Semi-automatic Tool for Thai Corpora Construction

Nattapol KRITSUTHIKUL
Dr. Krit KOSAWAT
Dr. Thepchai SUPNITHI

Background

- Corpus เป็นสิ่งจำเป็นสำหรับงานวิจัยด้าน NLP
- มีความต้องการ Corpus ขนาดใหญ่
  เพื่อนำมาใช้รองรับงานวิจัยด้าน NLP
- เครื่องมือสำหรับการวิจัยทาง NLP สำหรับภาษาไทย ยังมีอยู่น้อย

Problems

- การสร้าง Corpus ที่มีขนาดใหญ่ขึ้น ต้องการเวลาในการสร้างมาก
- ลักษณะพิเศษของภาษาไทย ที่ไม่มีเครื่องหมายกำกับขอบเขตของคำ ทำให้การสร้าง Corpus ภาษาไทย มีความยุ่งยากมากกว่าการสร้าง Corpus ภาษาอังกฤษ
- Corpus มีรูปแบบหลากหลาย ทำให้แลกเปลี่ยนข้อมูลได้ยาก

Motivation

- ต้องการเครื่องมือที่ช่วยสร้าง Corpus ขนาดใหญ่ได้ในเวลาที่เหมาะสม
- สามารถจัดเก็บข้อมูลในรูปแบบมาตรฐาน XML

Why “Semi-Automatic”?

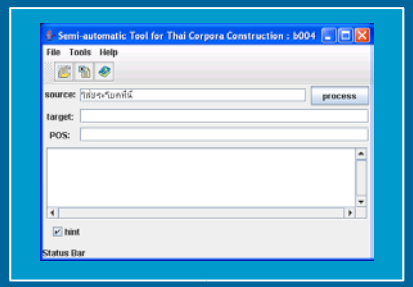
- เครื่องมืองานวิจัยทาง NLP สำหรับภาษาไทย
  ยังไม่สามารถกำกับขอบเขตของคำได้อย่างถูกต้องสมบูรณ์
- จำเป็นต้องใช้มนุษย์ในการตรวจสอบความถูกต้อง
  - ตัดสินใจเลือกขอบเขตของคำที่ดีที่สุด
  - ตัดสินใจเลือกหน้าที่ของคำที่ดีที่สุด

Features

- ความสามารถแบบอัตโนมัติ
  - สามารถแบ่งขอบเขตของคำได้โดยอัตโนมัติ
  - สามารถกำหนดหน้าที่ของคำได้โดยอัตโนมัติ
  - บันทึกข้อมูลในรูปแบบ XML
- ความสามารถกึ่งอัตโนมัติ
  - สามารถแนะนำขอบเขตของคำแบบต่าง ๆ ให้แก่ผู้ใช้
  - สามารถกำหนดหน้าที่ของคำโดยผู้ใช้ได้

System Architecture

Front End



Back End

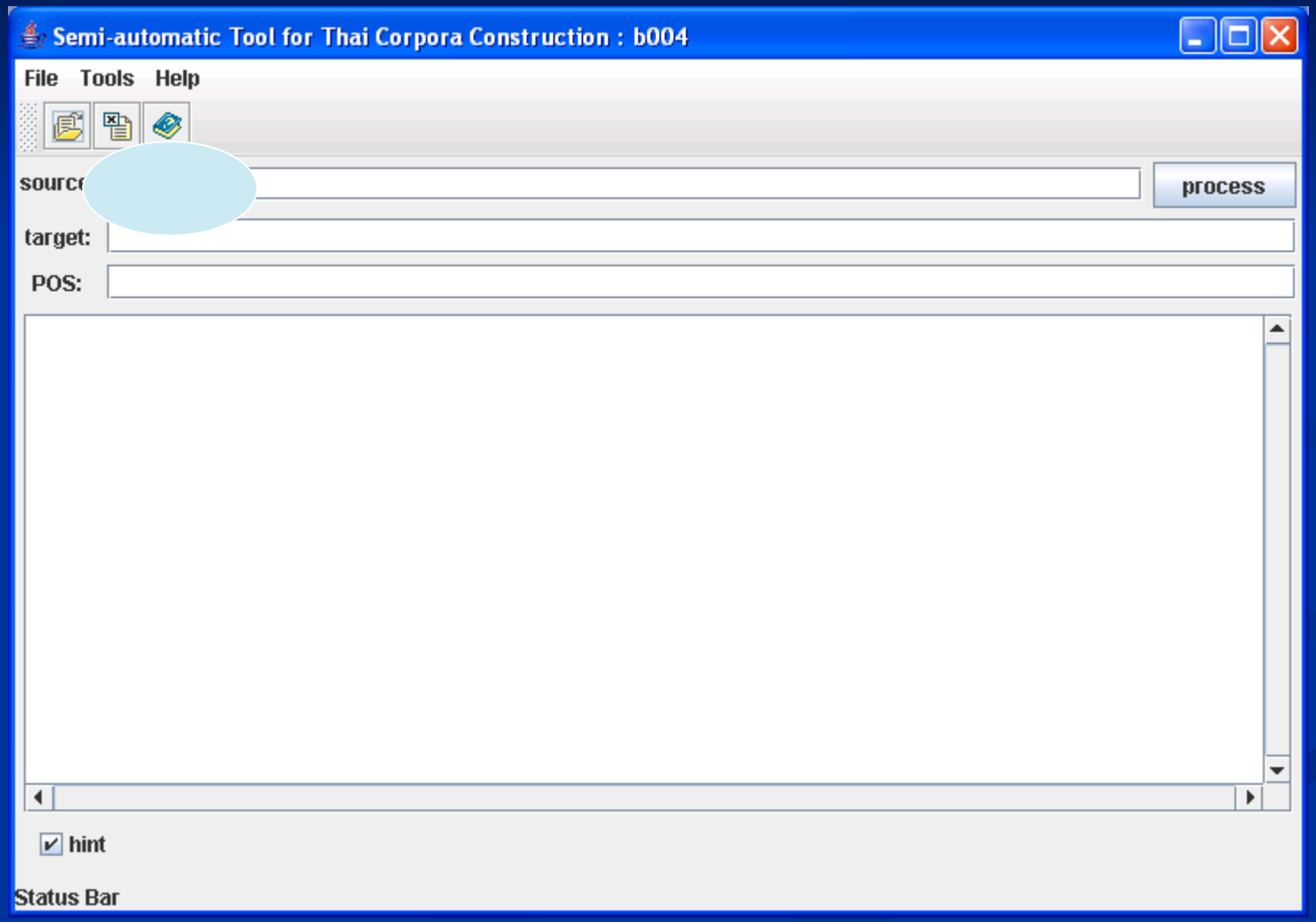
SWATH
XML parser

Semi-automatic tool for Thai Corpora Construction

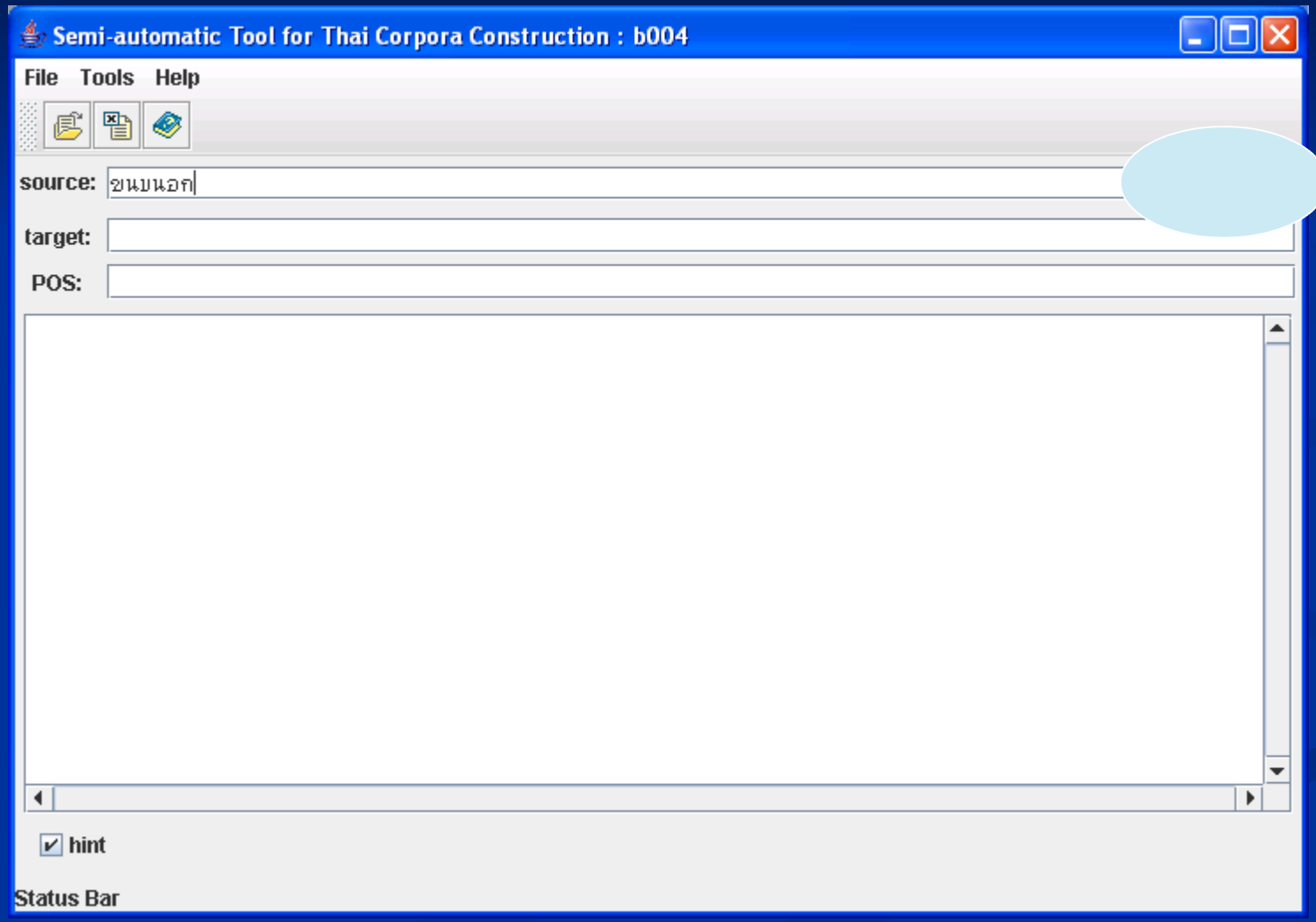
Developer Tools

- Front End
  - Sun Java 1.5.0 update 2 (jdk1.5.0_02)
- Back End
  - Microsoft Visual C++ 6.0 with SP6
  - XML parser (Apache Xerces Apache Xalan Apache FOP)

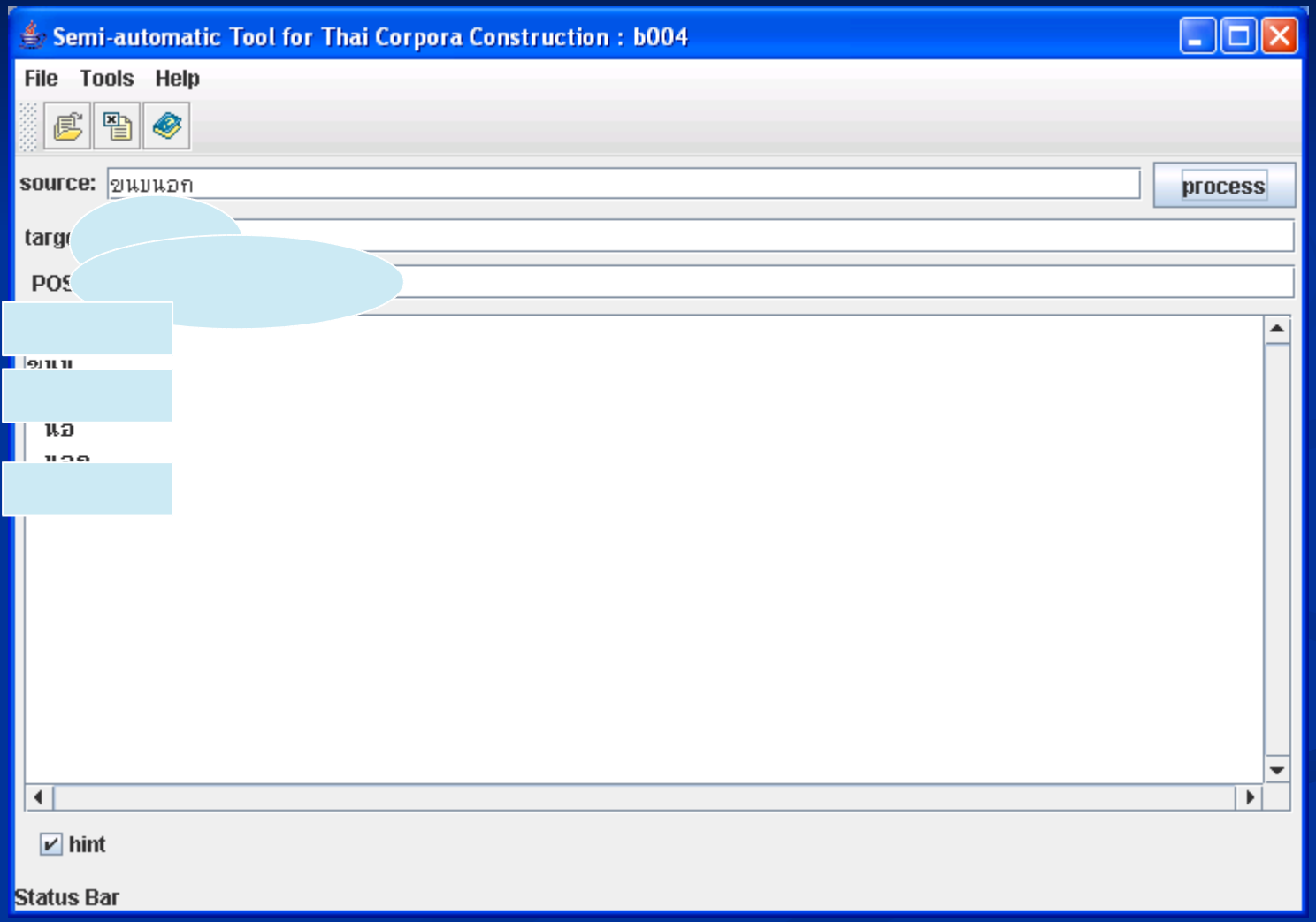
Front End : Main Screen



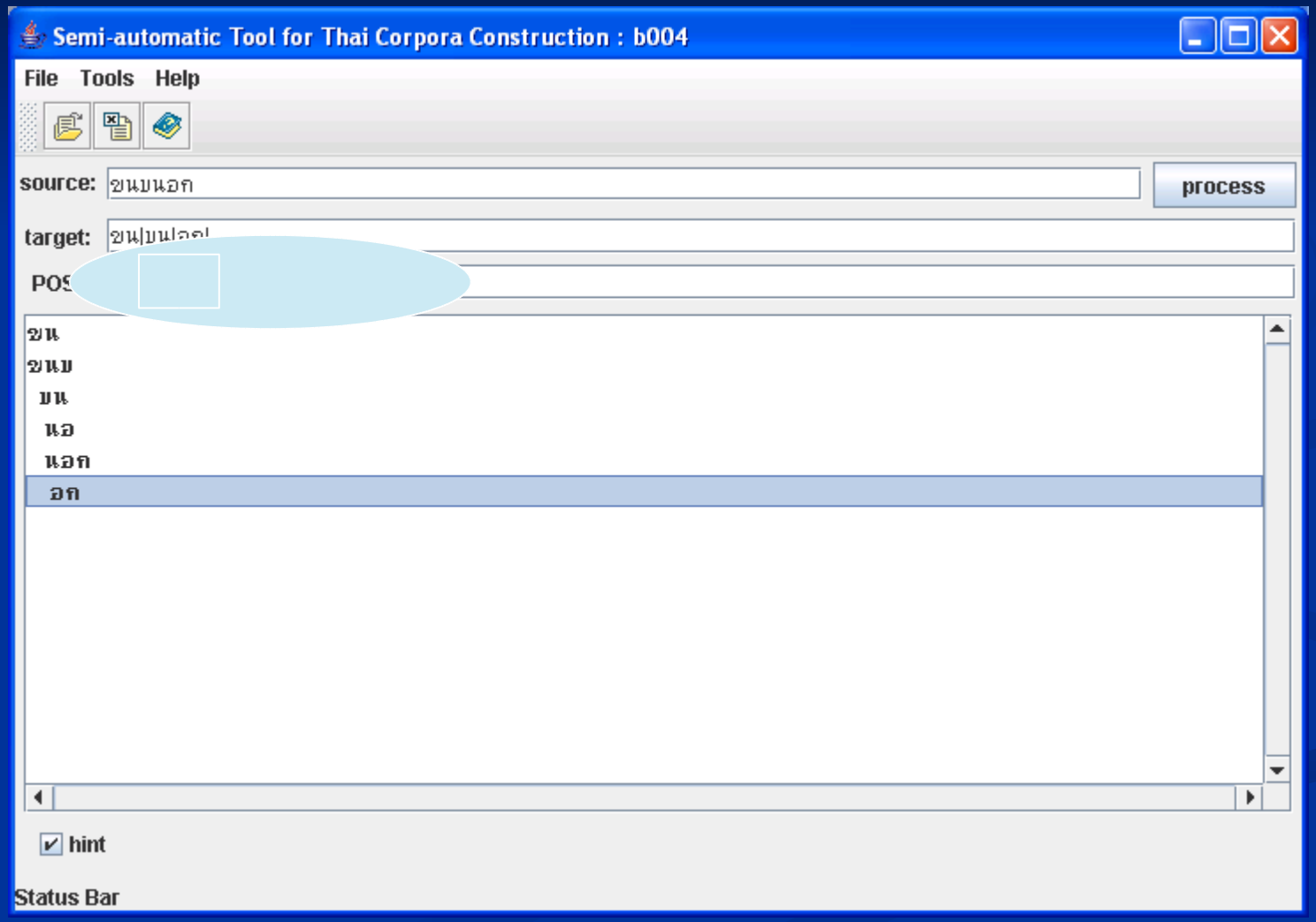
Front End : Input a sentence



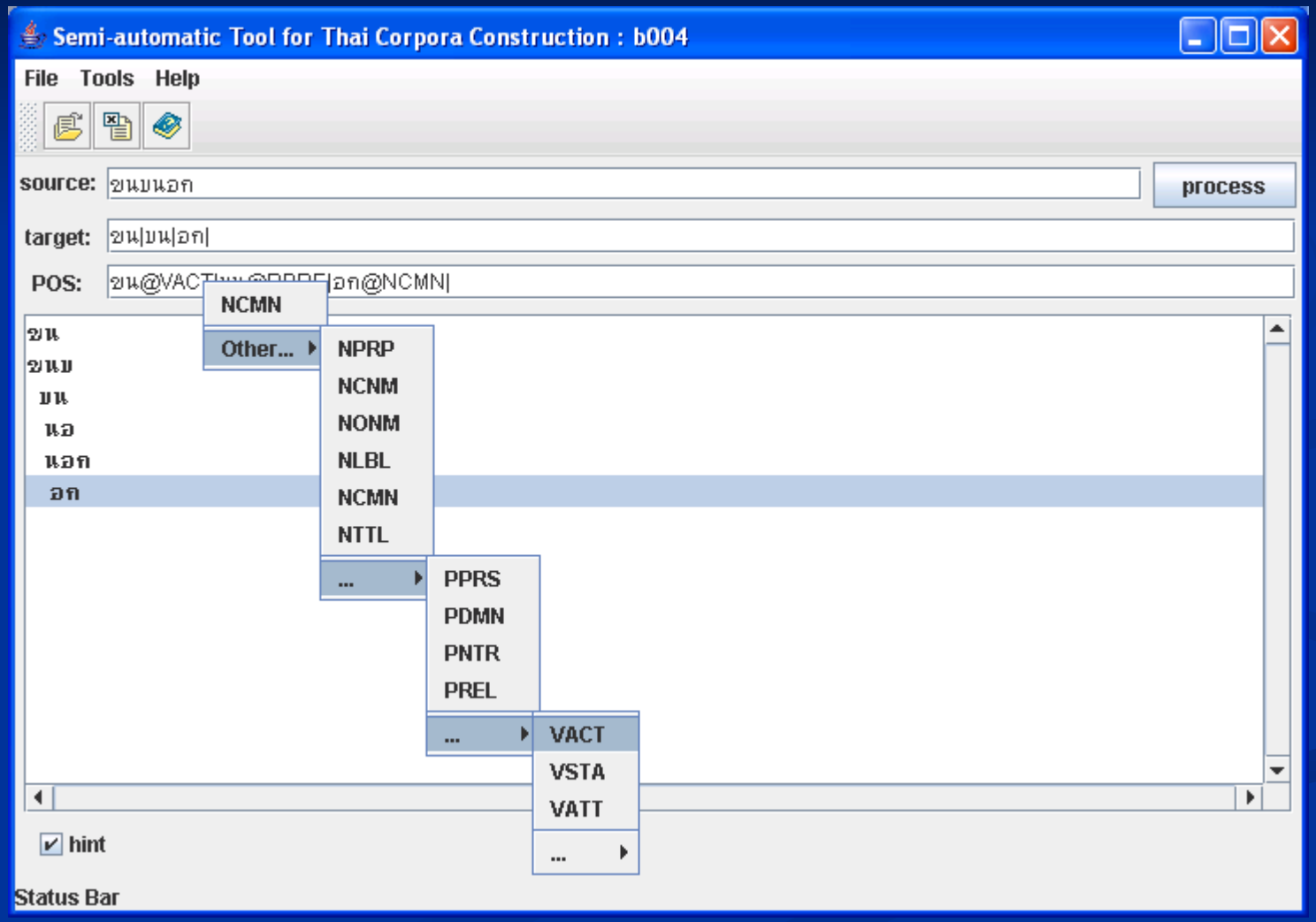
Automatic Word Segmentation



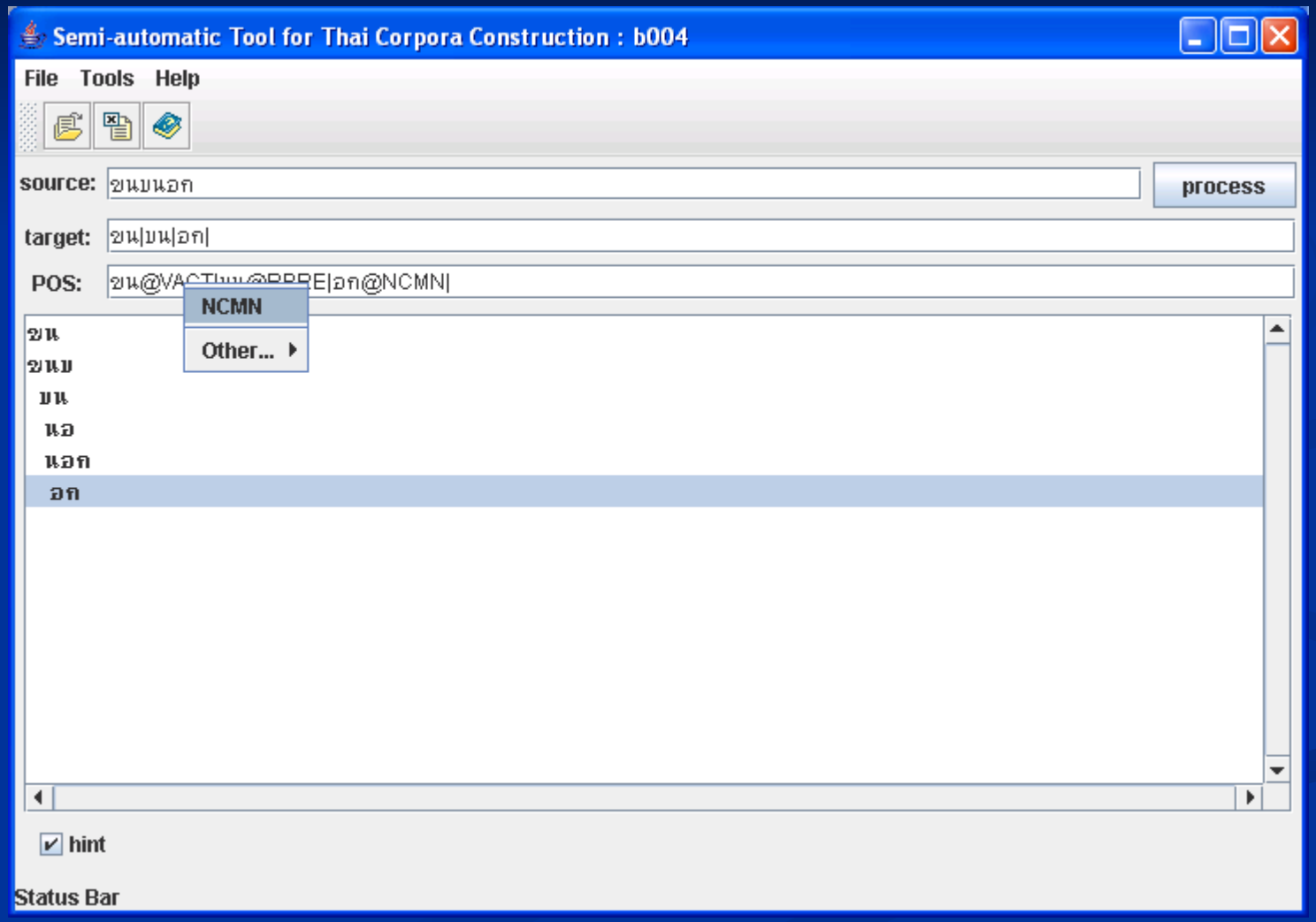
Word Segment Correction



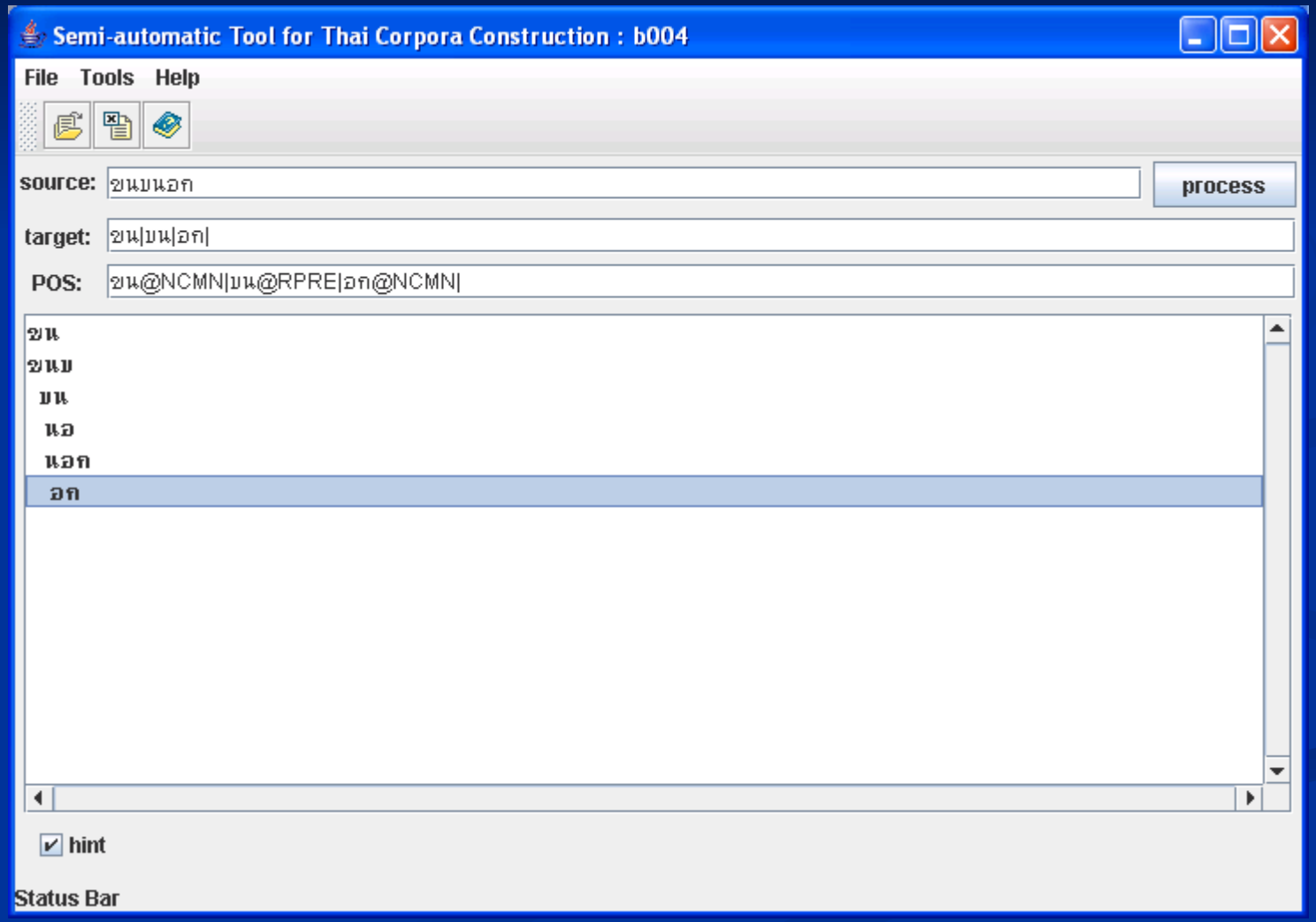
POS Correction



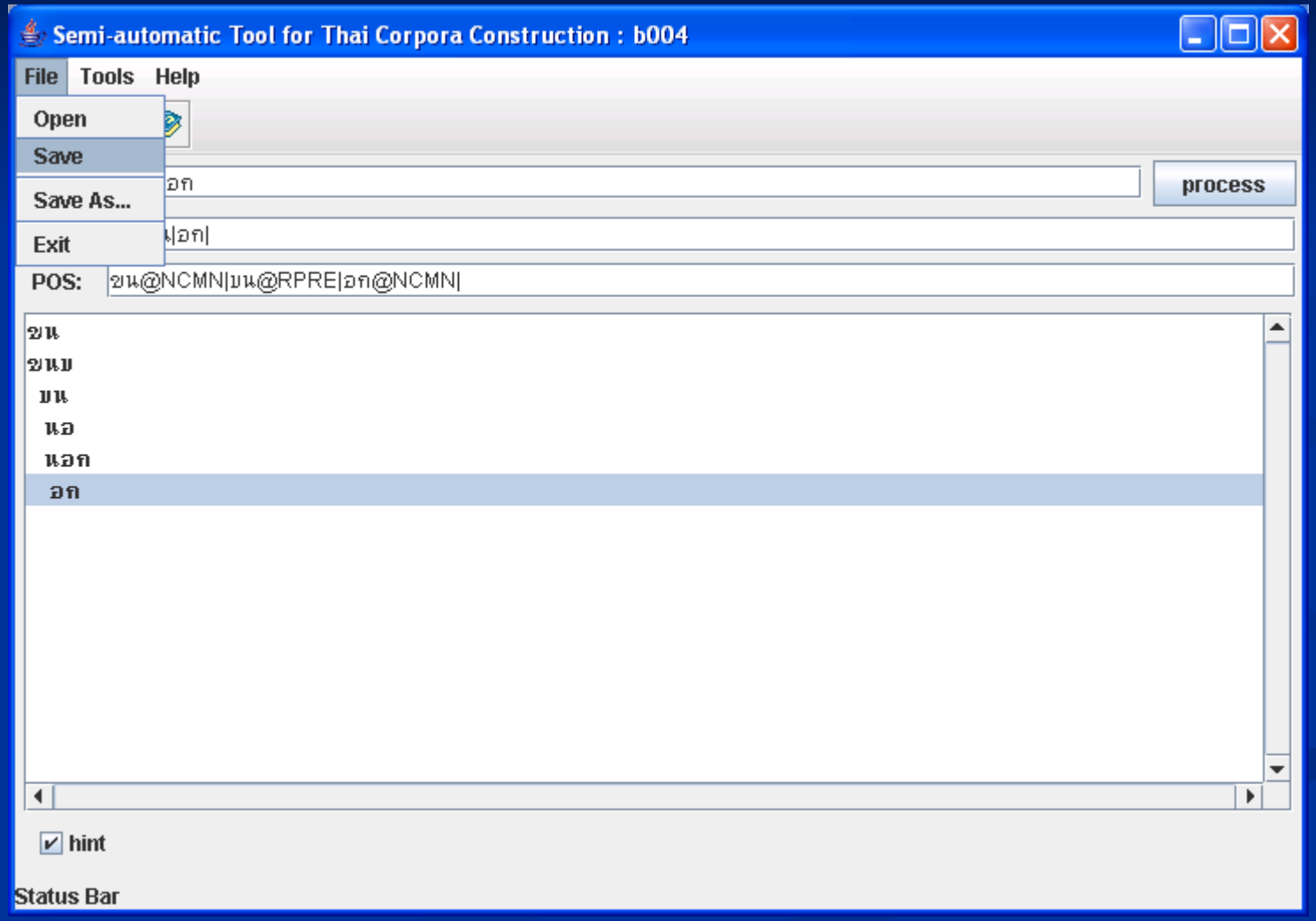
Select POS


Semi-automatic Tool for Thai Corpora Construction : b004
File
Tools
Help
source:
ขนบนอก
process
target:
ขน|บน|อก|
POS:
ขน@VACTbu@RRRE|อก@NCMN|
NCMN
Other...
ขน
ขนบ
บน
นอ
นอก
อก
hint
Status Bar

Result


Semi-automatic Tool for Thai Corpora Construction : b004
File
Tools
Help
source:
ขนบนอก
process
target:
ขน|บน|อก|
POS:
ขน@NCMN|บน@RPRE|อก@NCMN|
ขน
ขนบ
บน
นอ
นอก
อก
hint
Status Bar

Save to XML


Semi-automatic Tool for Thai Corpora Construction : b004
File
Tools
Help
Open
Save
Save As...
Exit
อก
process
น|อก|
POS:
ขน@NCMN|ขน@RPRE|อก@NCMN|
ขน
ขนข
ขน
นอ
นอก
อก
hint
Status Bar

XML Data file

```
<?xml version="1.0" encoding="utf-8"?>
<semiauto version="1.00">
    <sentence>ขนบนอก</sentence>
    <words>
        <word id="0">
            <content>ขน</content>
            <pos>NCMN</pos>
            <begin>0</begin>
            <end>2</end>
        <word>
        <word id="1">
            <content>บน</content>
            <pos>RPRE</pos>
            <begin>2</begin>
            <end>4</end>
        <word>
        <word id="2">
            <content>อก</content>
            <pos>NCMN</pos>
            <begin>4</begin>
            <end>6</end>
        <word>
    </words>
</semiauto>
```

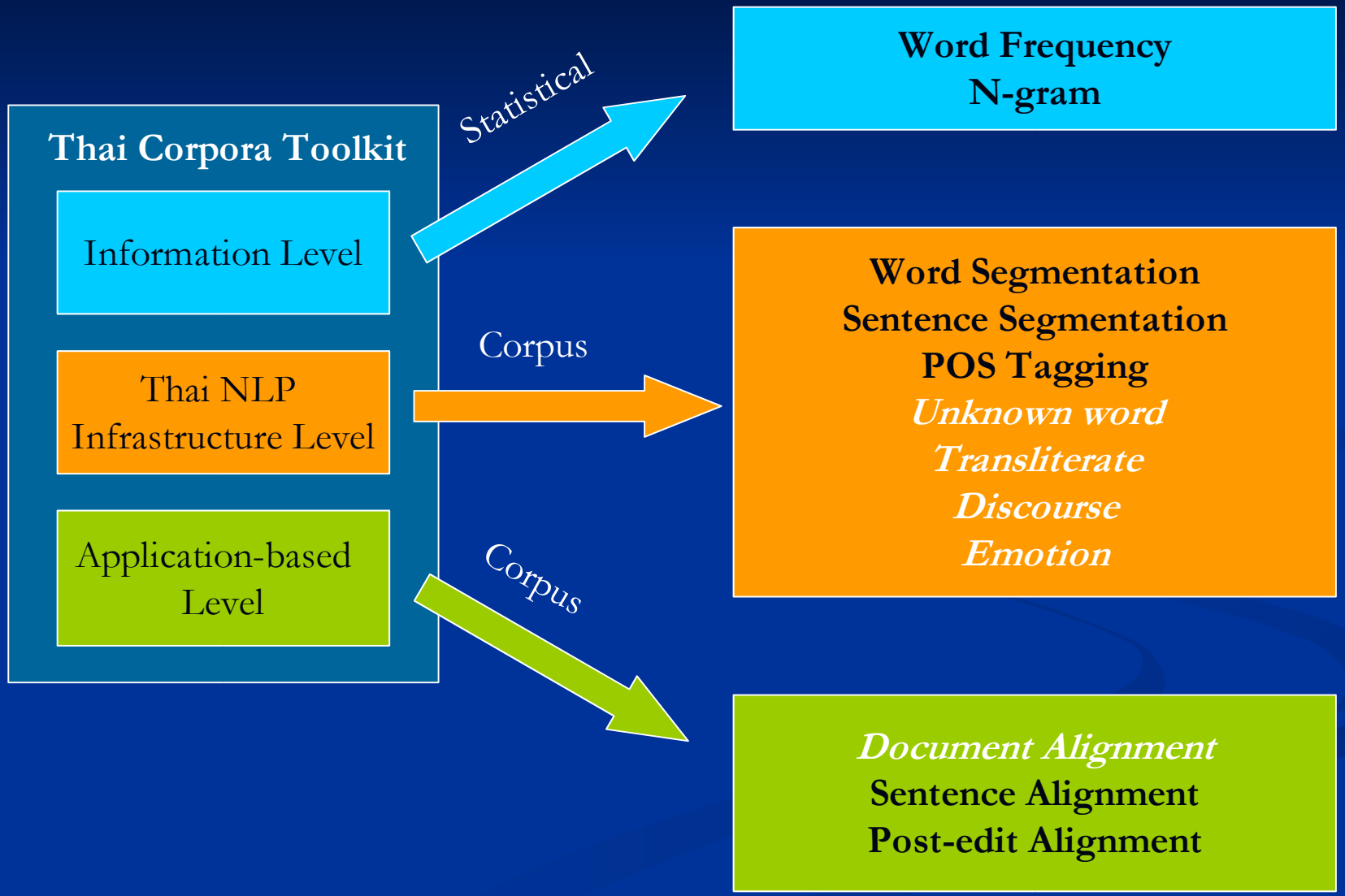
Back End

- SWATH : Smart Word Analysis for THai
- XML parser
  - Apache Xerces
  - Apache Xalan
  - Apache FOP

Conclusion

- เครื่องมือกึ่งอัตโนมัติ
  ที่มีความสามารถในการกำกับขอบเขตของคำเพื่อใช้ในการสร้าง Corpus สำหรับภาษาไทย
  - สามารถใช้งานง่าย
  - สามารถจัดเก็บข้อมูลในรูปแบบ XML ทำให้สามารถนำไปใช้ร่วมกับ Corpus อื่นได้ง่าย

Future Works...



Thank You