$2n^2 = m^2 = 4k^2 \implies n^2 = 2k^2$$

Since $m > 0$, if follows that $m^2 > 0$, $n^2 > 0$ and $n > 0$. Therefore $P(n)$. Moreover, $m^2 = n^2 + n^2 > n^2$, so $m^2 > n^2$ and hence $m > n$. So we can take $m' = n$.

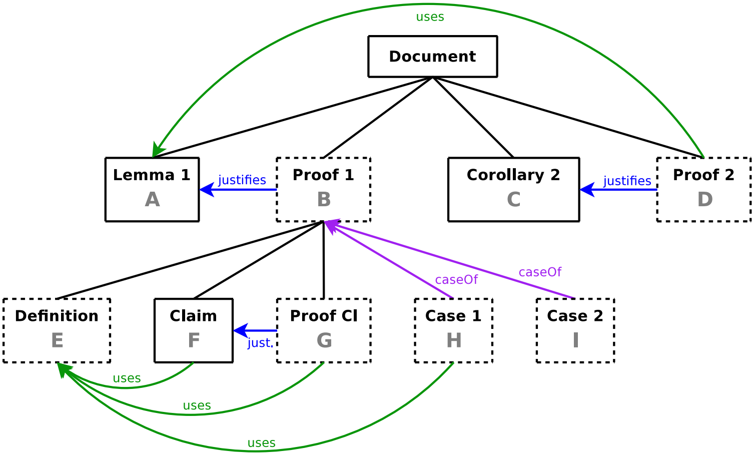
By the claim $\forall m \in \mathbb{N}. \neg P(m)$, since there are no infinite descending sequences of natural numbers. Case 1

Now suppose $m^2 = 2n^2$ with $m \neq 0$. Then $m > 0$ and hence $P(m)$. Contradiction. Therefore $m = 0$. But then also $n = 0$. Case 2

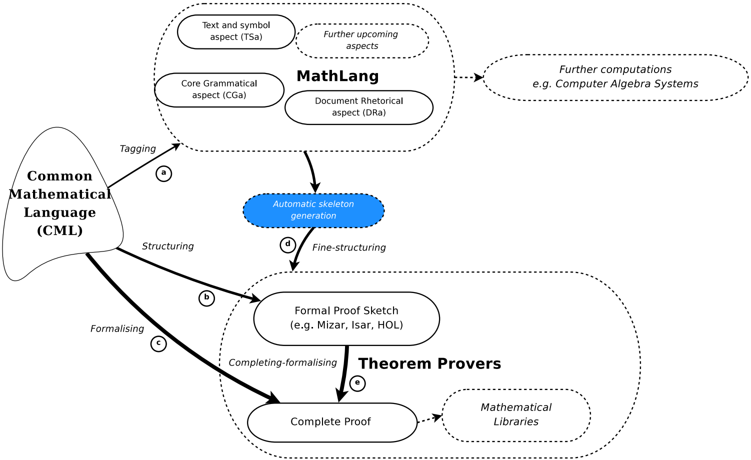
Corollary 1. $\sqrt{2} \notin \mathbb{Q}$. Corollary

Proof Suppose $\sqrt{2} \in \mathbb{Q}$, i.e. $\sqrt{2} = p/q$ with $p \in \mathbb{Z}, q \in \mathbb{Z} - \{0\}$. Then $\sqrt{2} = m/n$ with $m = |p|, n = |q| \neq 0$. It follows that $m^2 = 2n^2$. But then $n = 0$ by the lemma. Contradiction shows that $\sqrt{2} \notin \mathbb{Q}$. Proof

What does it look like? - The final DRa tree



Where are we?


Text and symbol aspect (TSa)
Further upcoming aspects
MathLang
Further computations e.g. Computer Algebra Systems
Core Grammatical aspect (CGa)
Document Rhetorical aspect (DRa)
Tagging
a
Common Mathematical Language (CML)
Automatic skeleton generation
Structuring
d
Fine-structuring
b
Formal Proof Sketch (e.g. Mizar, Isar, HOL)
Formalising
c
Completing-formalising
Theorem Provers
e
Complete Proof
Mathematical Libraries

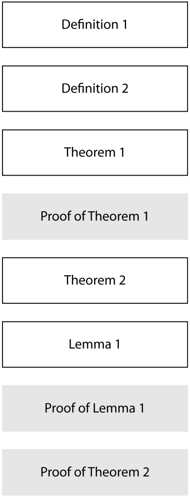
What is Skeleton Generation?

- Check the DRa annotations
  - Are there loops in the graph? (*A justifies B, B justifies A*)
  - Are there unproofed nodes which should be proofed, proofed nodes, which should not be proofed?
- Generate a proof skeleton of the annotated text for a certain theorem proofer
- highly configurable generic algorithm which takes the properties of different theorem proofers in account

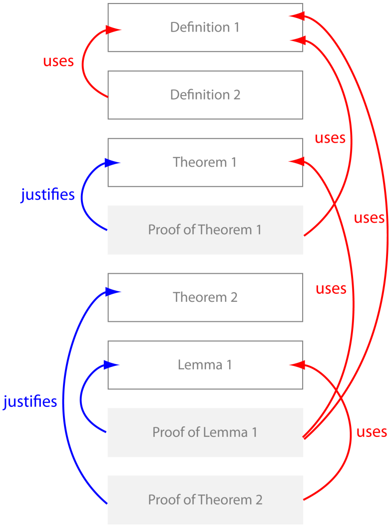
How does the skeleton generation look like?

Different provers have

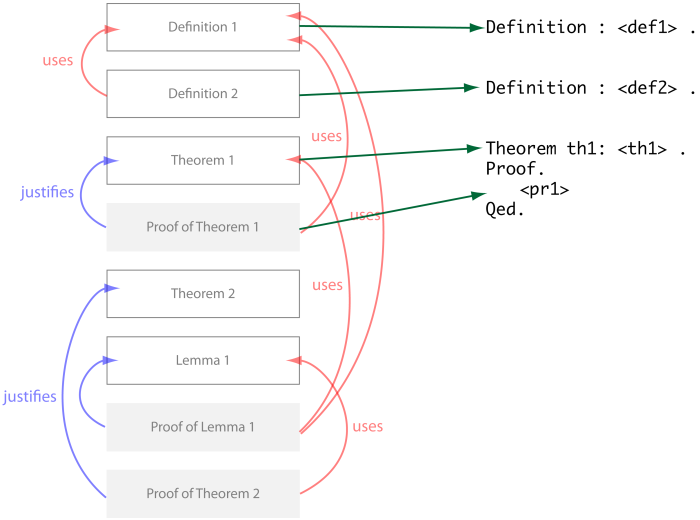
- different syntax
- different requirements to the structure of the text e.g.
  - no nested theorems/lemmas
  - only backward references
  - ...
- Aim: Skeleton should be as close as possible to the mathematician's text but with re-arrangements when necessary



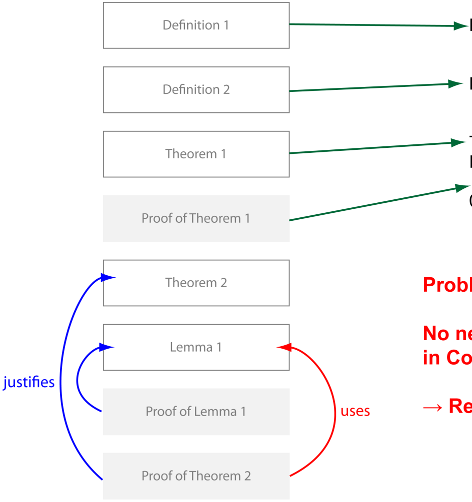
The relations of the example



Straight-forward translation of the first part



Problem: nested theorems



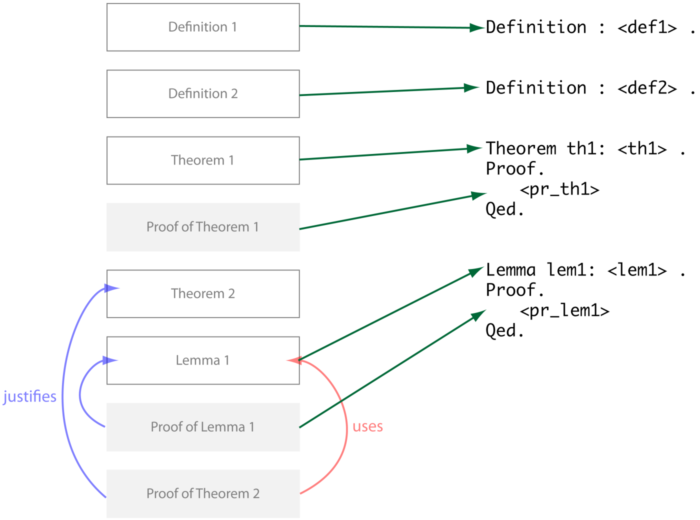

```
Definition : <def1> .
Definition : <def2> .
Theorem th1: <th1> .
Proof.
    <pr1>
Qed.
```

Problem

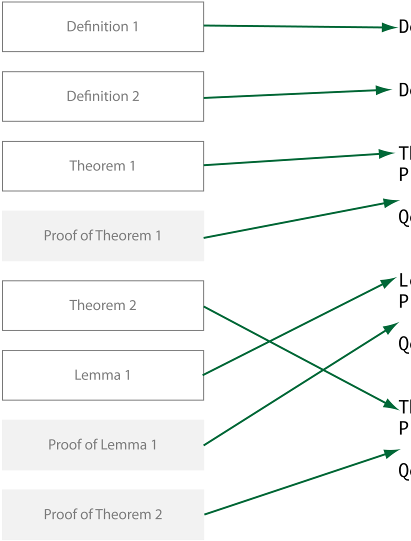
No nested theorems/lemmas in Coq e.g.

→ Re-Ordering !

Solution: Re-ordering

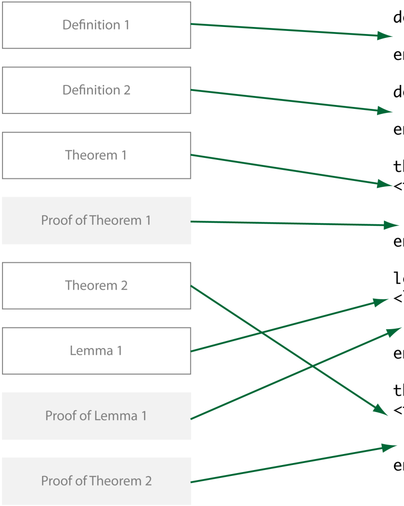


Finishing the skeleton (for Coq)




```
Definition : <def1> .

Definition : <def2> .

Theorem th1: <th1> .
Proof.
    <pr_th1>
Qed.

Lemma lem1: <lem1> .
Proof.
    <pr_lem1>
Qed.

Theorem th2: <th2> .
Proof.
    <pr_th2>
Qed.
```

Skeleton for Mizar




```
definition def1:
    <def1>
end;

definition def2:
    <def2>
end;

theorem th1:
<th1>
    proof
    <pr_th1>
end;

lem1:
<lem1>
    proof
    <pr_lem1>
end;

theorem th2:
<th2>
    proof
    <pr_th2>
end;
```

Chapter 1 of Landau's Grundlagen der Analysis

Facts

- 18 pages of plaintext CML
- 5 axioms
- 6 definitions
- 36 Theorems with proofs
- over 200 relations between chunks of text

Timeframe

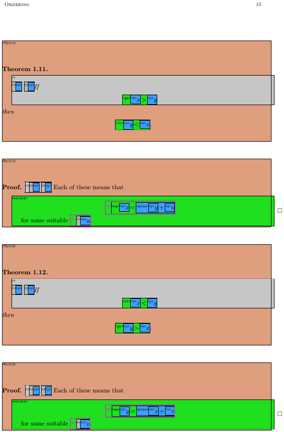
- CGa + TSa annotation: ca. 20h
- DRa annotation: ca 1h
- Skeleton generation: automatically ($<$ 1 sec)
- Completing the proofs in Coq: 5h

The CGa annotation (2 pages)

MathLang

ULTRA Group

- Introduction
- Overview over the system
- The aspects of MathLang
  - CGa
  - TSa
  - DRa
  - Skeleton Generation
- **A complete example**
- Outlook



1.3 ORDERING 15

Theorem 1.11. *If*

$$x > y$$

then

$$y < x.$$

Proof. Each of these means that

$$x = y + u$$

for some suitable u. □

Theorem 1.12. *If*

$$x < y$$

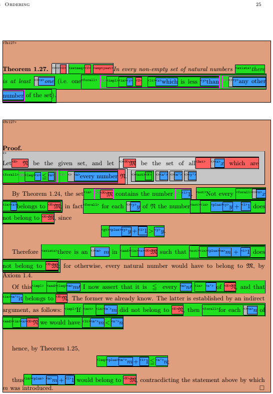
then

$$y > x.$$

Proof. Each of these means that

$$y = x + v$$

for some suitable v. □



1.3 ORDERING 25

Theorem 1.27. *In every non-empty set of natural numbers there is at least one* (i.e. one which is less than any other number of the set).

Proof. Let $\mathfrak{M}$ be the given set, and let $\mathfrak{N}$ be the set of all x which are $\leq$ every number of $\mathfrak{M}$.

By Theorem 1.24, the set $\mathfrak{N}$ contains the number 1. Not every x belongs to $\mathfrak{N}$; in fact for each y of $\mathfrak{M}$ the number $y + 1$ does not belong to $\mathfrak{N}$, since

$$y + 1 > y.$$

Therefore there is an m in $\mathfrak{N}$ such that $m + 1$ does not belong to $\mathfrak{N}$; for otherwise, every natural number would have to belong to $\mathfrak{N}$, by Axiom 1.4.

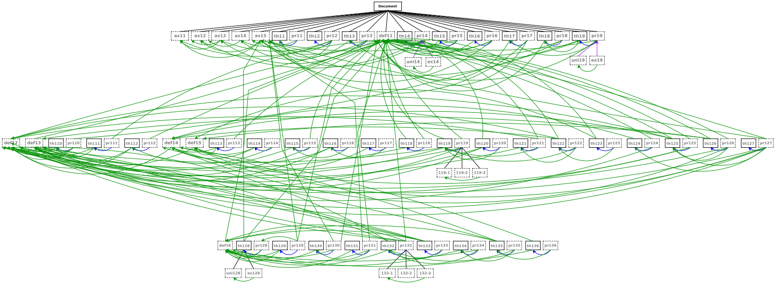
Of this m I now assert that it is $\leq$ every n of $\mathfrak{M}$ and that it belongs to $\mathfrak{M}$. The former we already know. The latter is established by an indirect argument, as follows: If m did not belong to $\mathfrak{M}$, then for each n of $\mathfrak{M}$ we would have $m < n$,

hence, by Theorem 1.25,

$$m + 1 \leq n;$$

thus $m + 1$ would belong to $\mathfrak{N}$, contradicting the statement above by which m was introduced. □

The graph of the DRa annotation



The proof skeleton (Coq)


```
Theorem th13: <th13> .
Proof.
         <pr13>
Qed.

Definition : <def11> .

Theorem th14: <th14> .
Proof.
         <pr14uniqueness>
         <pr14existence>
Qed.
```

The proof skeleton (Mizar)


```
theorem th13:
<th13>
          proof
          <pr13>
end;

definition def11:
          <def11>
end;

theorem th14:
<th14>
          proof

          uniqueness:
          <pr14uniqueness>

          existence:
          <pr14existence>
end;
```

The final proof (excerpt)


```
Theorem th11 (x y:nats) : neq x y -> neq (succ x) (succ y).
Proof.
intro x; intro y.
apply contraposition1.
intro.
rewrite <- triv1.
rewrite <- triv1 in H.
apply ax14.
assumption.
Qed.

Theorem th12 (x:nats) : neq (succ x) x.
Proof.
intro x.
elim x.
apply ax13.
intros n H.
apply th11.
assumption.
Qed.
```

What will happen in the future?

- Complete reimplementation of the existing system (ongoing)
- Extended annotation functions for easier TSa annotations (largest time effort in the translation process

$\Longrightarrow$ Get a community of MathLang users

What will happen in the future?

- Complete reimplementation of the existing system (ongoing)
- Extended annotation functions for easier TSa annotations (largest time effort in the translation process

$\Longrightarrow$ Get a community of MathLang users

- Rich skeleton generation $\Longrightarrow$ automatic translation not only of DRa parts but also of CGa parts (algorithm in preperation)
- Visualisation of the annotated content
- Building a library of mathematical texts and visualisation of the library (see links between documents)