

# IBM Dictionary & Linguistic Tools (a.k.a. LanguageWare)



**Brian O'Donovan**  
**April 2002**





Lotus®

## Agenda

- What is LanguageWare?
- Where/How is it Used?
- Some examples of difficult problems
- Where are the Development Team?
- What skills are needed?



Lotus

## What is LanguageWare?

- A suite of tools to assist in linguistic analysis
- Initially developed by IBM USA, but now developed by a Globally distributed team managed from Dublin
  - Used in internal and external products
  - Available as a toolkit



Lotus®

## 28 Languages Supported

- Afrikaans
- Arabic
- Catalan
- Simplified Chinese
- Traditional Chinese
- Czech
- Danish
- Dutch (x2)
- English
  - x4 regions & x2 domains
- Finnish
- French (x2)
- German (x3)
- Greek
- Hebrew
- Hungarian
- Icelandic
- Italian
- Japanese
- Korean
- Norwegian (x2)
- Polish
- Portuguese (x2)
- Russian
- Spanish
- Swedish
- Tamil
- Thai
- Turkish

Lotus

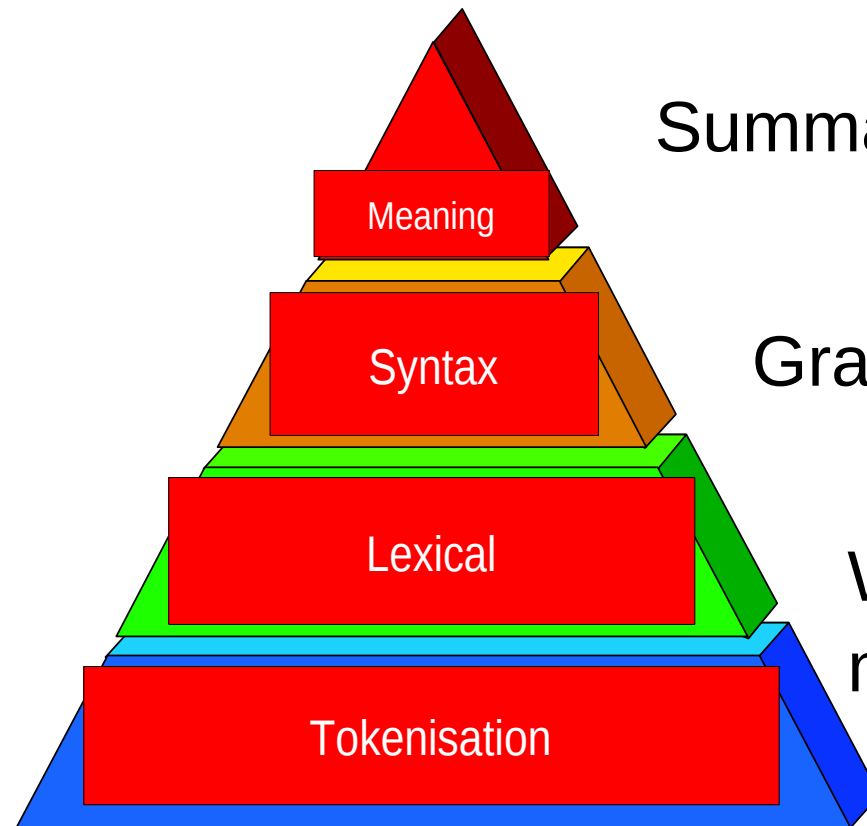
## Where are we?

- **Developers**
  - North Carolina, USA
  - Dublin, Ireland
  - Helsinki, Finland
  - Taipei, Taiwan
  - Yamato, Japan
- **Major Customers**
  - Various Locations, USA
  - Böblingen, Germany



Lotus®

# Analysis Hierarchy



Summarization

Grammar check

Words = spell check,  
morphology, etc.

IBM



Lotus

## Where/ How is it Used?

- Word Processors
  - spell aid, grammar checker
- Search Engines & Data Mining
  - Extracting Lemmatized form
  - Disambiguating terms
  - Assisting in Taxonomy Generation
- Translation Memory
  - Identification of linguistic units
- Machine Translation (not currently)
- Speech recognition (not currently)



Lotus®

## Challenges

- **Ambiguity**
  - **Words with multiple part of speech**
  - **Ambiguous word breaks**
- **Complex Orthographic rules**
  - **Casing rules**
  - **vowel dropping**



Lotus

## Ambiguity - POS

- Words have different meanings
  - shed (noun) = synonym of outbuilding
  - shed (verb) = synonym of molt
  - Meaning can be derived from context
    - It is in the shed
    - Trees shed their leaves
- English is one of the worst languages for POS ambiguity. Many other languages (e.g. Russian) inflect a word to indicate its syntactic role.



Lotus

## Ambiguity - Word Break

- Word break is not always obvious
  - French "pommes de terre"
    - as 3 words = "apples of the ground"
    - as 1 word = "potatoes"
  - English "red tape"
    - as 2 words = tape with red colour
    - as 1 word = synonym of "beuraucracy"



Lotus

## Languages with no spaces

- Some languages (e.g. Chinese, Japanese, Thai) do not put spaces between words
- It is difficult to figure out where the word breaks are
  - Sometimes there are multiple possible word segmentations.
  - Sometimes the choice of segmentation can change the meaning



Lotus

## Chinese Example

此路不通行不得在此小便

此 = this

路 = road

不 = no/not

通 = through

此路不通 = cul-de-sac

行 = walk

此路不通行 = cul-de-sac  
for walkers

不 = no/not

得 = allowed

不得 = forbidden

在此 = here

小便 = urinate

### ☀ Intended Interp

• 此路不通行

• 不得在此小便

### ☀ Other Interp

• 此路不通

• 行不得

• 在此小便



Lotus

## Decompounding

- German and related languages allow speakers to generate their own compound words by combining nouns and/or adjectives
- We need to break it into its constituent words and look these up in the dictionary
- Can be computationally expensive
- We can also sometimes get ambiguous results



Lotus

## Ambiguous sample 1

- Wachstube
  - Option 1 = wax tube
    - Wachs = wax
    - Tube = tube
  - Option 2 = awake room
    - Wach = awake
    - Stube = room
- Option 3
  - Wach(e) = Guard
  - Stube = room



Lotus

## Ambiguous sample 2

- Hochschulehrer = University Lecturer
- 3 component words
  - Hoch = High
  - Schule = school
  - Lehrer = teacher
- Could index under
  - Hochschule = High School (Univesity) & Lehrer = Teacher
  - Ober = Higher & Shulehrer = school teacher



Lotus

## Orthographic Variation

- A word may appear differently in a sample text.  
All languages have different rules
- English casing rules
  - a lowercase dictionary word can appear in text as titlecase or uppercase
  - a titlecase dictionary word could appear in text as uppercase
  - an uppercase or mixed case dictionary word must appear in text exactly the same.



Lotus

## Other languages have more complex rules

- In France letters lose their accent when capitalised
  - e.g. é is capitalised as E not É
  - but this rule does not apply in Fench Canada
  - So être becomes Etre in Paris but Être in Montreal
- In German capitalisation can change the letter count
  - sharp ß is sometimes capitalised as SS
- In English we optionally drop accents
  - The name Zoë can be written Zoe
  - In German umlauts are optionally represented by e so Böblingen can be written Boeblingen

IBM



Lotus

## Vowel dropping in semitic languages

- Semitic languages (Arabic and Hebrew), have an orthographic rule that vowels may be dropped.
- Imagine that Arabic has the words hit, hut, hat
- They will be written in dictionary as هِت, هُت, هَـت
- In any piece of text the string هـ could be any of these words



Lotus

## Arabic example in Large Font

- **hat** = هَات
- **hut** = هُت
- **hit** = هِيت
- **ht** = هَف



IBM



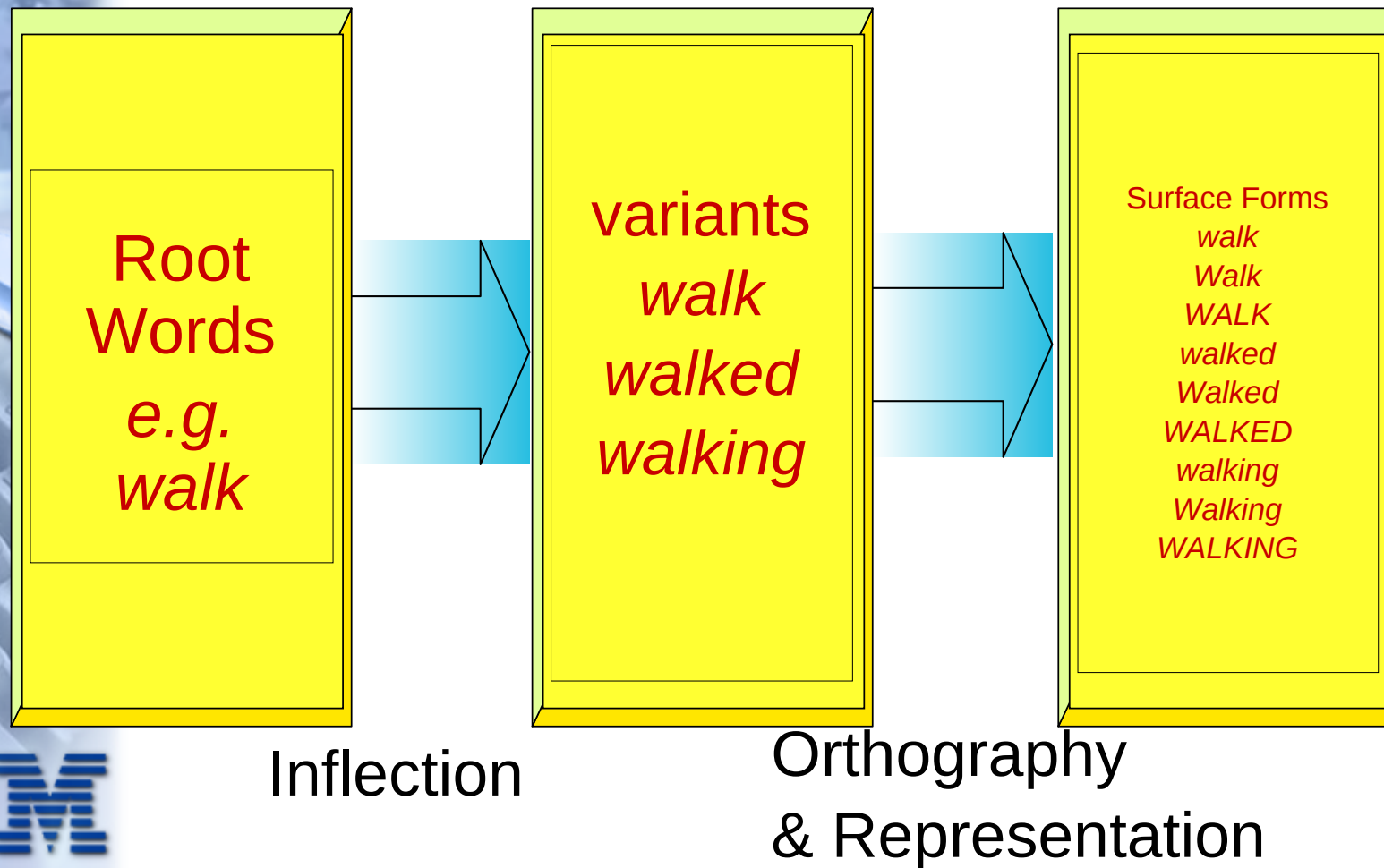
Lotus

## Representation Issues

- Our latest version is based entirely on utf-16, but representation issues still arise
- The letter ë is represented as 0x00EB
  - But could be e=0x0065 followed by umlaut=0x0308
- Arabic letter Heh (ه) = E5 in Windows code page is UNICODE 0x0645
  - But it has different forms e.g. هه
  - Unicode defines 4 code points for presentation forms
    - isolated=0xFEE9, final=0xFEEA, initial=0xFEEB and medial=0xFEEC

Lotus

# Model of How Orthographic variation handled



IBM



Lotus

## Future Challenges

- Improving quality and breadth of linguistic data
  - more languages
  - more words
  - richer relationships (e.g. part of, type of etc.)
- Increasing Accuracy of analysis
- Increasing performance speed
  - Latest version exceeds 2.5M Char/second
    - can scan 9GBytes in one hour?
  - allows .4 micro seconds/char

This document was created with Win2PDF available at <http://www.daneprairie.com>.  
The unregistered version of Win2PDF is for evaluation or non-commercial use only.