

---

# Localisation of Computer Software

***Alberto Escudero Pascual <aep@it46.se>***

***2005/11/15 Version 1.0***

---

# Table of Contents

- **Language Facts**
- What is localization and internationalization?
- Why is localization needed?
- The importance of FOSS
- **Locale**
- **Glossary**

---

# Table of Contents

- Writing systems
- Character encodings
- Fonts
- Input methods and keyboards
- Spell checkers
- **Localization Projects**

---

# Planet Earth



- Every two weeks, one language disappears
- With the languages also disappears a piece of our history!
- Without that history, we loose global and local knowledge!

---

# Planet Earth



- English is NOT spoken by  $2/3$  of the world's population
- Teaching people English is more difficult than teaching computers other languages

---

# Language Facts

- 6500 living languages in the world today.
- > 50% < 10 000 people.
- > 25% < 1 000 people
- Approx. 10% < 100 people

---

# Language Disparity Facts

- 5% of the world's languages are spoken by at least 1 million people and account for 94% of the world's population.
- The remaining 95% are spoken by only 6% of the world's people

---

# Future Facts?

The pessimists believe:

“In 100 years' time 90% of the world's languages will be gone, and that a couple of centuries from now the world may be left with only 200 tongues.”

---

# Most Widely Spoken Languages

Language	Number of native speakers
1. Chinese, Mandarin	885 million
2. Hindi	370 million
3. Spanish	350 million
4. English	340 million
5. Arabic (all forms)	206 million
6. Portuguese	203 million
7. Bengali	196 million
8. Russian	145 million
9. Japanese	122 million
10. Punjabi (east/west)	104 million

---

**Source: Ethnologue**

15 November 2005

© Creative Commons Deed. Attribution-NonCommercial-ShareAlike 2.0

Introduction to Computer Localisation

IT +46

---

I-18-n  
L-10-n

# Internationalization Localization

---

# What is Internationalization?

“Internationalization is the process of **generalizing** a product so that it **can handle multiple languages** and cultural conventions without the need for redesign. Internationalization takes place at the level of program design and document development.”

Source: The Localization Industry Standards Association

---

# What is localization?

Another process...

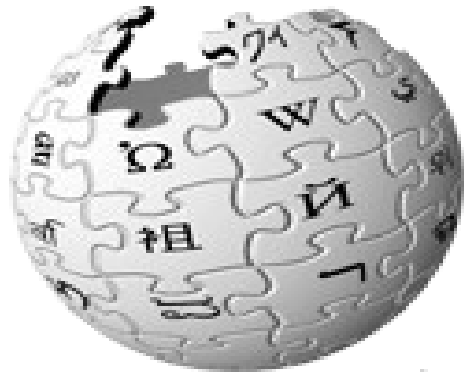
“Localization involves taking a product and **making it linguistically and culturally appropriate** to the target locale (country/region and language) where it will be used and sold.”

Source: The Localization Industry Standards Association

# Cultural Localization



# Graphical Localization



**WIKIPEDIA**  
*Cbū-iū ē Pek-kbo-choân-su*



**ویکی پدیا**  
دایرةالمعارف آزاد



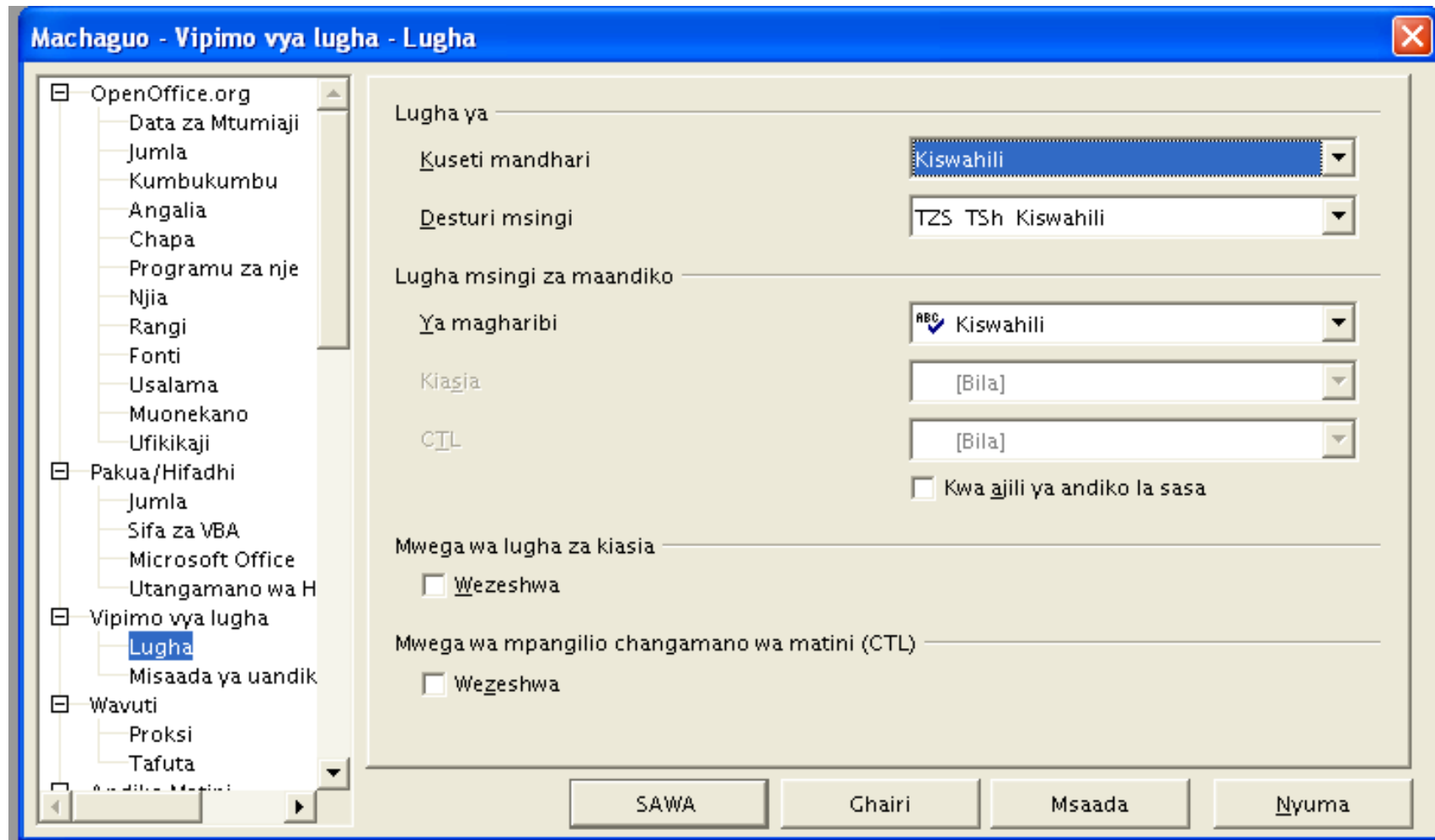
**ΒΙΚΙΠΑΪΔΕΙΑ**  
Η ελεύθερη εγκυκλοπαίδεια

---

# The Locale

- Representation of dates and times
- Abbreviations of days of the week, months
- Order of first day of the week
- Numerical notation

# The Locale



# The Locale



More than two simple  
time zones

AM

PM

---

# Glossary

- Find the ICT equivalent terms
- Requires Computer and Linguistic expertise
- Important element before starting localization
- Avoid start to translate without a glossary
- Consistency/QA
- Tools are needed!

---

# Creation of new terms

- Loanwords
  - Internet, CD-ROM (Swedish)
  - Chatta, maila (Swedish)
- Transliteration
  - Mouse » mus, **ratón**, **Maus**
- Semantic expansion
  - microorganism or malicious code?
- Metaphors

# Methaphors



---

# Writing Systems

## 1. **Logographic** (Chinese)

- One single symbol for an entire word

## 2. Syllabic (Cherokee, Katakana (jp))

- A set of written symbols represent syllables, which make up words

## 3. **Segmental scripts** (Arabic, Latin-based languages)

- A set of symbols represent the phonemes (basic unit of sound) of a language.

## 4. Featural (Hangul (kr))

- Each part of a symbol corresponds to a phonetic feature

---

# Logographic writing system

## Sample text in Chinese

繁體中文字 (Traditional Chinese characters)

人人生而自由，在尊嚴和權利上一律平等。他們賦有理性和良心，並應以兄弟關係的精神互相對待。

简体中文字 (Simplified Chinese characters)

人人生而自由，在尊嚴和權利上一律平等。他們賦有理性和良心，並應以兄弟關係的精神互相對待。

## Hànyǔ pīnyīn transliteration

Rénrén shēng ér zìyóu, zài zūnyàn hé quánlì shàng yìlǜ píngděng. Tāmen fùyǒu lǐxìng hé liángxīn, bìng yīng yǐ xiōngdì guānxì de jīngshén hùxiāng duìdài.

---

# Segmental Writing System

- **3.1 “True” Alphabet** (Latin, Greek, Mongolian)
  - A small set of separate letters (not diacritic marks) that represents both consonants and vowels.
- **3.2 Abjads** (Arabic, Hebrew)
  - One symbol per consonant
  - Vowels are usually not marked
- **3.3 Abugidas** (Ethiopic, Devanāgarī )
  - signs consists of consonants with an inherent vowel
  - modifications of the basic sign indicate other following vowels than the inherent one.

# Abugidas, Devanāgarī

## Vowels and vowel diacritics

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः	अँ	ऋ
a	ā	i	ī	u	ū	e	ai	o	au	aṅ	aḥ	ām	ṛ
[ə]	[a]	[i]	[i:]	[u]	[u:]	[e]	[æ:]	[o]	[ɔ:]	[aŋ]	[əh]	[ã:]	[r]
प	पा	पि	पी	पु	पू	पे	पै	पो	पौ	पं	पः	पाँ	पृ
pa	pā	pi	pī	pu	pū	pe	pai	po	pau	paṅ	paḥ	pām	pr

## Consonants

क	ka	[kə]	ख	kha	[kʰə]	ग	ga	[gə]	घ	gha	[gʰə]	ङ	ṅa	[ŋə]
च	ca	[tʃə]	छ	cha	[tʃʰə]	ज	ja	[dʒə]	झ	jha	[dʒʰə]	ञ	ña	[ɟə]
ट	ṭa	[tʰə]	ठ	ṭha	[tʰʰə]	ड	ḍa	[dʰə]	ढ	ḍha	[dʰʰə]	ण	ṇa	[ɳə]
त	ta	[tə]	थ	tha	[tʰə]	द	da	[də]	ध	dha	[dʰə]	न	na	[nə]
प	pa	[pə]	फ	pha	[pʰə]	ब	ba	[bə]	भ	bha	[bʰə]	म	ma	[mə]
य	ya	[jə]	र	ra	[rə]	ल	la	[lə]	व	va	[və]			
श	śa	[ʃə]	ष	ṣa	[ʃʰə]	स	sa	[sə]						
ह	ha	[ɦə]												

# Abugidas, Ethiopic Gee'z

ሀ	ለ	ሐ	መ	ሠ	ረ	ሰ	ሸ	ቀ	ቄ	በ	ተ	ቸ	ኀ	ኁ	ነ	ኘ	አ
h	l	h	m	s	r	s	š	q	qu	b	t	č	h	hu	n	ñ	'
[h]	[l]	[h]	[m]	[s]	[r]	[s]	[j]	[kʰ]	[kʷ]	[b]	[t]	[tʃ]	[h]	[hʷ]	[n]	[ɲ]	[ʔ]
ከ	ኸ	ወ	ዐ	ዘ	ዠ	የ	ደ	ጀ	ገ	ኀ	ጠ	ጪ	ጰ	ጸ	ፀ	ፊ	ፐ
k	h	w	'	z	ž	y	d	ǰ	g	gu	t	č	p	s	z	f	p
[k]	[h]	[w]	[ʔ]	[z]	[ʒ]	[j]	[d]	[dʒ]	[g]	[gʷ]	[t]	[tʃ]	[p]	[ts]	[ts]	[f]	[p]



# Character encodings UTF-8

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

**Latin scripts and symbols**

Other European scripts

Linguistic scripts

**Middle Eastern and SW Asian scripts**

**African scripts**

**South Asian scripts**

Southeast Asian scripts

East Asian scripts

Unified CJK Han

Ah original scripts

Symbols

Diacritics

UTF-16 surrogates and private use

Miscellaneous characters

Roadmap of Unicode Basic Multilingual Plane.

Each numbered box represents 256 codepoints.

---

# Fonts

“The **visual representation of characters** from a particular character set.”

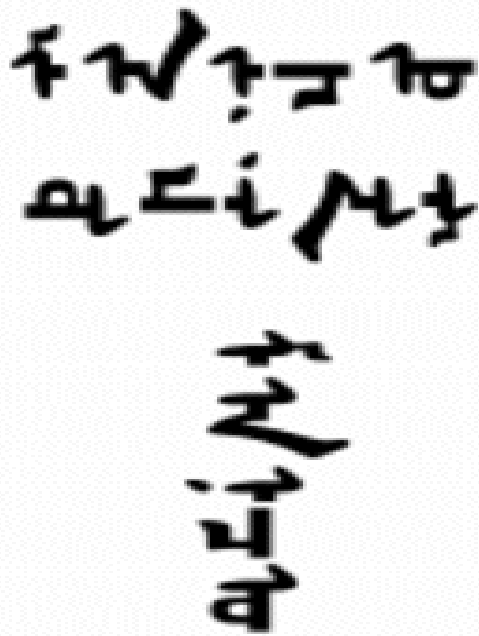
or

“A description of how to display a set of characters that includes the **shape** of the characters, **spacing** between characters, **type** of characters (bold, italics, underline) and the **size** of the characters.”

# Fonts

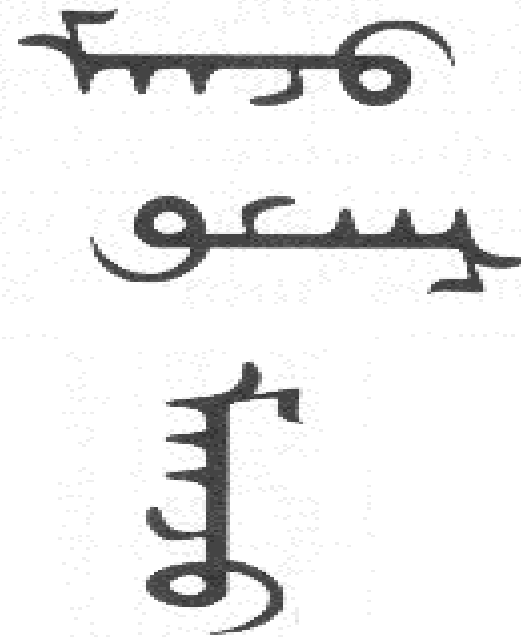


# Fonts and joining (ligatures)



The image shows three lines of Manchu script. The top line is written horizontally from left to right. The middle line is written horizontally from right to left. The bottom line is written vertically from top to bottom.

"Manchu" in Mongolian  
LTR, RTL, and Vertical  
182E, 1820, 1828, 1834, 1824  
(No ligatures or position variants)



The image shows three lines of Arabic script. The top line is written horizontally from left to right. The middle line is written horizontally from right to left. The bottom line is written vertically from top to bottom.

# Input Method

*Hãy thử nhìn nhiệm vụ của chúng ta, vì nó sẽ thành công!*

$= \sim + e$   
 $= ^ + . + e$   
 $= . + ^ + e$







# Bidi

כאשר העולם רוצה לדבר, הוא מדבר  
ב־Unicode. הירשמו כעת לכנס  
Unicode הבינלאומי העשירי,  
שייערך בין התאריכים 10-12 במרץ  
1997, ב־מִינְיָץ שבגרמניה. בכנס  
ישתתפו מומחים מכל ענפי התעשייה  
בנושא האינטרנט העולמי  
וה־Unicode, בהתאמה לשוק  
הבינלאומי והמקומי, ביישום  
Unicode במערכות הפעלה  
וביישומים, בגופנים, בפריסת טקסט  
ובמחשוב רב־לשוני.

# Keyboards



---

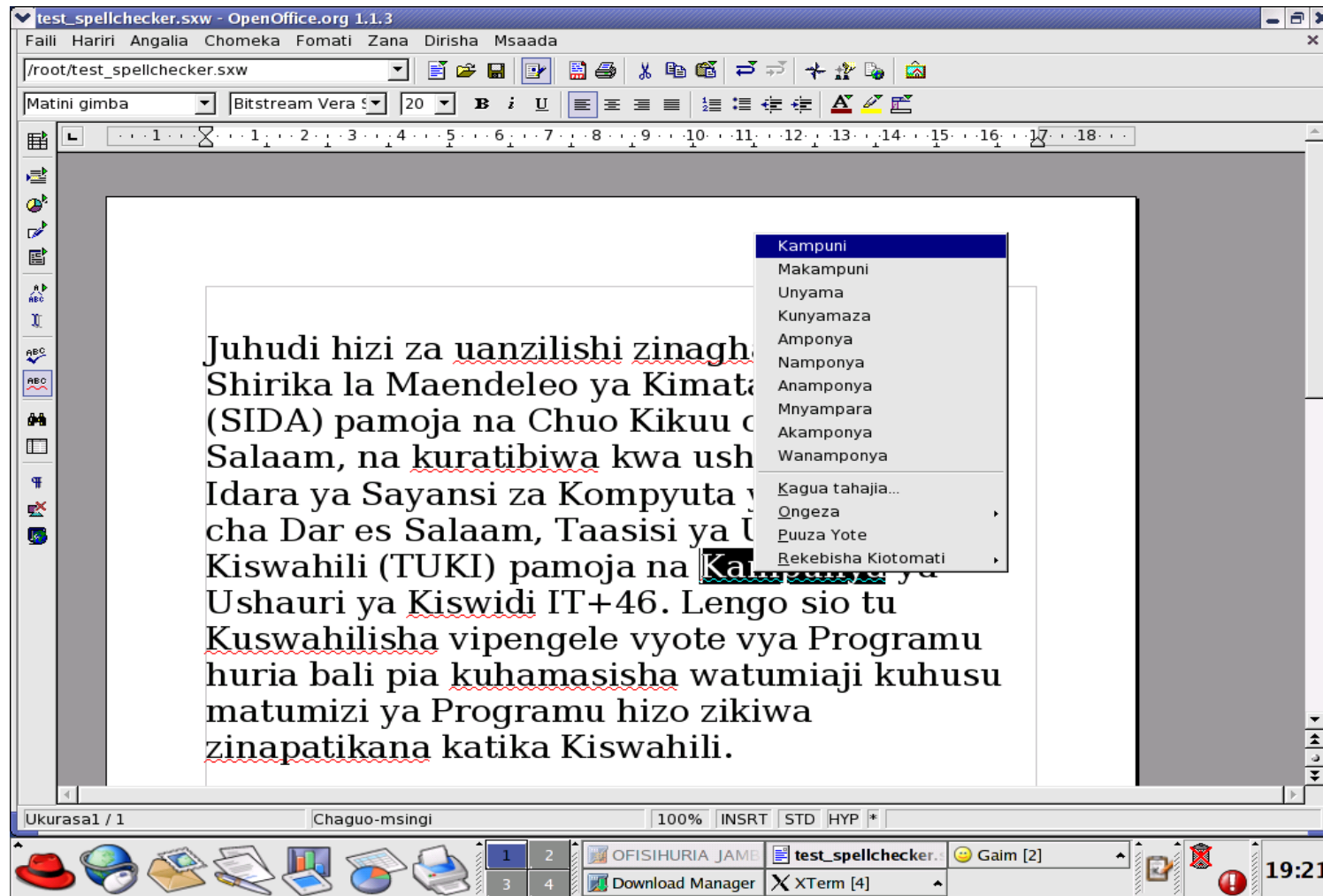
# Spellcheckers

- Most minority languages are using Scannell Word crawler
  - e.g. Kinyarwanda, Swahili dictionary for Myspell
- Compression is a difficult issue as implies strong linguistic expertise and “patience”
- Lack of <spaces > is major issue for some language
- Hunspell is designed to eventually replace Myspell in OpenOffice.org.

# Thai Spellchecking <spaces>

ณ ยามที่โลกต้องการเชื่อมต่อคำใดๆ โลกจะใช้เพียง Unicode เราจึงขอเชิญชวนท่านรับลงทะเบียนงาน International Unicode Conference ครั้งที่ 10 ซึ่งจะจัดให้มีขึ้น ณ เมือง Mainz ประเทศเยอรมัน ในระหว่างวันที่ 10-12 มีนาคม ค.ศ. 1997 เสียแต่บัดนี้ โดยในงานประชุมดังกล่าว ท่านจะมีโอกาสได้พบกับบรรดาผู้เชี่ยวชาญจากธุรกิจอินเทอร์เน็ตและ Unicode ธุรกิจ Internationalization และ Localization จากทุกมุมทั่วโลก พร้อมรับทราบการใช้ประโยชน์จาก Unicode ร่วมกับระบบปฏิบัติการและโปรแกรมต่างๆ ฟอนต์ รูปแบบข้อความ รวมทั้งวิทยาการด้านคอมพิวเตอร์ในภาษาต่างๆ

# Vantu Spellchecking



---

# Collation

- Collation is the general term for the process and function of determining the sorting order of strings of characters.

a < b < ... < z < å < ä < ö (se\_SE)

a < b < c < ch < ... < n < ñ < ... < z (es\_ES)

Source: <http://www.unicode.org/reports/tr10/tr10-10.html>

---

# Collation

- Phonetic sorting of Han characters requires:
  - lookup dictionary of words
  - databases to maintain an associated phonetic spelling for the words in the text.
  
- For example, the text string "A " (A \u0e02\u0e40) is processed internally in collation as "A " (A \u0e40\u0e02).

Good collation charts:

<http://developer.mimer.com/collations/charts/index.tml>

---

# Localization of strings

- After including all the other parts... finally the translation (a knowledge area by itself)
  - Extract strings from source
  - Translate the strings
  - Merge back and Build the software
  - Review QA
  - Translate the strings
  - Build the software (release)

---

# Planning Localization

35% Management

25% New tools development

25% Translation itself (typing), QA

15% Education and Dissemination

# Lots of work to do!

