

# Patterns in language for POS disambiguation in a style checker

Correct: The input is correct.  
Not correct: You must input the data.

Starting check in English (American)...

**1. Line 2, column 23**

**Message:** Microsoft style: Make sure that 'input' is a noun. [\(deactivate\)](#)

**Context:** ...input is correct. Not correct: You must **input** the data.

# What this presentation is about

A lookup tool is not sufficient

LanguageTool: customizable structure

Finding misused terms: structure of a grammar rule

Part of speech disambiguation:

nouns, verbs, adjectives

Some difficult problems

How good are the rules?

Demonstration

Questions: interrupt and at the end

# A lookup tool is not sufficient

Correct: · The · **input** · is · satisfactory. ¶

Not · correct: · You · must · **input** · the · data. ¶

The lookup tool has limits:

- Slow (5 minutes to check a 50-page document)
- No explanation of problem
- No linguistic intelligence.

# LT is open-source proofreading software

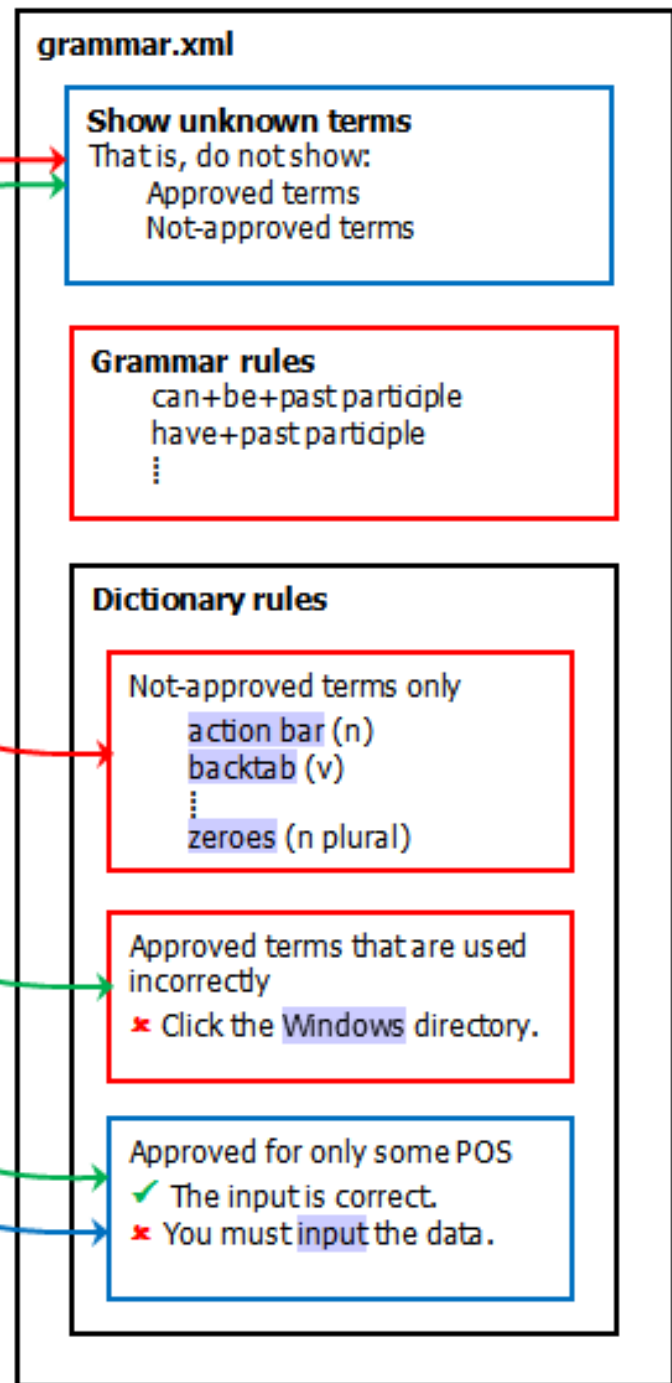
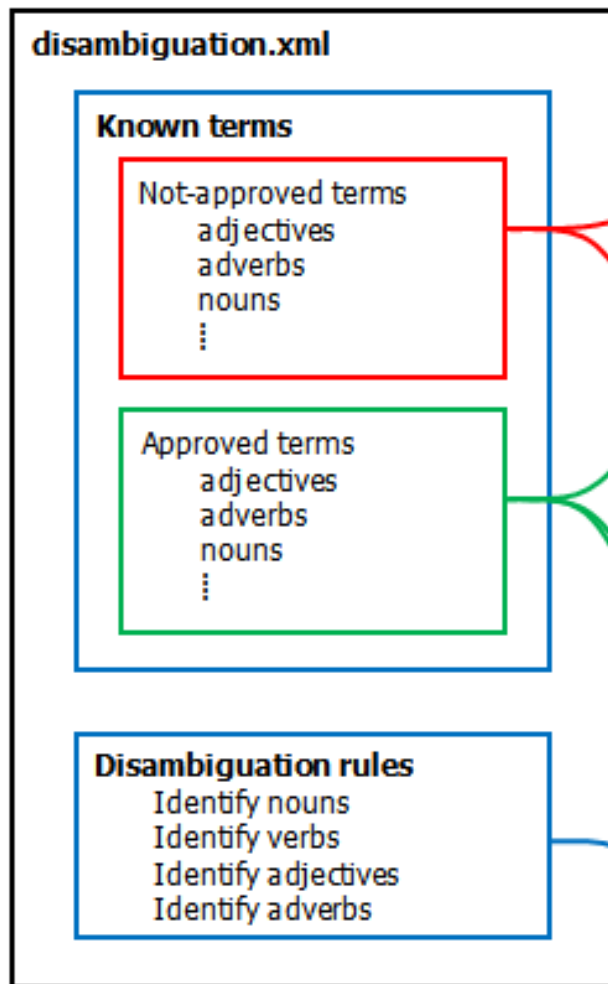
LanguageTool: [www.languagetool.org](http://www.languagetool.org).

LT is fully customizable.

LT has rules for style and for grammar.

Can embed LT into other software.

Term checker uses LT: [www.simplified-english.co.uk](http://www.simplified-english.co.uk).



Disambiguation.xml: what a term is

Grammar.xml: the problem with a term

# The structure of a grammar rule

```
<rule id="MS_POS_INPUT" name="Microsoft, noun approved: input">
  <pattern>
    <token>input<exception postag="IS_NOUN"/></token>
  </pattern>
  <message>Microsoft. Make sure that '<match no="1"/>' is a noun.</message>
  <example type="incorrect"><marker>Input</marker> the data.</example>
  <example type="incorrect">If you <marker>input</marker> the data...</example>
  <example type="correct">The <marker>input</marker> was correct.</example>
</rule>
```

# The alternative method is not good

- 1) `<token>input<exception  
                  postag="IS_NOUN" /></token>`
- 2) `<token postag="IS_VERB">input</token>`

Disambiguation is not 100% accurate.

Primary design decision: no false negatives.

Option 2 is not safe.

Alternative: [www.acrolinx.com/checking.html](http://www.acrolinx.com/checking.html).

# Typical types of conflict in MMoS

Approved	Not approved	Example
verb	noun	install
noun	verb	input
adjective	verb	checked

Grammarians use language patterns to parse text.



# Simple POS disambiguation for nouns (1)

In 'an + X + was', X is a noun.

Use sets of simple patterns to disambiguate.

General or specific rules:

- Specific: The X was; The X is; Some X takes
- General: MODIFIER + X + VERB

Trade-off:

- Specific: not practical
- General: need fewer rules, but sometimes get disambiguation errors.

# Simple POS disambiguation for nouns (2)

Microsoft style: Use 'input' as a noun.

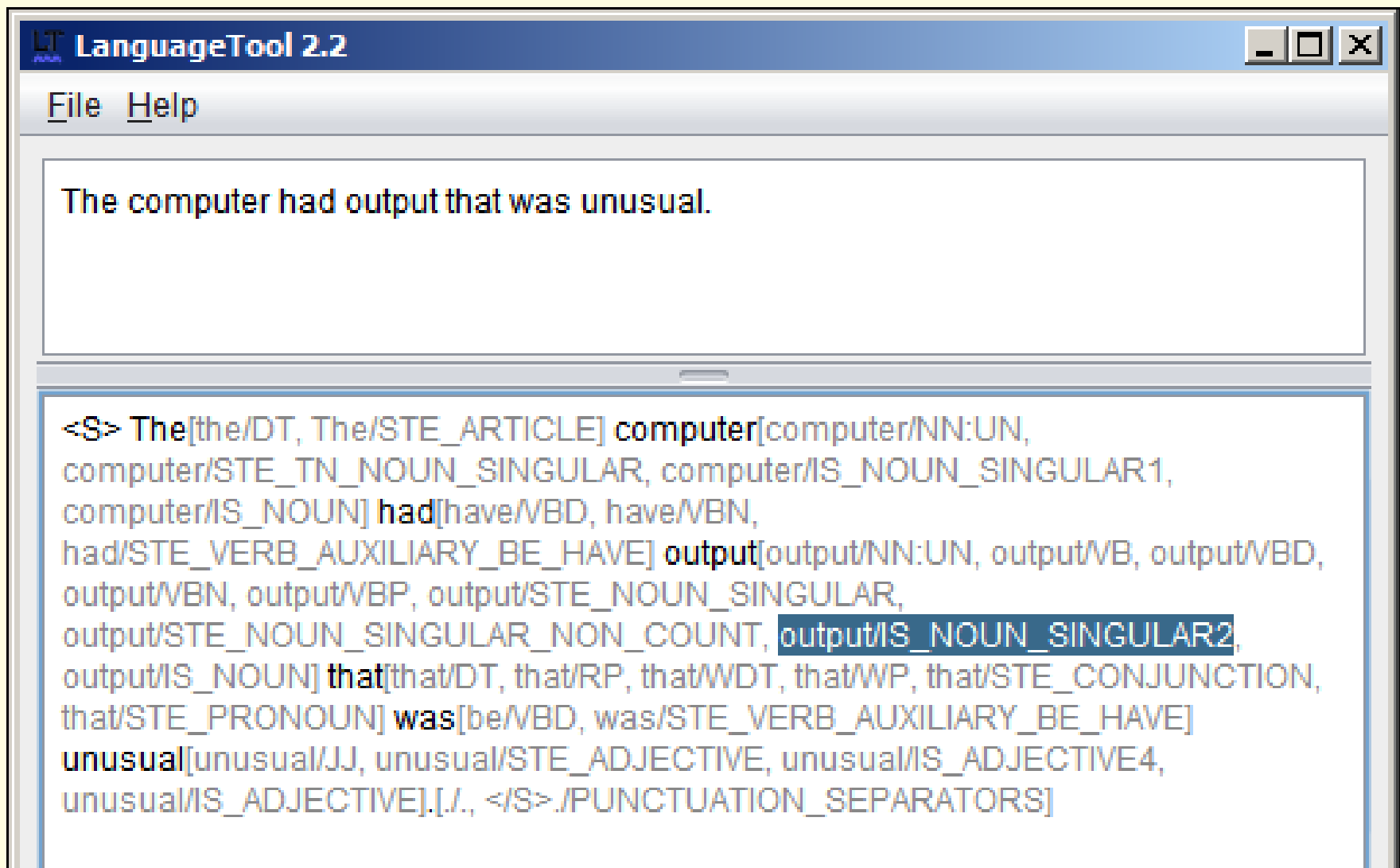
- ✗ You must **input** the data.
- ✓ The computer had **input** that was unusual.
- ✓ The device had **inputs** that were unusual.

Rule: HAVE + NOUN SINGULAR NON-COUNT + THAT|WHICH

Rule: HAVE + NOUN PLURAL + THAT|WHICH

Rules are in groups for nouns, adjectives, verbs.

# Example of postags in LT



The screenshot shows a window titled "LanguageTool 2.2" with a menu bar containing "File" and "Help". The main text area contains the sentence "The computer had output that was unusual." Below this, a detailed list of part-of-speech tags is shown for each word in the sentence. The word "output" is highlighted in blue in the original image.

The computer had output that was unusual.

<S> The[the/DT, The/STE\_ARTICLE] computer[computer/NN:UN, computer/STE\_TN\_NOUN\_SINGULAR, computer/IS\_NOUN\_SINGULAR1, computer/IS\_NOUN] had[have/VBD, have/VBN, had/STE\_VERB\_AUXILIARY\_BE\_HAVE] output[output/NN:UN, output/VB, output/VBD, output/VBN, output/VBP, output/STE\_NOUN\_SINGULAR, output/STE\_NOUN\_SINGULAR\_NON\_COUNT, output/IS\_NOUN\_SINGULAR2, output/IS\_NOUN] that[that/DT, that/RP, that/WDT, that/WP, that/STE\_CONJUNCTION, that/STE\_PRONOUN] was[be/VBD, was/STE\_VERB\_AUXILIARY\_BE\_HAVE] unusual[unusual/JJ, unusual/STE\_ADJECTIVE, unusual/IS\_ADJECTIVE4, unusual/IS\_ADJECTIVE].[./., </S>./PUNCTUATION\_SEPARATORS]

# Simple POS disambiguation for verbs

Microsoft style: Use 'install' as a verb:

- ✗ If the **install** was a problem...
- ✓ You must **install** all the...
- ✓ You must never **install** all the...

If X can be used as a verb, then in 'must + X',  
X is a verb.

General rule: MODAL AUXILIARY VERB + X.

If a counter-example exists, then 3 options:

- Re-write the rule for better disambiguation.
- Make X an exception to the rule.
- Do nothing.

# Simple POS disambiguation for adjectives

Microsoft style: Use 'checked' as an adjective:

- ✘ If you **checked** the box...
- ✔ If the **checked** command is...

In 'ARTICLE + X + NOUN', X is an adjective.

# A difficult problem: noun or verb?

Noun cluster: plastic bucket, fire engine, oil sample

Sub-pattern:

NOUN SINGULAR + NOUN PLURAL + END OF SENTENCE

Is the last word a noun or a verb?

- Use the metal covers. Noun
- The device analyses the oil samples. Noun
- The alarm covers. Noun
- The alarm sounds. Ambiguous
- The oil system leaks. Ambiguous
- The electrical equipment sparks. Ambiguous

What pattern identifies the nouns?

# A difficult problem: noun or verb?

For the previous sub-pattern, if a word can be both a noun and a verb AND

- If the verb is transitive only, the word is a noun.
- If the verb is intransitive, the POS is ambiguous.

Transitive = has an object:

The metal covers the hole.

Intransitive = does not have an object:

She laughs.

Some verbs are both transitive and intransitive:

- Transitive: The heat melted the snow.
- Intransitive: The snow melted.

# Some text is always ambiguous

Microsoft style avoids the passive voice.

- ✗ The wire **was disconnected** by the technician.
- ✗ The wire **was disconnected** quickly.
- ? The wire **was disconnected**.
  - Passive voice?
  - Adjective *disconnected* describes the wire? (Compare, "The wire was **dirty**.")

Real-world knowledge:

- The water **was drunk**. (Passive voice)
- The waiter was **drunk**. (Adjective)



# How good are the rules?

Context is important:

- Winemaker: **The must** is contaminated.
- Informal: Warm clothes are **a must** in cold weather.

Term checker:

- Approximately 150 sentences a second
- Approximately 6% false positives for STE rules.

Demonstration: Microsoft examples.

# Questions

Questions?

[mike@techscribe.co.uk](mailto:mike@techscribe.co.uk)

[www.techscribe.co.uk](http://www.techscribe.co.uk)