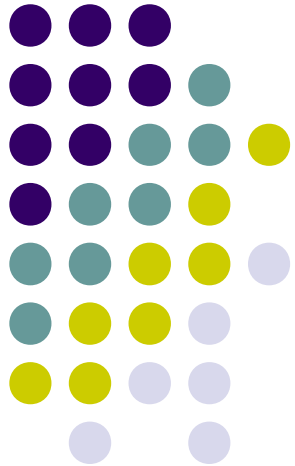
# Transcription, transliteration, transduction, and translation

## A typology of crosslinguistic name representation strategies

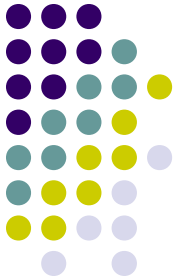


Deryle Lonsdale

BYU Linguistics

lonz@byu.edu

# The crossroads



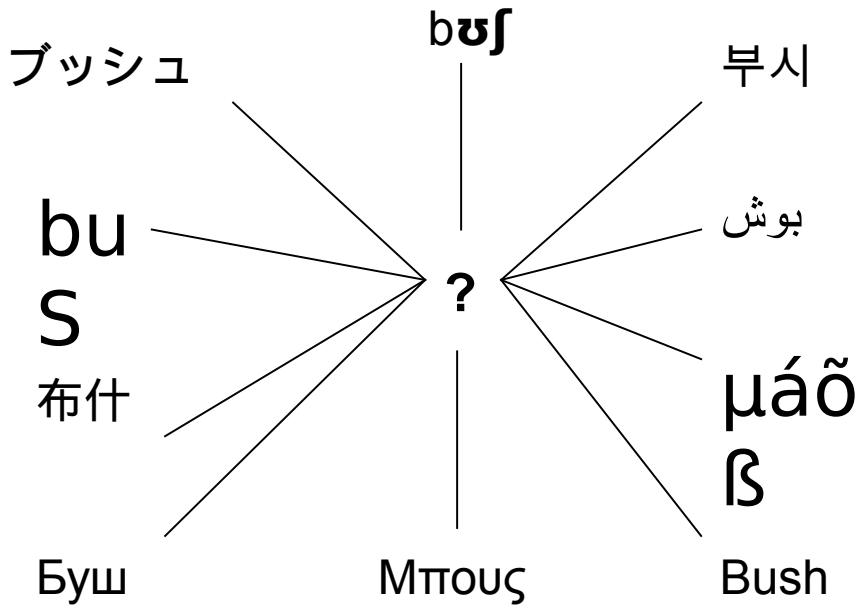
- Many NLP applications treat personal names
  - (CL)IR of text (MUC, TREC, TIPSTER)
  - (CL)IR of spoken documents (TDT)
  - Information extraction (ACE)
  - i18n, l10n
  - OCR/digitization
  - Semantic Web annotation
  - Homeland security and DoD (Aladdin, REFLEX)

and, of course,

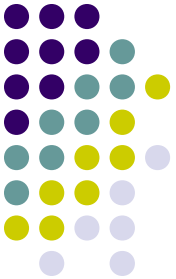
  - Family history research (PAF, TMG, etc.)

# The problem

- Storing and accessing proper nouns crosslinguistically



# What we won't address...

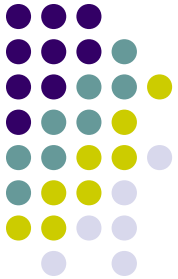


- Other types of proper nouns (organizations, countries, etc.)
- Position and title modifiers
- Selection and ordering of name components (surname, patronymics, etc.)
- Nicknames and hypocoristics
- Morphological variants (case, honorifics)
- Coreference, reduced forms, subsequent mentions

# Issues

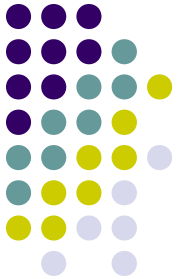
- Scope: some 6,000 languages
- Various types of writing systems
- Conventions: culturally/linguistically set
- Crosslinguistic: migrations, minorities
- Diachrony: spelling changes over time
- Innovation: names are continually invented
- Borrowings: names cross barriers

# Writing systems



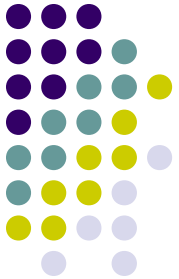
- Alphabetic: (roughly) one symbol / sound
  - Roman (Bush), Armenian (µáõß) , Georgian, etc.
- Syllabic: (usually) one symbol / syllable
  - Hiragana, Katakana ( ブッシュ ), Cherokee, etc.
- Abugidic (alphasyllabic): CV*
  - Devanagari (buS), Inuktitut, Lao, Thai, Tibetan, etc.
- Logographic: (roughly) one symbol / word
  - Hieroglyphs, Hieratic, Cuneiform, Hanzi ( 布什 ), etc.

# Special cases



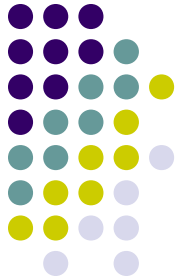
- Hangul
  - underlyingly alphabetic
  - sounds are arranged compositionally into syllabic symbols ( 부시 )
- Abjads
  - alphabetic, but without (some/all) vocalization
  - e.g. Arabic, Hebrew, Persian (بوش)

# Normalization



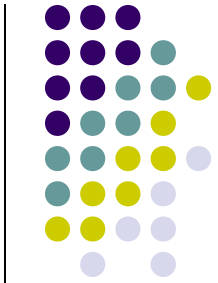
- Direction
  - left-right vs. right-left
  - horizontal vs. vertical
  - boustrophedonic
- Case
  - DeVon vs. Devon
- Vocalization
  - McConnell, St. John
- Diacritics
  - Étienne vs. Etienne
- Punctuation
- Abbreviations

# Related computational aspects



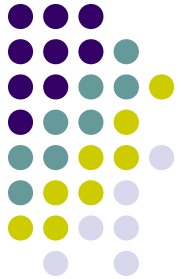
- Character sets, fonts, glyphs
- Input/output (keyboard, display)
- Collation (ordering, alphabetization)

# A few mapping strategies



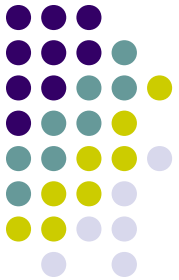
- Don't bother: lexical lookup
- Transcoding
- Transcription
- Transliteration
- Transduction
- Translation

# Lexical lookup



- Rote, literal access (e.g. hash tables)
  - Unending, expensive lexicon management task
  - Some automation possible (bitext, text mining)
- Bush → 布殊
- Some large-scale commercial undertakings
  - Hundreds of millions of names and variants, primarily European
  - Similar efforts exist for CJK conversion via lookup

# Transcoding

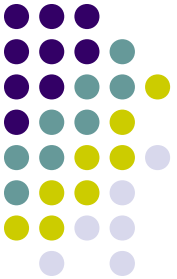


- Rote (mostly) character-by-character symbol conversion (e.g. Unix recode)
- x44 x61 x6e → xee xb3 xdd
- Even codes within a language vary
  - 布什 (Mainland China)
    布希 (Taiwan)
    布殊 (Hong Kong)
  - Osama bin Laden: 10 Hanzi variants
- Unicode helps, but does not solve the problems

# Transcription

- Conversion: (spoken) words → script
  - SAMPA (ASCII)
  - International Phonetic Alphabet (linguistics)
    - Bush → bʊʃ
  - Usually spoken language = transcribed language
- Sometimes as a strategy for crosslinguistic textual conversion
- Variation is a problem: whose dialectal/idiolectal pronunciation should be used?

# Transliteration



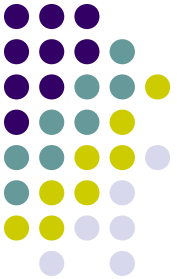
- Rewrite symbols of source language in target alphabet
- Bush → Буш
- Source/target sounds don't always align
  - 32 English spellings for Muammar Gaddafi
  - 6 Arabic spellings for Clinton
- Sensitive to properties of target language
  - e.g. Yuschenko vs. Iouchtchenko
- Romanization chaos: scores of schemes

# Transduction

- Mapping variable correspondences (transcription, transliteration), often (probabilistic) rule-based
- Implemented via algorithmic finite-state automata
  - e.g. Soundex (Russell, American, Daitch-Mokotoff), others
- Bush → buS

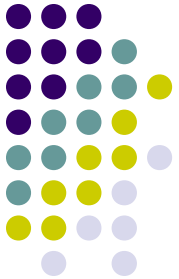
| Alternate spellings based upon easily confused letters | American soundex alternatives | Daitch-Mokotoff soundex alternatives |
|---|---|---|
| Bcller, Bebler, Beiler, Belber, Belier, Bellcr, Bellen, Bellor, Boller, Bcbler, and 152 others... | Beler, Beller | Aueler, Beler, Fbeler, Feler, Peler, Pfeler, Ppheler, Veler, Weler |

# Problems with Soundex



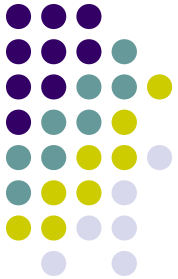
- Long names: Sivaramakrishnarao, Sivaramakrishnan, Sivaramarao
- Implausible collapses
- Anglocentric
- Alphabetic-based
- Not very efficient distributionally

# Translation



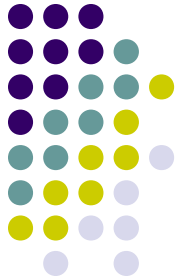
- Most widely used when logographic system is used
  - Names are rendered non-literally, non-phonemically to/from logograph (sequence)
- Great Salt Lake → 大鹽湖
- Creative, most opaque of mapping schemes

# Common techniques used



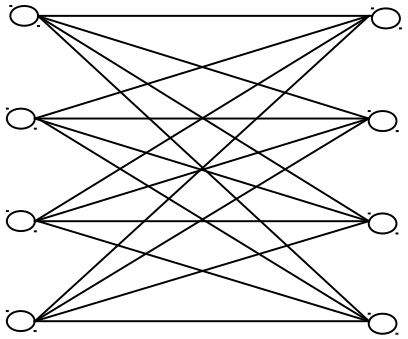
- Machine learning
  - Statistical/stochastic approaches (e.g. n-grams)
  - Entropy/noisy channel approaches
  - Rule-based transformational approaches
- String matching algorithms
  - Levenshtein edit distance (similarity measure)
  - Dynamic programming techniques
- Speech processing (recognition, TTS)
- Bitext mining, alignment metrics, indexing

# What's the best method?



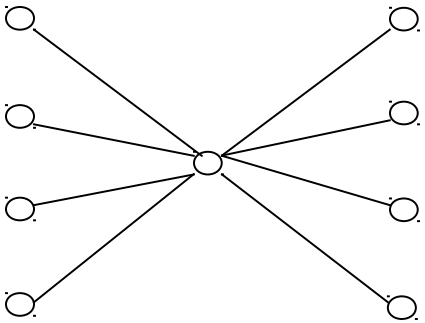
- One of schemes listed previously
  - All approaches are information-losing propositions
- Hybrid approaches combining several of these
  - Pipeline results
  - Poll different engines for optimal results
- How to generalize beyond a handful of languages?

# The direct model



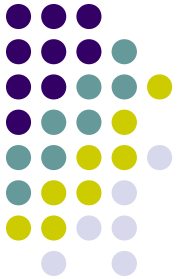
- Pairwise conversion between specific languages
- Potentially n x m components
  - Not all pairs will likely be needed, though
- Developer expertise a problem

# The pivot model



- Neutral "interlingua" or pivot
- n + m components
- What could serve as the pivot?
- Some small-scale examples exist
  - ISCII for Dravidian-script (South Asian) languages

# Pivot desiderata



- Neutral representation scheme
- Should address all possible writing systems
- Should assure as lossless a conversion as possible
- Should encode all necessary information
- Principled enough to allow algorithmic implementation
- Generative capability necessary
- Is it even possible to have only one pivot?

# Pivot = alphabet?

- English?
  - Consistency: very bad sound/symbol mapping
  - Anglocentricity
- IPA?
  - Transparency: difficult for non-linguists
  - Comprehensive, but not totally adequate
- Logographs would be problematic

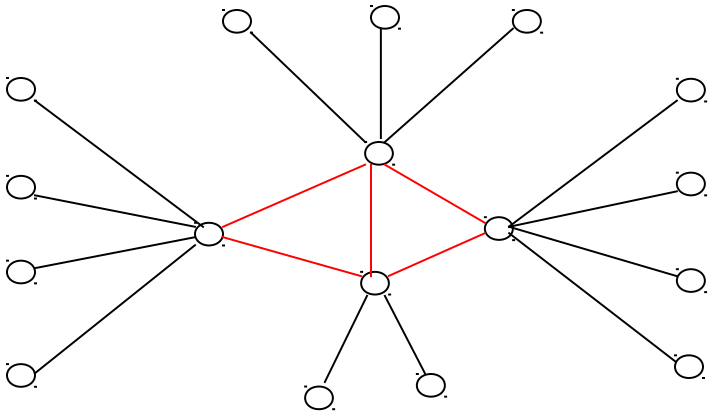
# Pivot = syllabic?

- Not as intuitive to alphabet users
- Syllable definition is still debated in some languages
- Ambisyllabicity
  - Mary, Brigham, Deryle

# Pivot = logographic?

- Need to invent character (sequences)
  - Meaning is not always obvious
- Impracticality: complexity of representation, script

# An articulated pivot approach



- More than one "pivot", feed into each other
- n + m + p components
- Allows grouping of typologically similar languages
- Intra-pivot links could represent current research results (most commonly used languages)

# Conclusions

- Rich area for current research
- The issues are daunting
- Various approaches are being implemented
- MT has tackled some of the same problems
- A principled solution might involve some type of articulated pivot
- Open annotation environment, sharable resources, algorithm libraries
- Genealogists can contribute