



NUANCE

The experience speaks for itself™

Corpus Linguistics and Culturally Relevant Localization for the Global Mobile Market

David Rojas, LRC XII – Localisation Research Forum, 2007-09-28



Agenda

- T9 and XT9 from 30,000ft / 9144m
- Historical and geo-linguistic situation of our software
- Corpus linguistics principles and procedures
- Input methods

T9 in a Nutshell

- Disambiguated key presses = words
- “One key per letter”
- With T9
– Your ⁵phone is ⁷smarter than you ⁵think !
- Without T9
– Is your ¹⁰phone ¹⁵smarter than you ¹⁰think ?

T9 and XT9 Highlights

- Some features of T9
 - Word completion
 - Word prediction
 - Simultaneous bilingual mode
 - User personalization
- Some features of XT9
 - Regional disambiguation (i.e. “Sloppy Type”)
 - Spell correction
 - Auto substitution
 - thxab → thanks a bunch!
 - nest → n’est



Ten Years of T9 Text Input

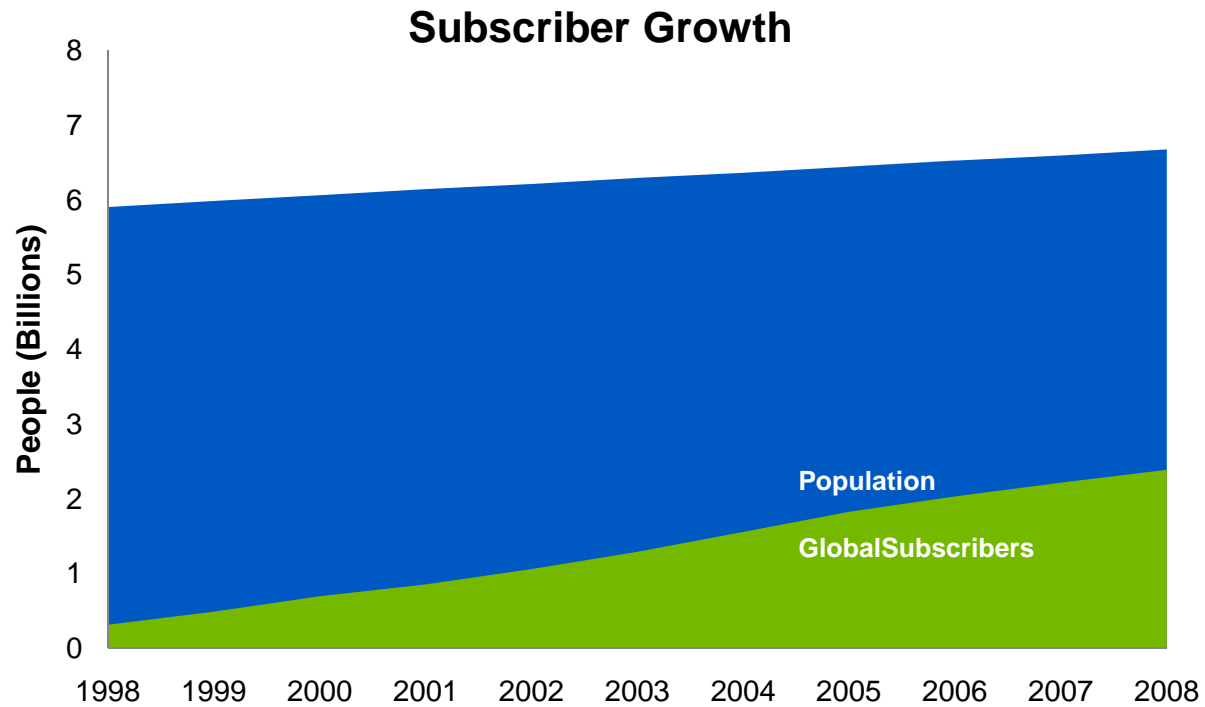
- T9 English first released: fall of 1997
 - Texas Instruments Avigo PDA
- T9 Text Input in 66 different languages
- Our T9 and XT9 languages are written in 19 scripts



David Rojas, LRC XII, 2007-09-28
© 2007 Tegic Communications



Mobile Market Worldwide



(Sources: U.S. Census Bureau, Gartner)

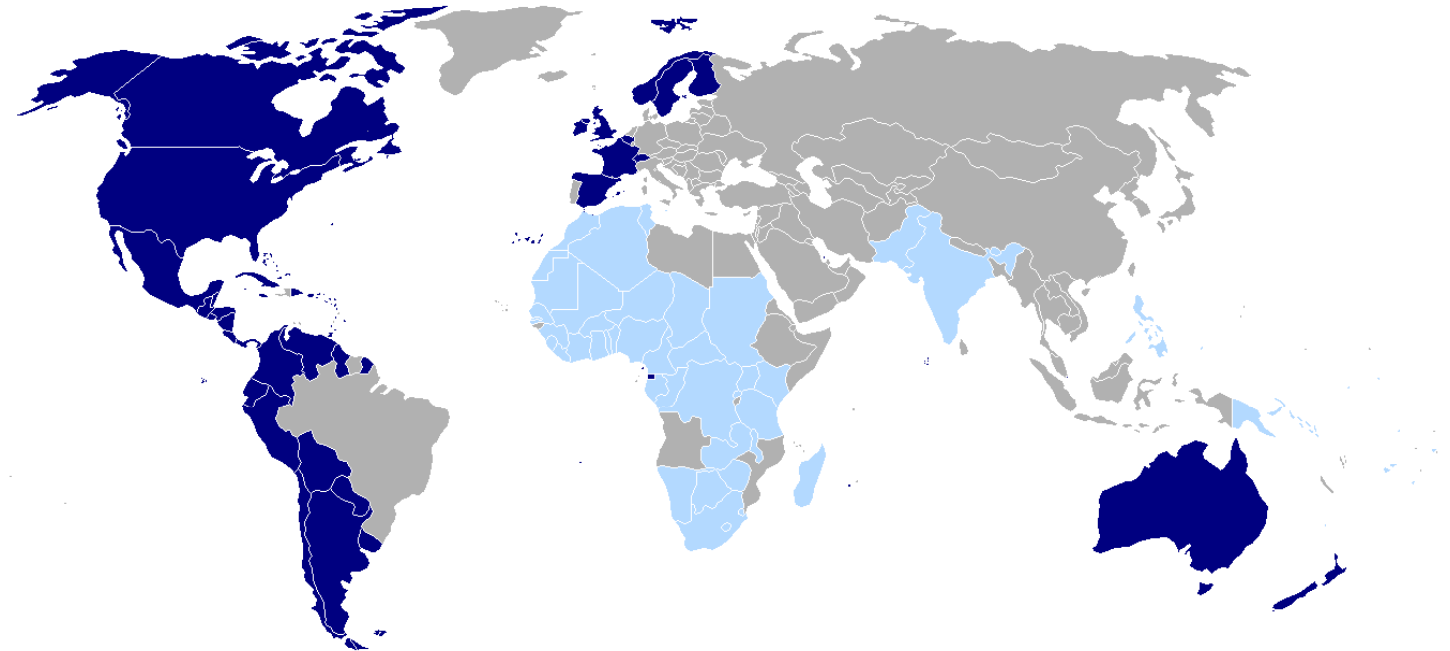


David Rojas, LRC XII, 2007-09-28
© 2007 Tegic Communications



T9 Chronology I

- English
- Spanish
- French
- Finnish
- Norwegian
- Swedish

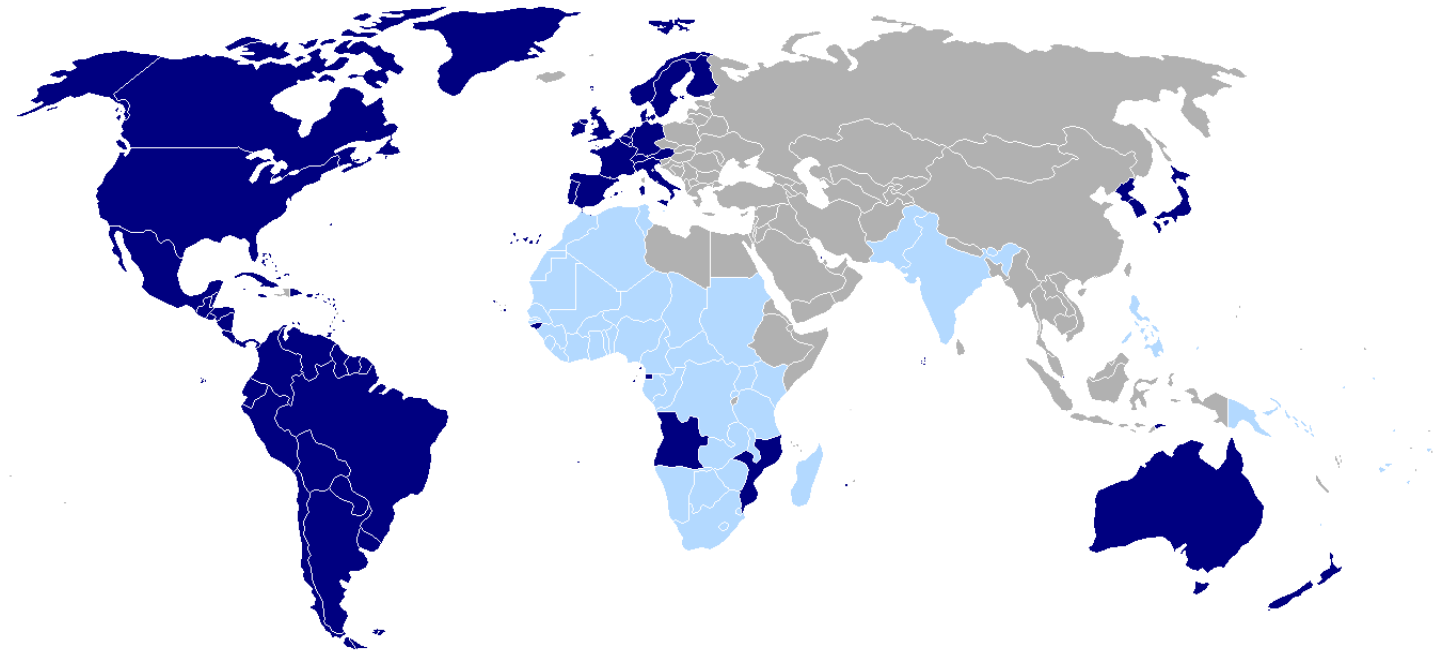


David Rojas, LRC XII, 2007-09-28
© 2007 Tegic Communications



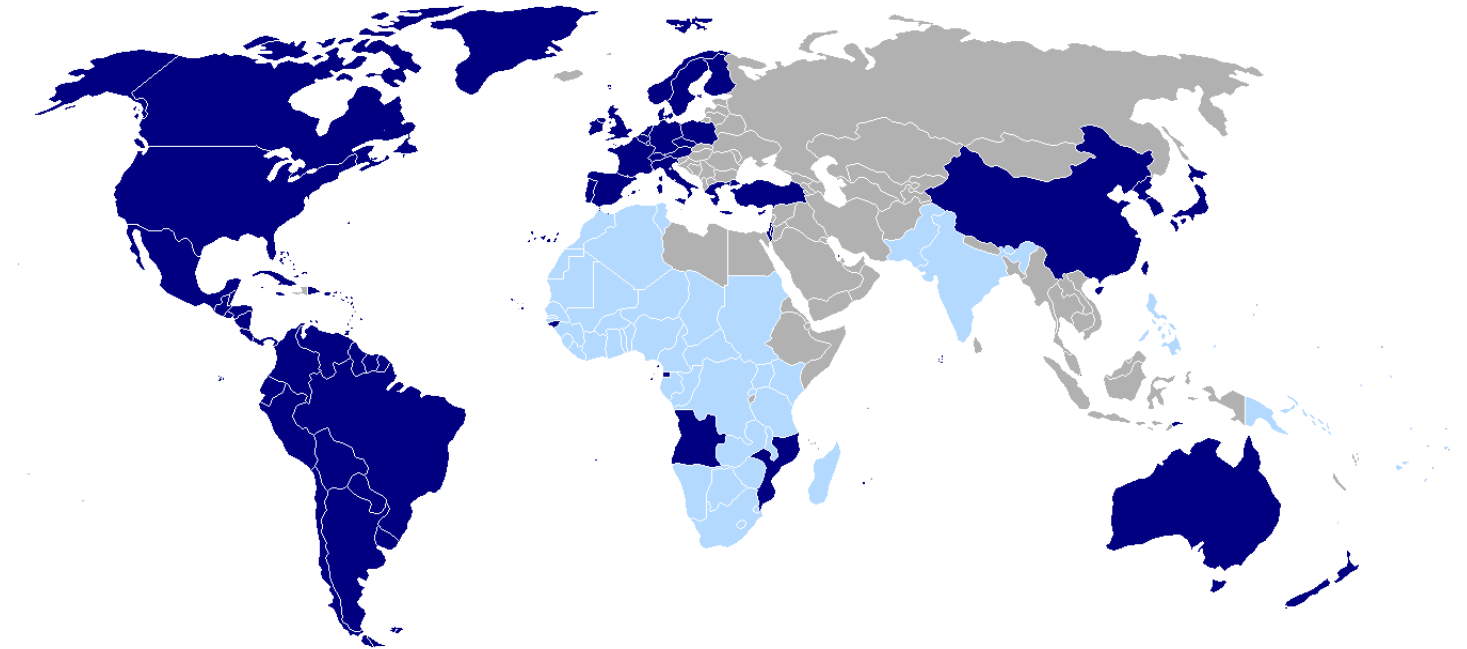
T9 Chronology II

- Danish
- German
- Italian
- Portuguese
- Dutch
- Korean
- Japanese



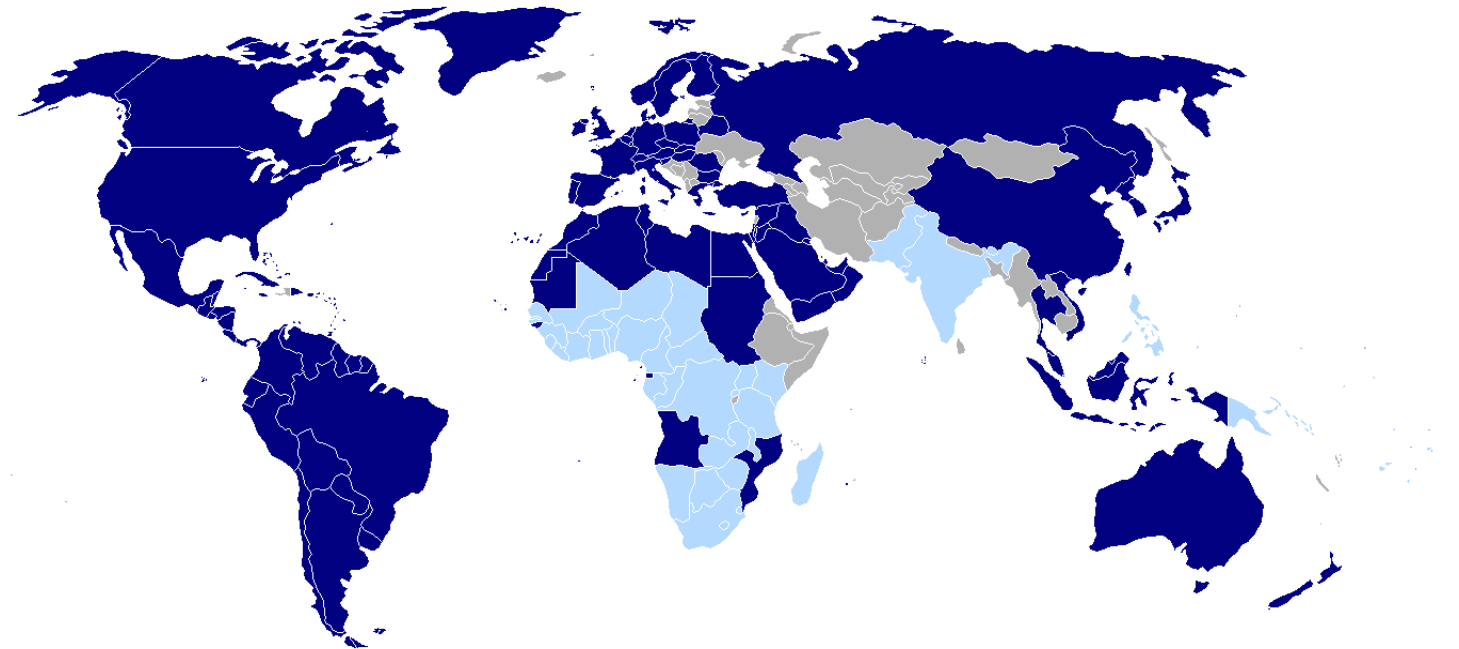
T9 Chronology III

- Chinese
- Greek
- Turkish
- Czech
- Polish
- Hebrew



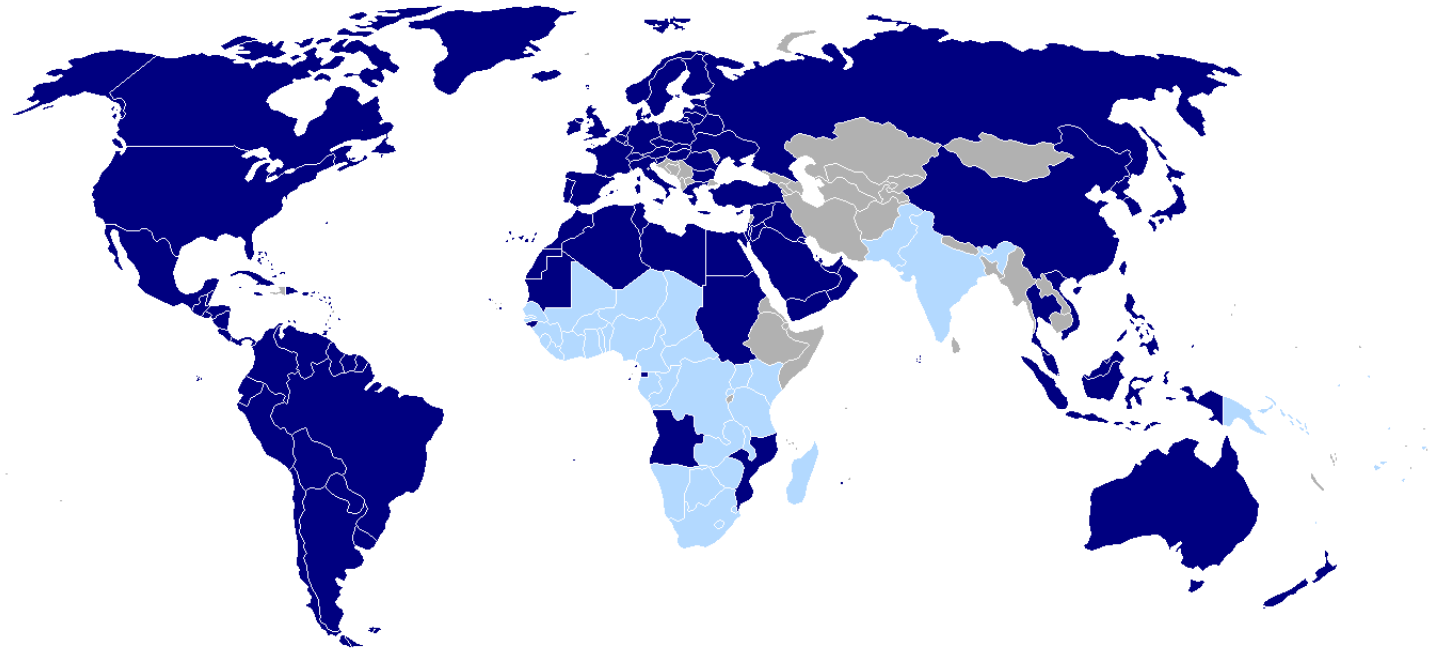
T9 Chronology IV

- Arabic
- Hungarian
- Slovak
- Russian
- Bulgarian
- Romanian
- Slovenian
- Malay
- Indonesian
- Thai
- Vietnamese



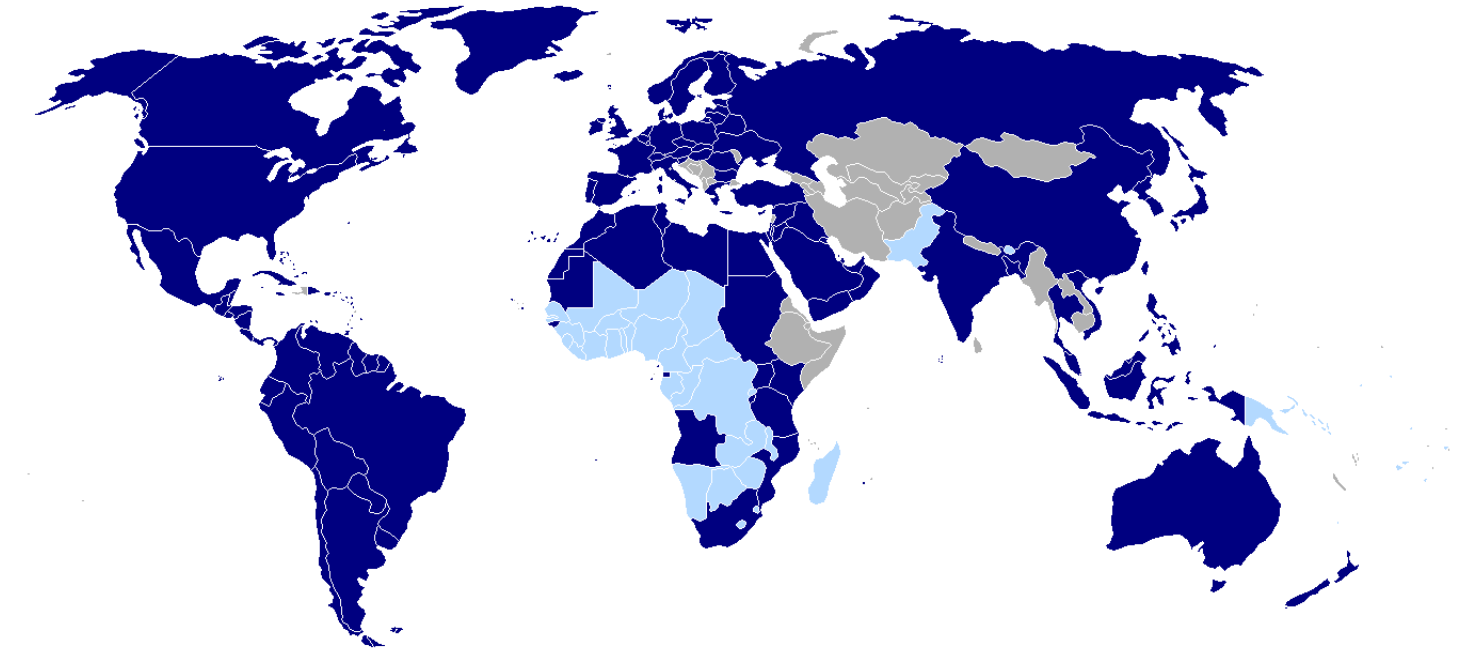
T9 Chronology V

- Lithuanian
- Latvian
- Estonian
- Ukrainian
- Catalan
- Icelandic
- Tagalog



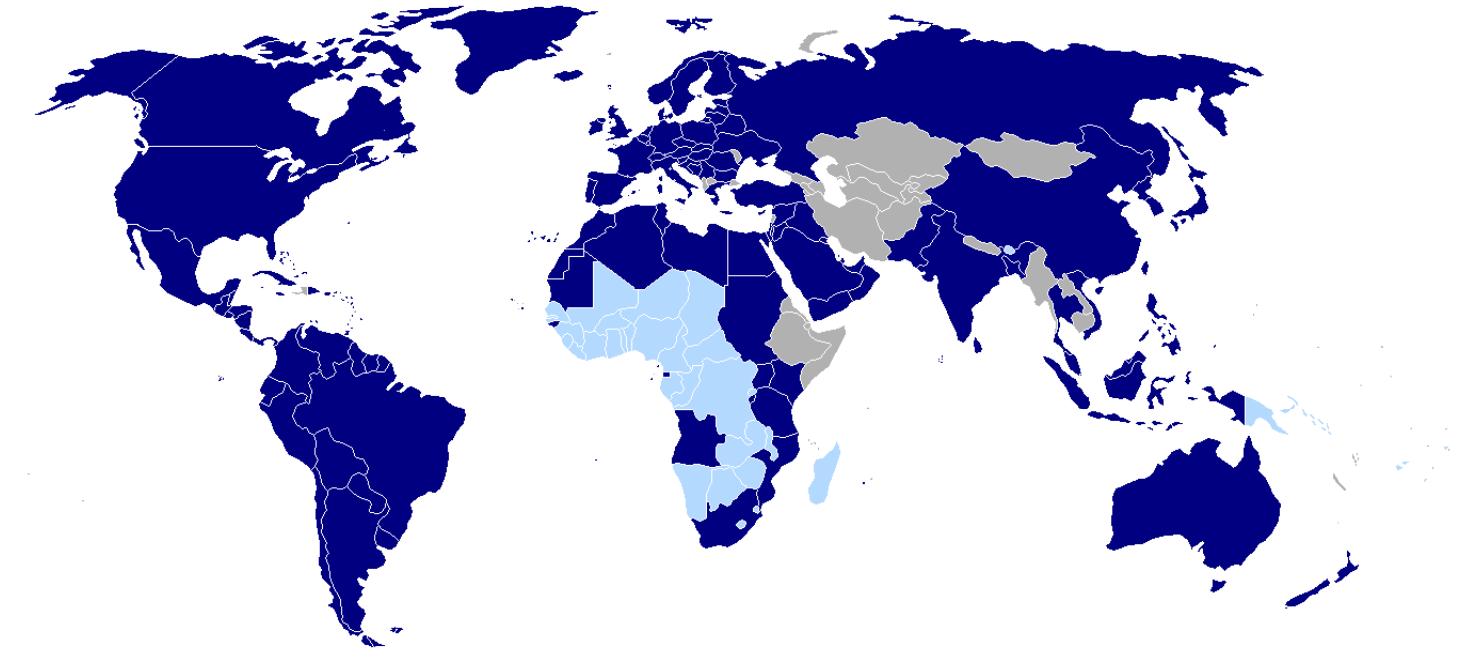
T9 Chronology VI

- Hindi
- Bengali
- Afrikaans
- Swahili



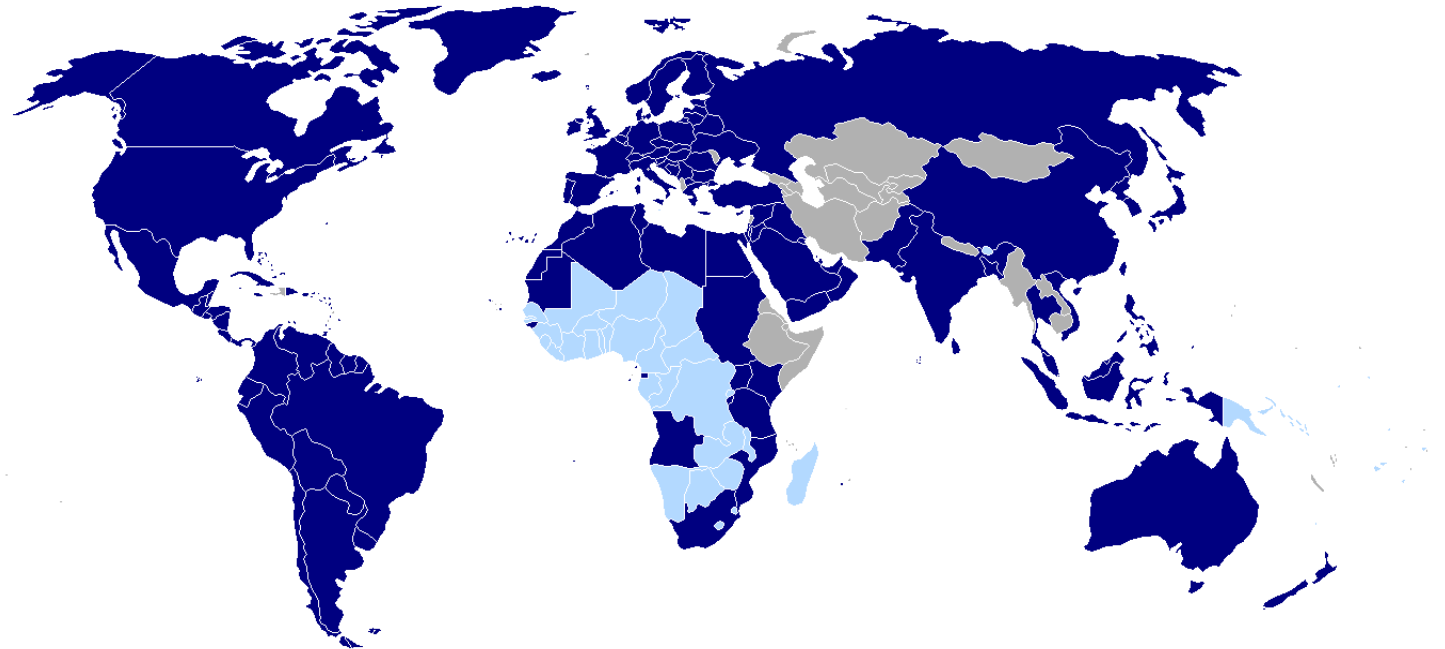
T9 Chronology VII

- Tamil
- Urdu
- Punjabi
- Croatian
- Serbian



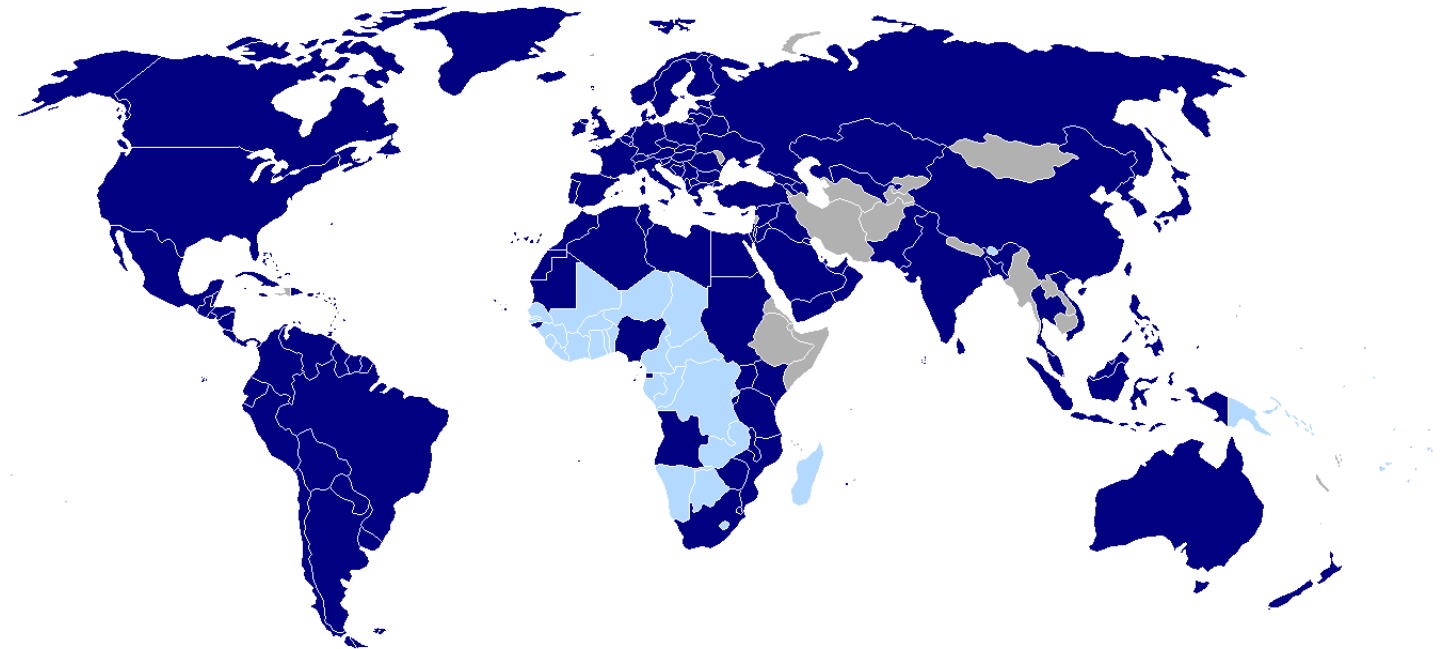
T9 Chronology VIII

- Basque
- Galician
- Macedonian
- Marathi
- Gujarati



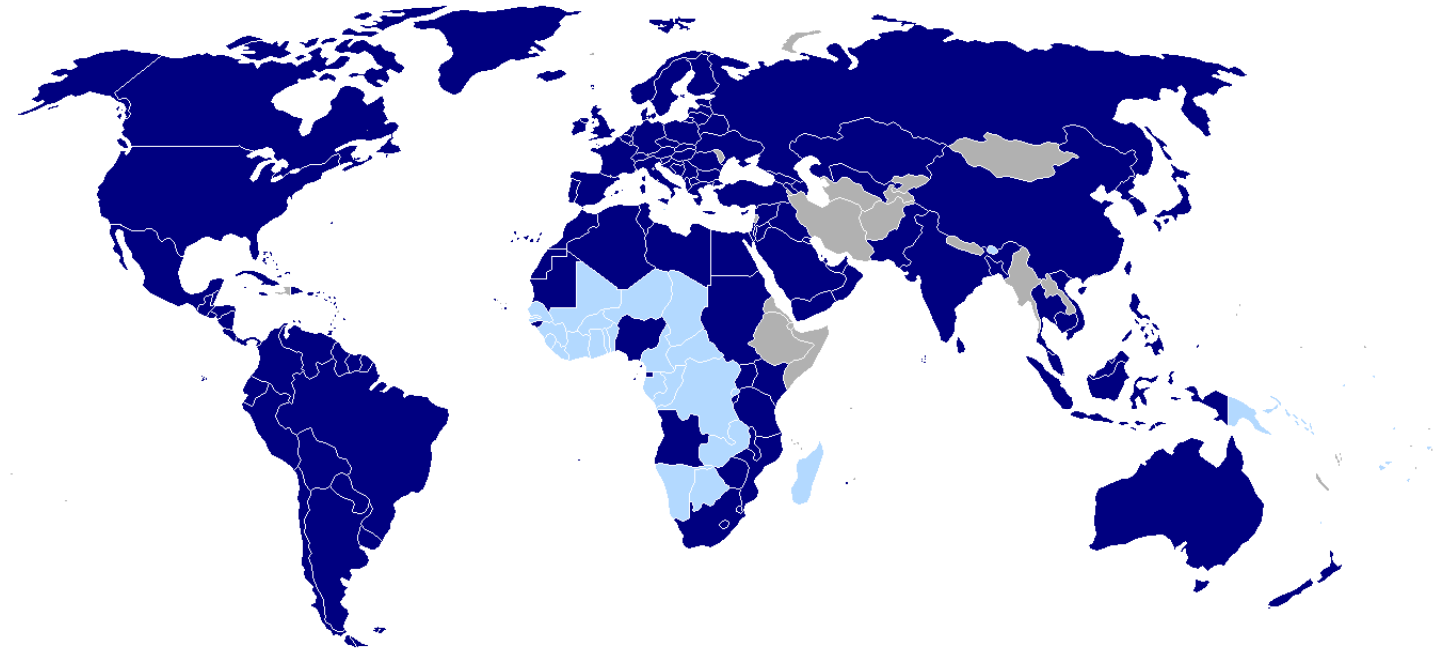
T9 Chronology IX

- Kannada
- Kazakh
- Azerbaijani
- Uzbek
- Georgian
- Albanian
- Zulu
- Xhosa
- Hausa
- Yoruba
- Igbo



T9 Chronology X

- Welsh
- Sesotho
- Khmer
- Telugu



David Rojas, LRC XII, 2007-09-28
© 2007 Tegic Communications



Bilingualism and Text Input

- T9 bilingual mode
- Not everyone is monolingual
- Spanglish
 - Large market in the US
 - Code switching is common
- Hinglish
 - Transliterated Hindi
 - English lexical content

Corpus Linguistics

- Localization
 - Culturally and linguistically appropriate translation of strings
 - Provision of contextually relevant concepts (not by translation)
- Patterns of language use
 - Zipfian distribution expected
- Challenges
 - Collection
 - Clean & Normalize
- Native speaker involvement along the way

Which Words Are Important?

Hojaldre con morcilla y manzana

1 lámina de hojaldre 1 huevo
3 morcillas de cebolla 5 almendras
3 manzanas

Enharinar la mesa de trabajo y extender la lámina de hojaldre con el rodillo.

Cubrir el fondo de una bandeja de horno con papel sulfurizado y colocar encima una capa de hojaldre.

Quitar la piel de las morcillas, chafarlas con la ayuda de un tenedor y poner por encima del hojaldre.

Pelar y triturar las manzanas y distribuir el puré por encima de la morcilla.

Tapar con otra capa de hojaldre, decorar con los recortes de hojaldre y con las almendras.

Batir el huevo y pincelar con él la coca.

Meter en el horno precalentado a 200º centígrados, aproximadamente durante 1 hora.

(<http://www.recetas.net/receta.asp?ID=6624SI>)

West seeks U.N. sanctions on Myanmar

UNITED NATIONS (Reuters) - The U.N. Security Council on Wednesday urged Myanmar to admit a top U.N. envoy immediately, but China immediately ruled out calls for sanctions or a U.N. condemnation of the ruling junta's use of force.

The United States and the 27-member European Union had asked the council to consider punitive measures and demanded that the junta in the former Burma open a dialogue with jailed opposition leader Aung San Suu Kyi and ethnic minorities.

"We believe that sanctions (are) not helpful for the situation down there," China's U.N. Ambassador Wang Guangya told reporters.

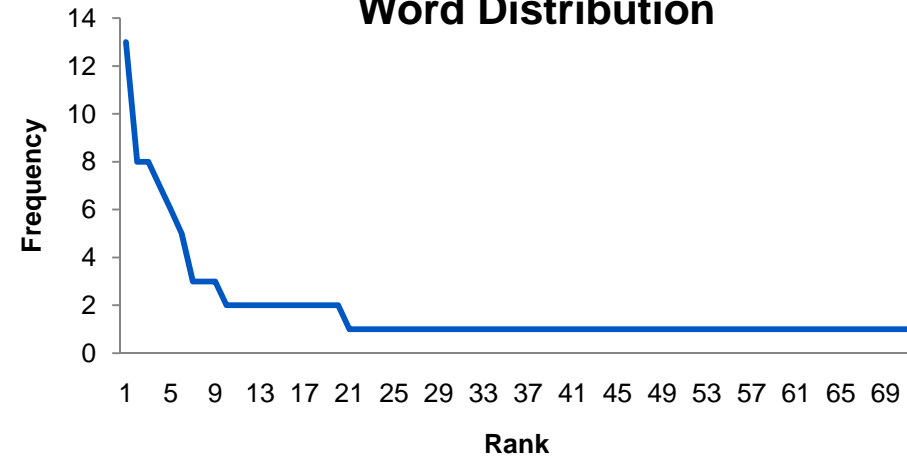
While expressing concern at events, Wang said the situation in Myanmar did not "constitute a threat to international peace and security," the main mandate of the Security Council and the reason China in the past has prevented council action.

(<http://www.reuters.com/article/worldNews/idUSN2621598020070926>)

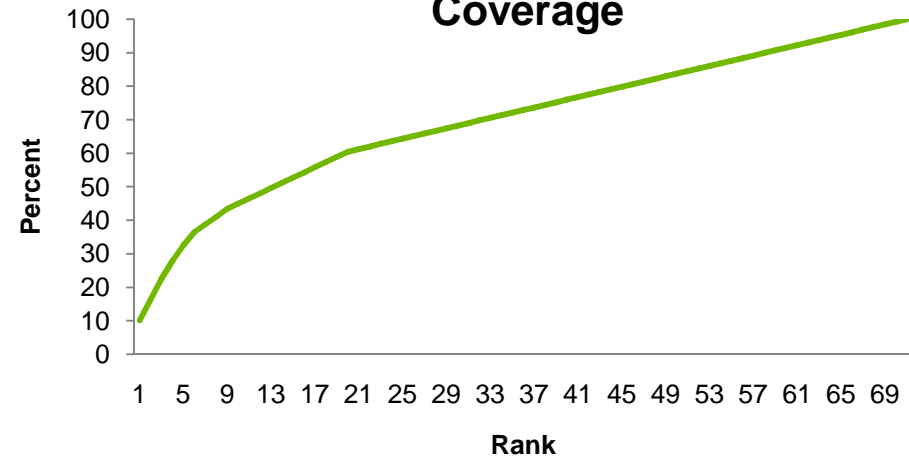
Spanish Example

13 de	2 horno	1 poner
8 y	2 capa	1 pincelar
8 con	2 almendras	1 piel
7 hojaldre	2 3	1 personas
6 la	1 él	...
5 el	1 un	1 extender
3 las	1 triturar	1 española
3 encima	1 trabajo	1 enharinar
3 1	1 tenedor	...
2 una	1 tapar	1 durante
2 por	1 sulfurizado	1 distribuir
2 morcillas	1 rodillo	1 del
2 morcilla	1 recortes	...
2 manzanas	1 quitar	1 4
2 lámina	1 puré	1 200°
2 huevo	1 precalentado	

Word Distribution



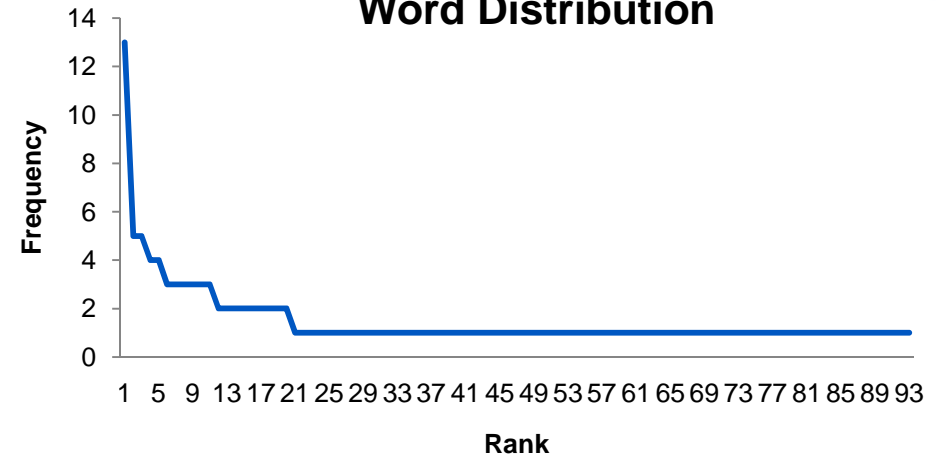
Coverage



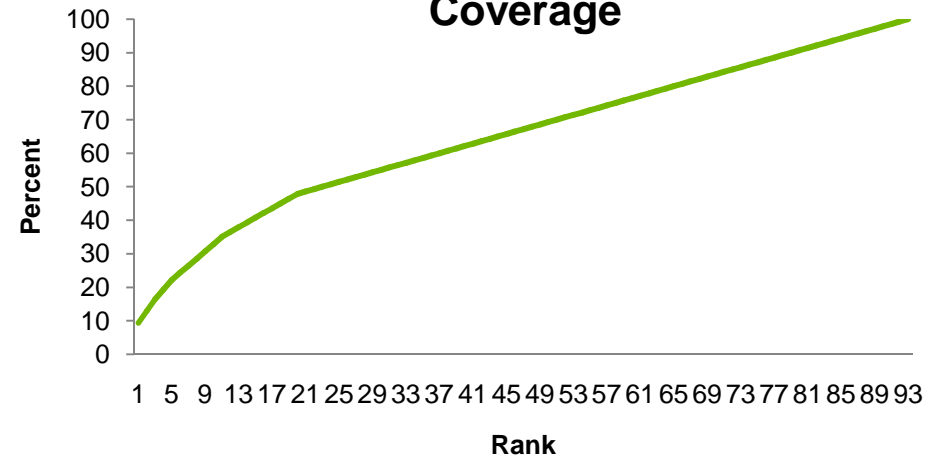
English Example

13 the	2 on	1 threat
5 u.n.	2 not	1 there
5 and	2 immediately	1 suu
4 council	2 for	1 states
4 a	2 china	1 seeks
3 to	1 with	1 san
3 security	1 while	1 said
3 sanctions	1 west	1 ruling
3 of	1 wednesday	1 ruled
3 myanmar	1 we	...
3 in	1 use	1 are
2 wang	1 urged	1 ambassador
2 united	1 union	1 admit
2 that	1 top	1 action
2 situation	1 told	1 27-member

Word Distribution



Coverage



Cultural Relevance

- Patterns of natural usage
 - Limited amount of content guaranteed
 - Certain semantic equivalents are expected across languages
- Humans in the mix
 - Usenet groups, blogs, etc.
 - Customer and consultant requested modifications

Language & Literacy Effects

- Assume literate users
- Compensate for poor spellers
 - T9: include common misspellings
 - XT9: spell-correction
- Philippines using T9 to promote literacy
 - “wl u b thr l8r”
 - Errors become acceptable by habit
 - “Text Right...T9 It!” campaign
- The *next* key and slang...totally book!

19 Scripts Supported

- Alphabetic
- Logographic
- Syllabic
- Abjads (+ bi-directional)
- Abugidas



Localising the Text Input Method

- Abugida input
 - Syllabic/Quasi-syllabic
 - Knowledge of alphabetic order & location of combining chars
- Non-alphabetic input
 - Chinese
 - Japanese
 - Korean (not exactly alphabetic, that is)




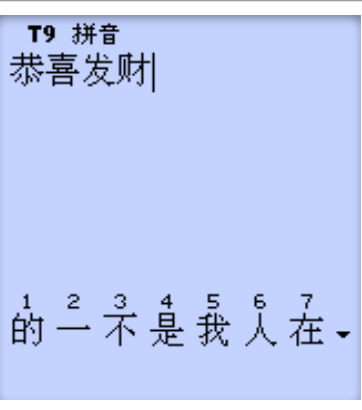
Chinese Solution: Phrasal Text Input

- Closer to native text input solution on PC
- Fewer keypresses, easy to learn
- Bilingual Chinese – English support
- Numbers, symbols, emoticons without changing mode
- Various input methods
 - Simplified Chinese: phrasal input and stroke
 - Traditional Chinese: phrasal BoPoMoFo, stroke, Pinyin
 - Automatically detect syllable boundary (delimiter is not required in Pinyin/BoPoMoFo input)



Chinese Input: Pinyin

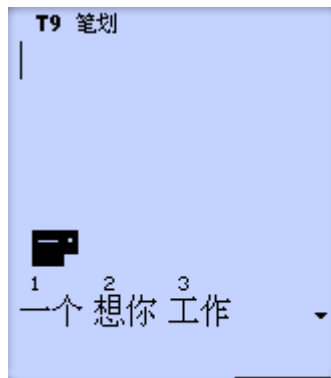
- Enter text using Pinyin
 - 恭喜发财(**English:** wishing you prosperity, **Pinyin:** GongXiFaCai)

			
Press 4664	Press 94	Press 32	Long press 1 to select phrase

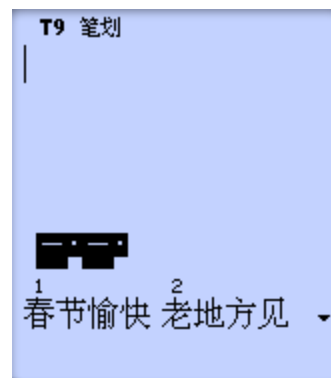
Chinese Input: Stroke

- Enter text using strokes by direction and correct order
 - 恭喜发财(**English:** wishing you prosperity, **Stroke:** 1221342444. 121251431251.53544. 2534123)

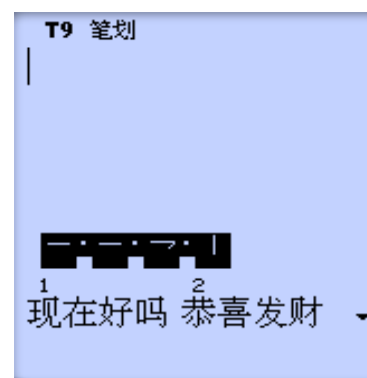
- Delimiter required



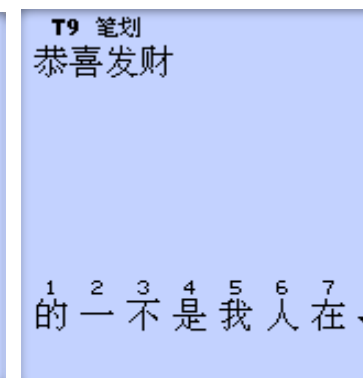
Press 1#



Press 1#



Press 5#2



Long press 2
to select phrase

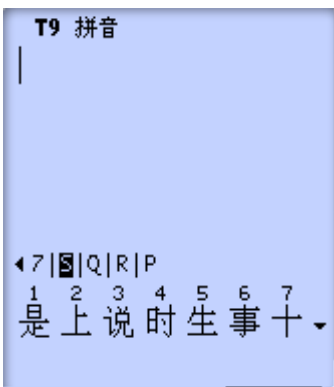
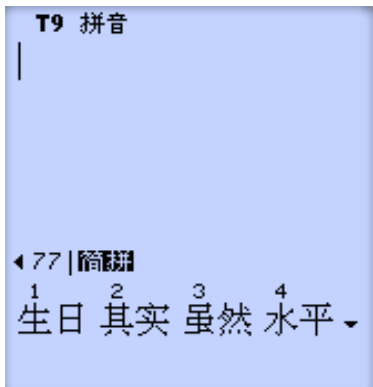
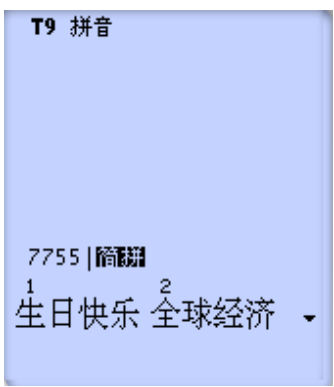
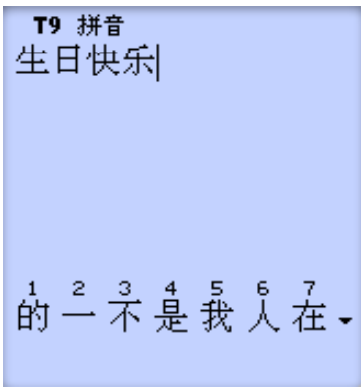
Chinese Input: BoPoMoFo

- Enter phrase using BoPoMoFo
 - 恭喜發財 (**English:** wishing you prosperity, **BoPoMoFo:** ㄍㄨㄥ ㄒㄧ ㄈㄞ ㄇㄞ ㄉㄨㄞ ㄉㄞ ㄉㄞ ㄉㄞ, 259820162)

<p>Press 259</p>	<p>Press 82></p>	<p>Press 0</p>	<p>Long press 1 to select phrase</p>

Chinese Input: Jianpin

- Enter phrase using Jianpin
 - 生日快乐(**English:** happy birthday, **Pinyin:** ShengRiKuaiLe, **Jianpin:** SRKL)
 - Enter first letter of each syllable to retrieve all phrases

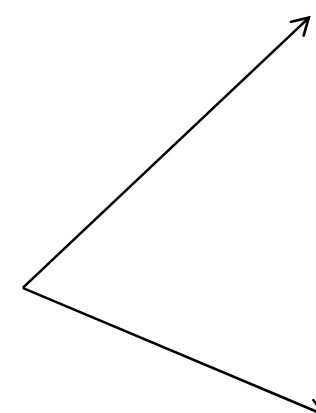
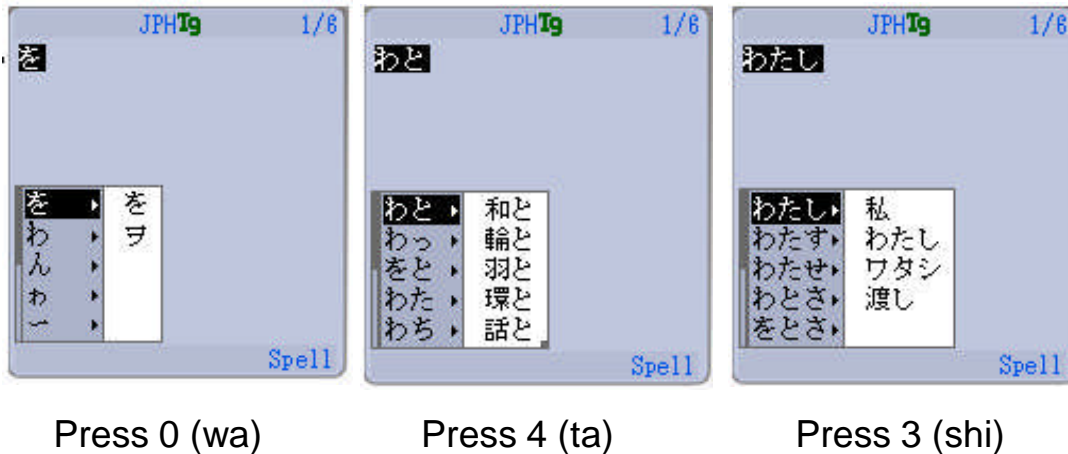
			
Press 7	Press 77	Press 55	Long press 1 to select phrase

Japanese Solution

- Multiple input methods
 - Initial text entry in either Hiragana or Katakana
 - Kanji conversion engine suggests complete words
- Word completion
- Word prediction
- Individual characters can be changed after selection

Japanese Solution: Example

- Enter the Japanese word わたし “watashi”
- 043ok
 - Updates screen with desired word
 - Displays possible next words



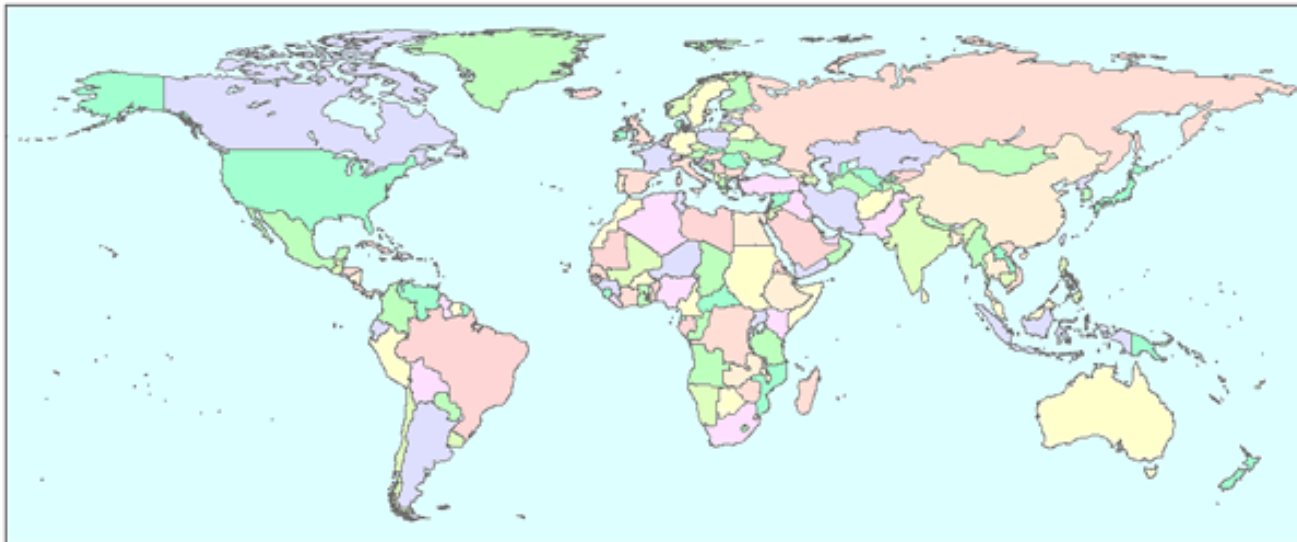
Select Kanji



Press OK

Mind the Gap!

- 194 independent nations
- Approximately 6800 living languages
- Approximately 6.6 billion people



Thank You

*Questions
And
Comments*

More information at <http://www.nuance.com/t9/>

david.rojas@nuance.com



David Rojas, LRC XII, 2007-09-28
© 2007 Tegic Communications

