Lexical Tools
ASCII Conversion

Dr. Chris J. Lu

The Lexical Systems Group
NLM. LHNCBC. CGSB

March, 2011

Table of Contents

- Introduction
- ASCII conversion
  - Character
  - Document
  - Corpus
  - Software/APIs
  - Example
- Questions

ASCII Character Set

- **ASCII:** American Standard Code for Information Interchange
- Contains 128 7-bit coded characters
- Value range: U+0000 ~ U+007F
- Includes:
    - alphabetic characters: A, B, C, …
    - numeric characters: 0, 1, 2, 3, …
    - control characters: ESC, FS, CR, …
    - graphic characters: #, $, %, &, *, (,), ..
- The most common used standard code (before Unicode)

Unicode

- A character encoding specification published by the Unicode Consortium
- Includes all of the major world’s writing systems
- Becomes the industry standard
- Allows data to be transported through different systems
- Very useful when dealing with multilingual NLP
- Latest version Unicode 6.0.0, 2011

Unicode Transformation Format

- Unicode Encoding
  - Including UTF-7, UTF-8, UTF-16, UTF-32

- UTF-8 has become the dominant character encoding
  - Backward-compatible with ASCII
  - Avoiding the complications of endianness
  - No need to use byte order marks (BOM)

Lexicon & Lexical Tools

- Released in UTF-8 format since 2006
- Provides functions to convert UTF-8 to ASCII
  - Character
  - Text
  - Document

Why ASCII Conversion?

- Non-ASCII Unicode are commonly seen even in English documents, such as “Déjà Vu “, “Café”, “resumé”, etc.

- Some NLP projects still only deal with ASCII

The Challenges

- Not one-to-one mapping:
  - Many to one: å, â, ã, á, à, ä to a
  - One to many: © to ![COPYRIGHT SIGN]!, (c), or just simply removed
  - One to none: French borrowing “divorcé” means a man who is divorced. This word has no pure ASCII spelling variant in Webster’s Dictionary, while the converted ASCII word, “divorce”, is another closely related word
- Misused Unicode characters (before the conversion)
  - μ (mu, U+03BC) and µ (micro sign, U+00B5)
  - ß (Sharp S , U+00DF) and β (beta, U+03B2)
  - ¶ (Pilcrow Sign, U+00B6) and π (PI, U+03C0)
- Wrong conversions (meaning changed)
  - © to (c): copyright or cellular phone number?
  - divorcé to divorce

Conversion Guidelines

- Preserve semantic and/or graphic representation
- Example ™:
  - Graphic: TM
  - Semantic: ![TRADE MARK SIGN]!
  - Graphic and Semantic: (TM), or (tm)
  - NLP: empty string, consider ™ as a stopword

- Different NLP applications might apply different methods due to different requirements and objectives
- There is no best method for ASCII conversion

Character Conversion

- Strip diacritics:

  å, â, ã, á, à, ä, ê, é, è, ë, î, í, ì, ï, ô, õ, ó, ø, ò, ö, û, ú, ù, ü, ý, ç, ñ, etc.
- Split ligatures:

  Æ, æ, Œ, , œ, ﬀ, ﬂ, ﬃ, etc.
- Punctuation mapping:

  “double quotation”, ‘single quotation’, – , -, etc.
- Symbols mapping:

  © ,®, ™, °, ÷, ≤, ≥, etc.

- Combinations:

  ǽ [U+01FD], ǅ [U+01C5], ¾ [U+00BE], etc
- Others:

  α, β, etc

Lexical Tools

- Unicode related functions (flow components)

LVG Flow	Description	Input (UTF-8)	Output (ASCII)
-f:q	Strips diacritic	Déjà Vu	Deja Vu
-f:q0	Symbols & punctuation	“Quote”	"Quote"
-f:q1	Unicode mapping	⅔	2/3
-f:q2	Splits ligatures	spælsau	spaelsau
-f:q3	Unicode names	©	![COPYRIGHT SIGN]!
-f:q4	Unicode Synonym	μ (mu, U+03BC)	µ (Micro sign, U+00B5)
-f:q5	Normalize Unicode (-f:q7:q3)	UMLS®	UMLS![REGISTERED SIGN]!
-f:q6	Normalize Unicode w Synonyms (-f:q4:q7:q3)	UMLS®	UMLS![REGISTERED SIGN]!
-f:q7	Core Norm (recursive -f:q0:q1:q2:q)	Ǣ	AE
-f:q8	Strip or Map (not ICU)	Zadaxin™	Zadaxin
-f:q8	Strip or Map (not ICU)	α	alpha

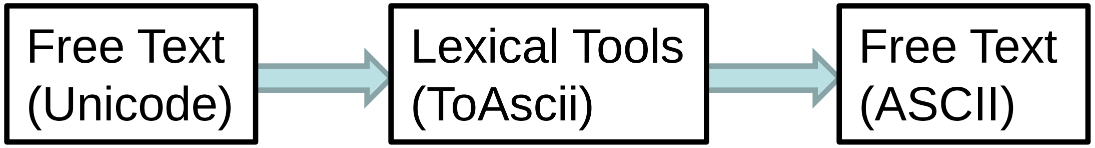
Lexical Tools (Cont.)

- Pure ASCII conversion

LVG Flow(s)	Desc.	Pure ASCII	Outputs
-f:q5	Normalize Unicode	Yes	Single
-f:q6	Normalize Unicode with Synonyms	Yes	Single
-f:N	Normalize	Yes	Multiple
-f:N3	Lui-Norm	Yes	Single
-f:q7:q8	Serial Flows	Yes	Single
ToAscii	ASCII conversion	Yes	Single

Text Conversion

- Many different ways for ASCII conversion
- The SPECIALIST Lexical Tools
  - Provides various powerful functions
  - Is configurable according to the specifications
  - Use ToAscii

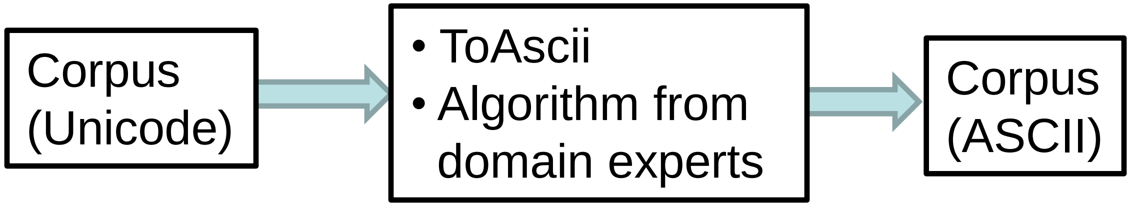


Free Text (Unicode)

Lexical Tools (ToAscii)

Free Text (ASCII)

Corpus Conversion

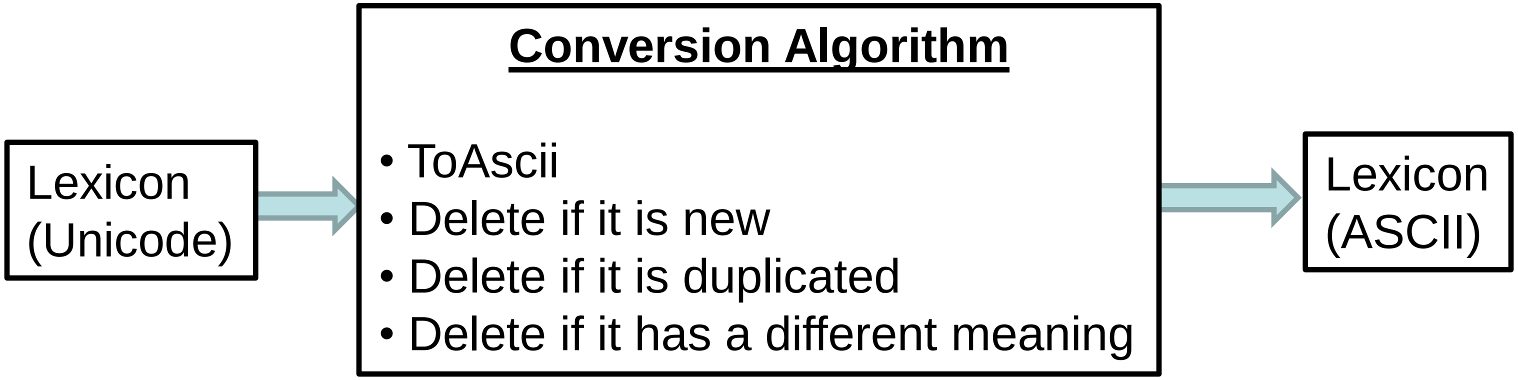


Corpus
(Unicode)

- ToAscii
- Algorithm from domain experts

Corpus
(ASCII)

Corpus Conversion - Lexicon



Lexicon (Unicode)

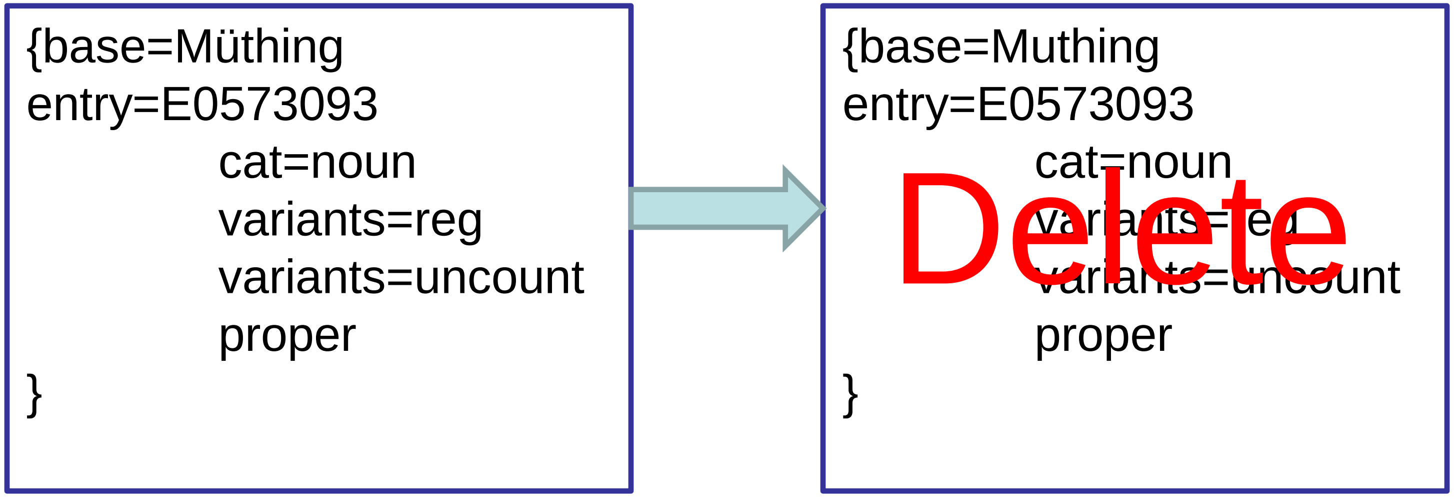
Conversion Algorithm

- ToAscii
- Delete if it is new
- Delete if it is duplicated
- Delete if it has a different meaning

Lexicon (ASCII)

Delete: If New

- Delete the conversion if it is new (not known to Lexicon)
  - Theoretically, the ASCII Lexicon is a subset of Unicode Lexicon since ASCII is a subset of Unicode
  - All converted bases should be known to (contained inside) Lexicon
- Example - Müthing” [E0573093]:
  - The record is deleted (“Muthing” is not know to Lexicon)



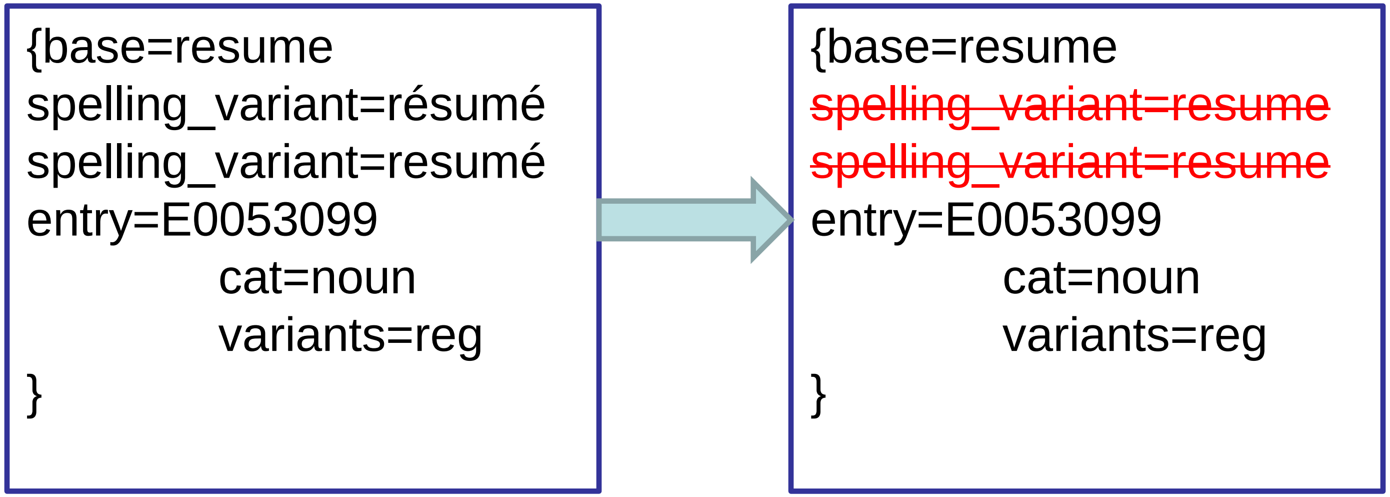
```
{base=Müthing
entry=E0573093
		cat=noun
		variants=reg
		variants=uncount
		proper
}
```

```
{base=Muthing
entry=E0573093
		cat=noun
		variants=reg
		variants=uncount
		proper
}
```

Delete

Delete: If Duplicated

- Delete the conversion if it is a duplication
- Example – resume [E0053099]
  - Spelling variants are removed

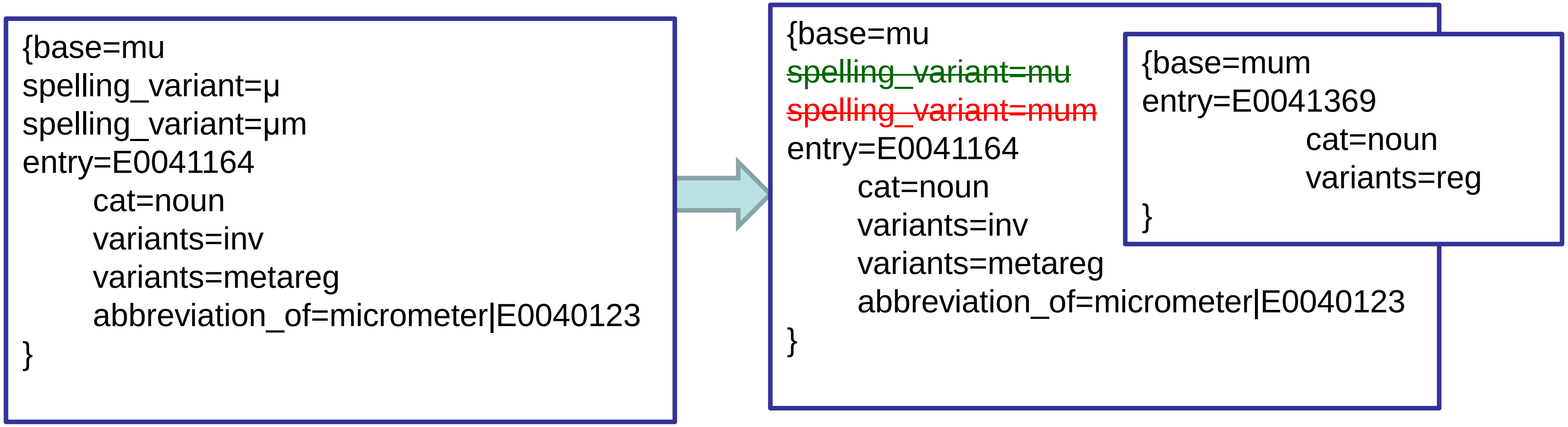


{base=resume
spelling_variant=résumé
spelling_variant=resumé
entry=E0053099
    cat=noun
    variants=reg
}

{base=resume
~~spelling_variant=resume~~
~~spelling_variant=resume~~
entry=E0053099
    cat=noun
    variants=reg
}

Delete: If Meaning Changed

- Delete the conversion if it has a different meaning
- Example – mu [E0041164]:
  - Spelling variant “µm” is deleted because its ASCII conversion, “mum” [E0041369], is a different record

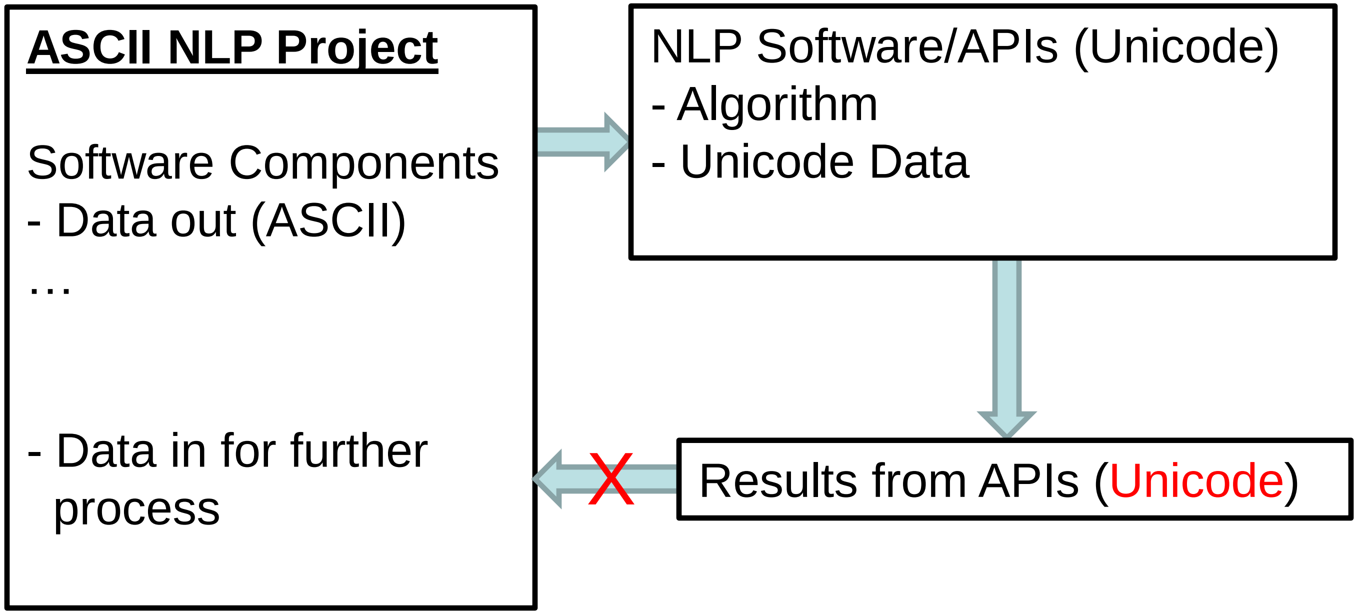


```
{base=mu
spelling_variant=µ
spelling_variant=µm
entry=E0041164
        cat=noun
        variants=inv
        variants=metareg
        abbreviation_of=micrometer|E0040123
}
```

```
{base=mu
~~spelling_variant=mu~~
~~spelling_variant=mum~~
entry=E0041164
        cat=noun
        variants=inv
        variants=metareg
        abbreviation_of=micrometer|E0040123
}
```

```
{base=mum
entry=E0041369
                cat=noun
                variants=reg
}
```

NLP Software Conversion



ASCII NLP Project

Software Components
- Data out (ASCII)
…

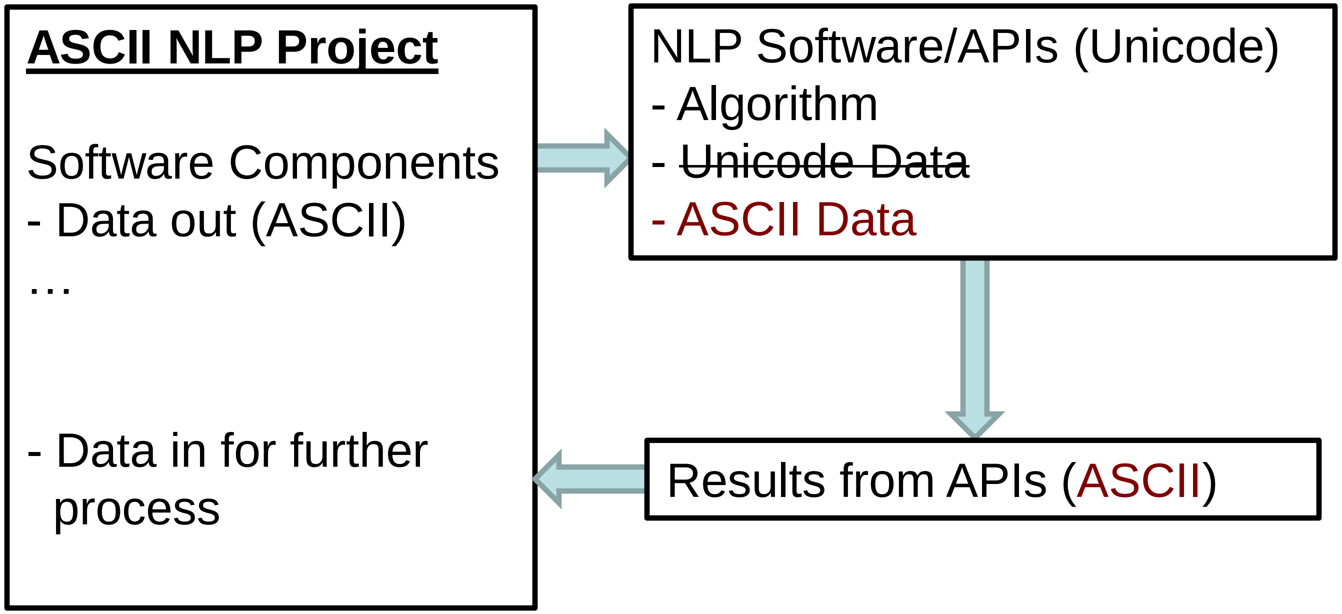
- Data in for further process

NLP Software/APIs (Unicode)
- Algorithm
- Unicode Data

Results from APIs (Unicode)

- Traditional approach
- Interface approach

Traditional Approach



ASCII NLP Project

Software Components
- Data out (ASCII)

…

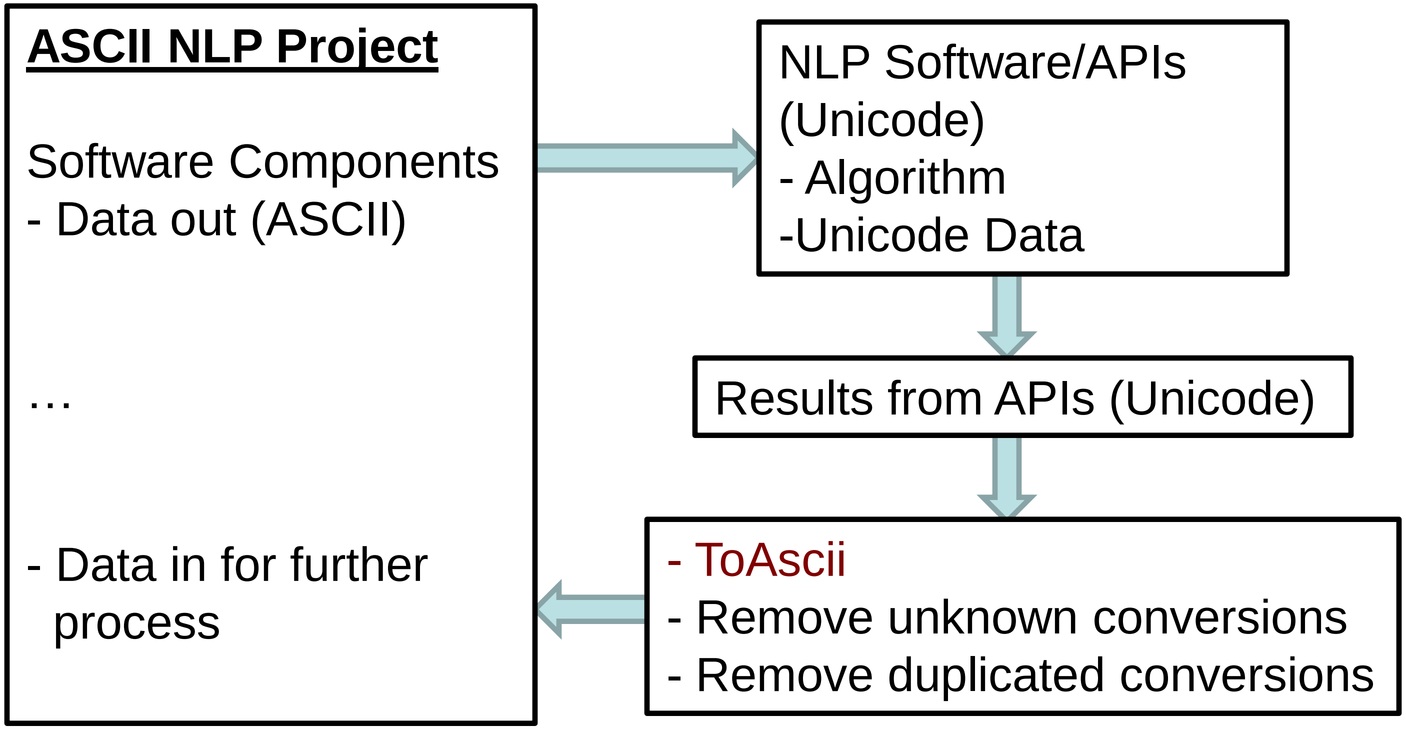
- Data in for further process

NLP Software/APIs (Unicode)
- Algorithm
- ~~Unicode Data~~
- ASCII Data

Results from APIs (ASCII)

- This traditional approach is tedious and not practical

Interface Approach



ASCII NLP Project

Software Components
- Data out (ASCII)

…

- Data in for further process

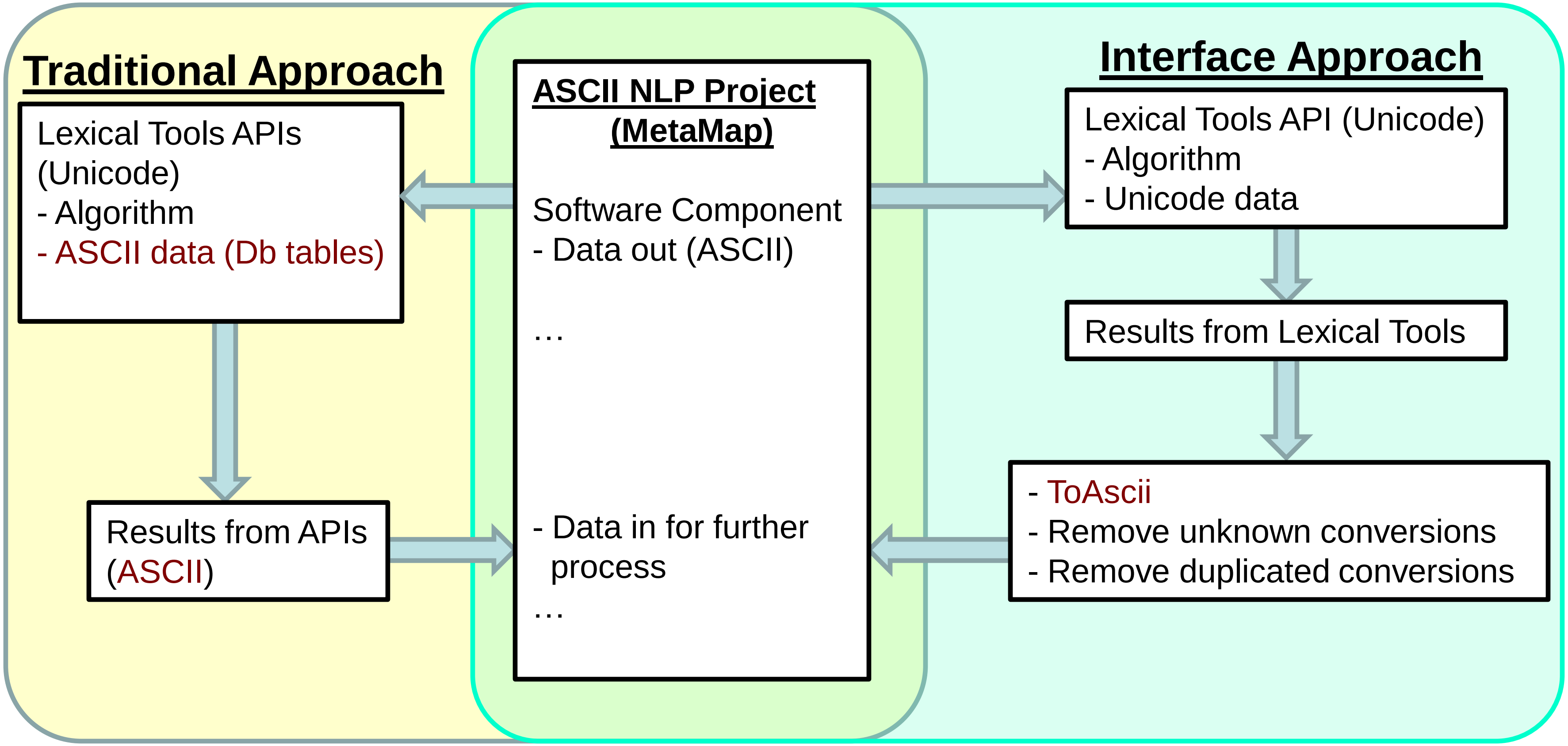
NLP Software/APIs (Unicode)
- Algorithm
-Unicode Data

Results from APIs (Unicode)

- ToAscii
- Remove unknown conversions
- Remove duplicated conversions

- The interface approach is easy and generic

Application Example



Traditional Approach

Lexical Tools APIs (Unicode)
- Algorithm
- ASCII data (Db tables)

Results from APIs (ASCII)

ASCII NLP Project (MetaMap)

Software Component
- Data out (ASCII)

…

- Data in for further process

…

Interface Approach

Lexical Tools API (Unicode)
- Algorithm
- Unicode data

Results from Lexical Tools

- ToAscii
- Remove unknown conversions
- Remove duplicated conversions

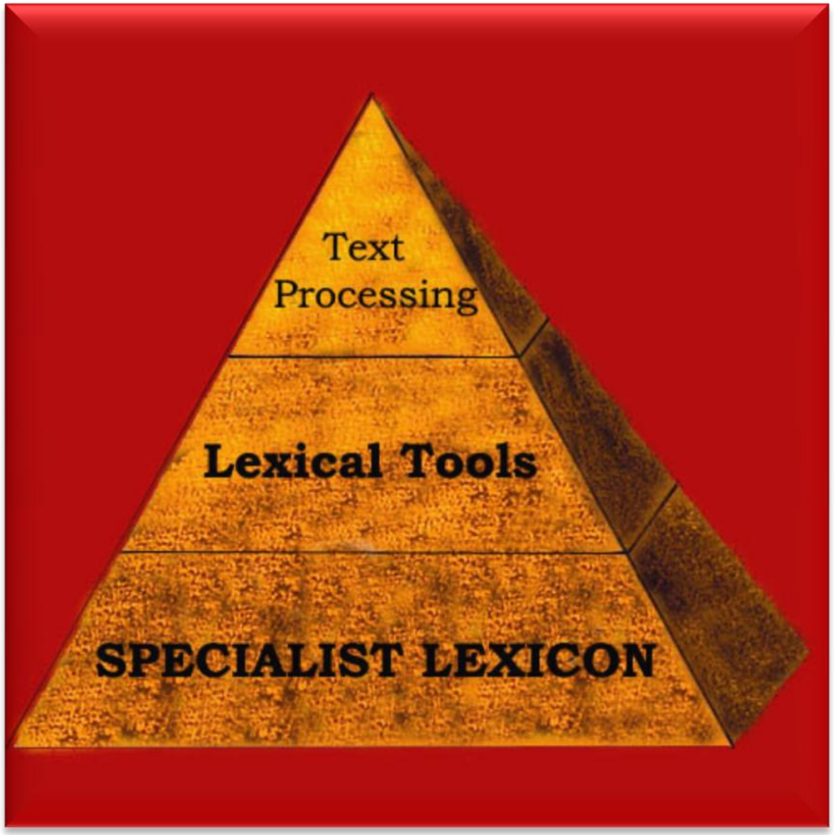
- Identical results from both approaches over 0.5M test cases for 2010 release

References

- Unicode Consortium - http://www.unicode.org
- ICU (International Components for Unicode) - http://site.icu-project.org
- Lexical Tools Unicode Documents -

  http://lexlsrv1.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/designDoc/UDF/unicode/index.html
- Lu, Chris J.; Browne, Allen C.; Divita, Guy, "Using Lexical Tools to Convert Unicode Characters to ASCII", Proceeding of AMIA 2008 Annual Symposium, Nov. 8-12, 2008, Washington DC, p. 1031
- Lu, Chris J. and Browne, Allen C., "Converting Unicode Lexicon and Lexical Tools for ASCII NLP", Submitted for publication in Proceeding of AMIA 2011 Annual Symposium, Oct. 22-16, 2011, Washington DC

Questions



- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov