MORPHOLOGICAL ANALYSIS OF INUKTITUT

STATISTICAL NATURAL LANGUAGE PROCESSING FINAL PROJECT

Gina Cook

Concordia University
ENCS
Engineering & Computer Science

Why Inuktitut?

- Official language of Nunavut
  - Government
  - Education

- Search Engines
- Spellcheckers
- Dictionaries, Thesaurus
- Grammar checkers

Inuktitut Resources

National Research Council Canada | Conseil national de recherches Canada
Institute for Information Technology | Institut de technologie de l'information

Inuktitut Computing dot C A

Français

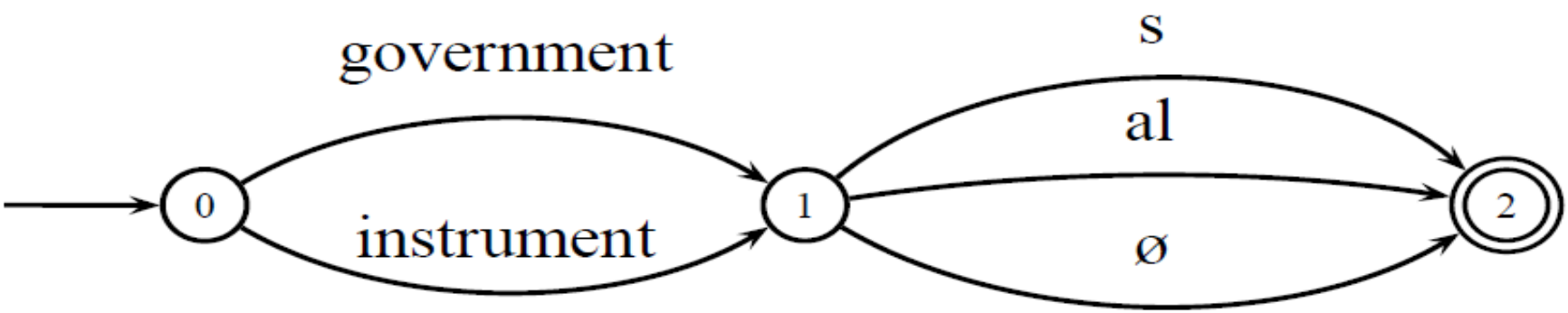
Mission: To facilitate the use of Inuktitut in its written form on computers and the web by providing useful tools and links to important resources.

Morning at Iqaluit, Nunavut

Inuktitut Morphological Analyser ***Run***
Bibliographic References
Dictionary: "Inuktitut - A Multi-dialectal Outline Dictionary" by Alex Spalding
Display and Input of Inuktitut Syllabic Characters - Unicode and Legacy Fonts
Inuktitut-English Parallel Corpus
Inuktitut Linguistics for Technocrats by Mick Mallon
Linguistic Data Base
NANIVARA - Inuktitut Search Engine ***Run***
Searching the Nunavut Hansard ***Run***
Transcoder ***Run***
Transliteration of Web Pages

Morphology 101

- Most languages have morphology
- Most morphology consists of either suffixes or prefixes



Why Morphological Parsing?

- Information Retrieval
  - Remove morphs
- Machine Translation
- Named Entity Recognition
- Natural Language Understanding
  - Use morphs

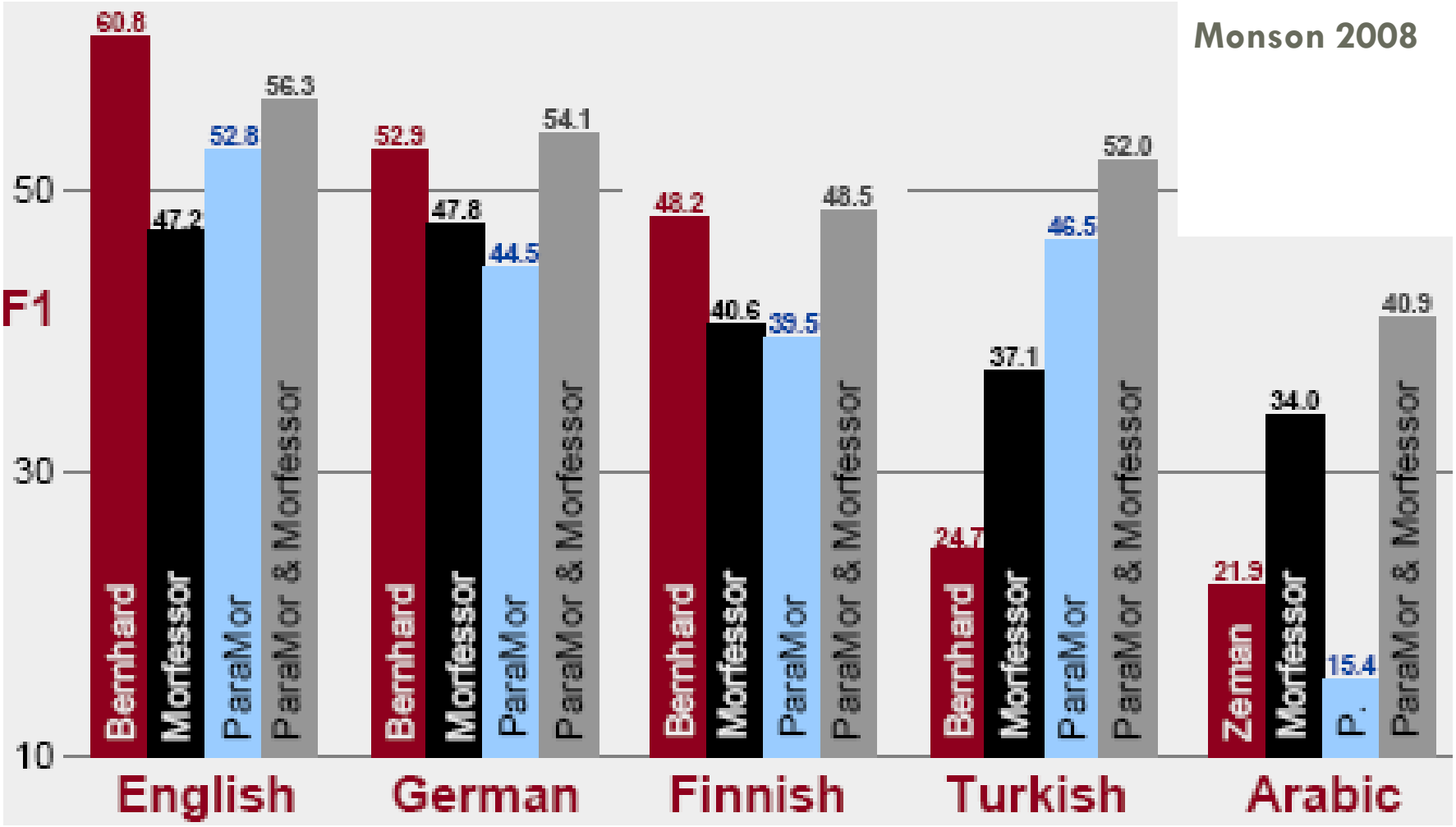
Stemming

- Useful for Information Retrieval
- Reduces feature space

Full Parsing

- Natural Language Generation
- Machine Translation
- Named Entity Tagging
- Text Summarization

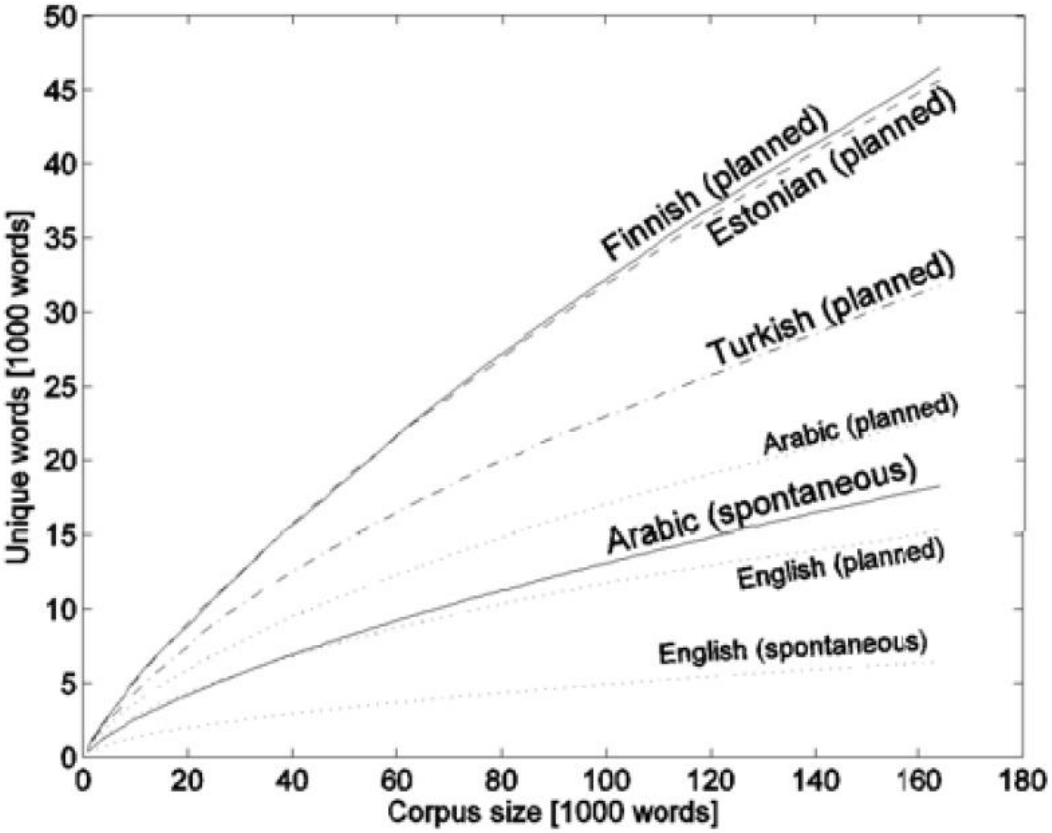
Low accuracy (F scores < 50%)
Performance heavily dependant on language type



Morphology 102

- Agglutinative languages more morphemes per word (Pirkola 2001)
- And unlimited words (Kurimo 2008)

Vietnamese	1,06
Yoruba	1,09
English	1,68
Old English	2,12
Swahili	2,55
Turkish	2,86
Russian	3,33
Inuit (Eskimo)	3,72



Five Approaches

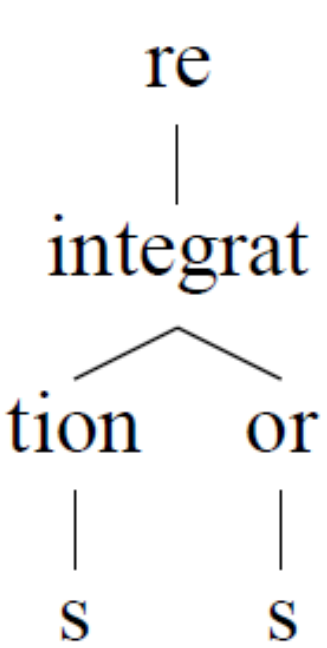
1. Transition Likelihood
    1. **Harris 1958**
    2. Johnson & Martin 2003 (English, Innuktitut) HubMorph
    3. Bernhard 2007 (English, German, Finnish)
2. Minimum Description Length
    1. Brent 1995 MBDP-1
    2. De Marken 1995 Composition and Perturbation
    3. Goldsmith 2001,2006 Linguistica
    4. **Creutz 2006 Morfessor**
3. Paradigms
    1. Goldsmith 2001,2006 Linguistica
    2. Snover 2002
    3. Monson 2008 ParaMor
4. Word Edit Distance & Latent Semantic Analysis of word context
    1. Yarrow & Wicentowski 2000
5. Phonotactic/ Allomorphy
    1. Heinz MBDP-Phon-Bigrams

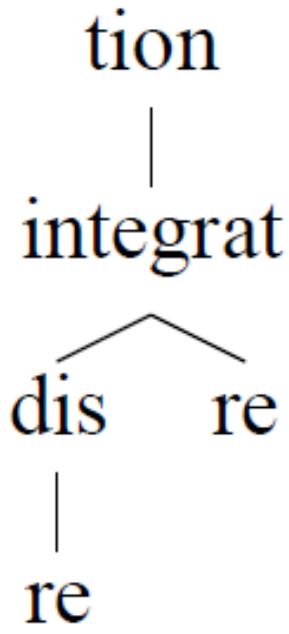
Compression

- Morphological Parsing as Compression
  - Tries
  - Minimum Description Length

Harris 1955

- Forward and Backward Tries





Minimum Description Length

- The more morphs you find, the smaller the key

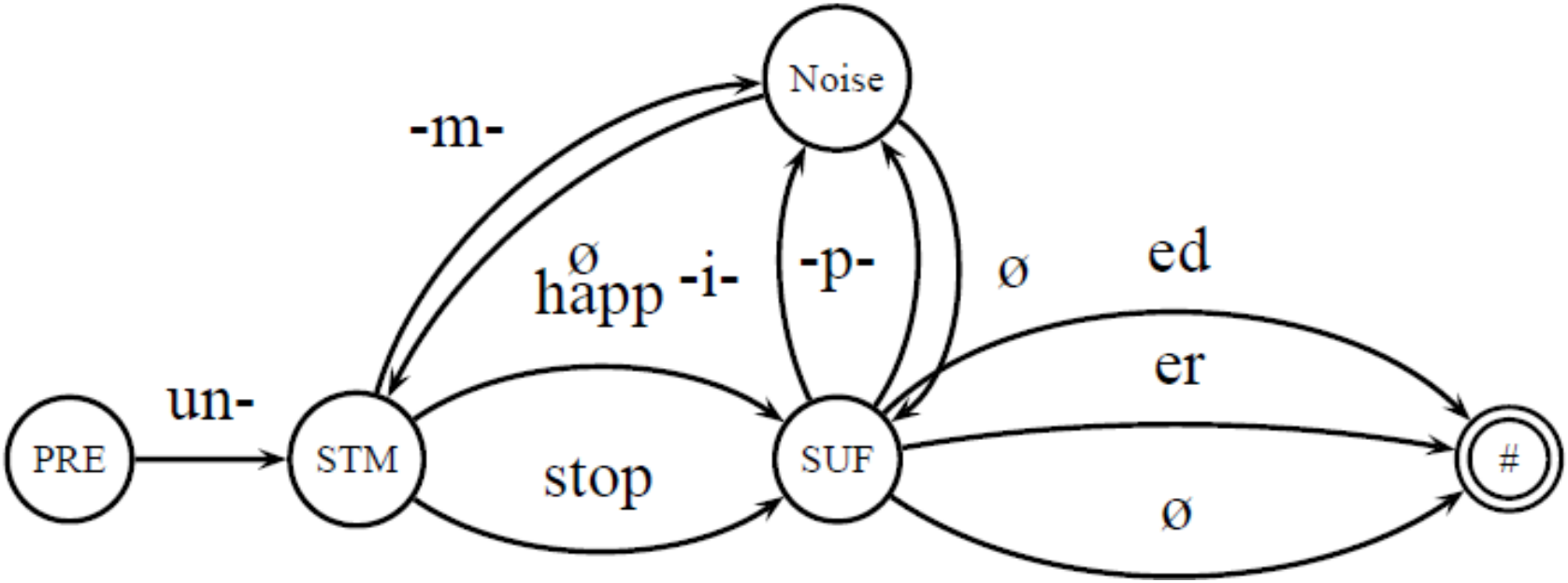
```
Data:
@ kati rsui sima giaqanir mik
@ kati maniu laur tu mik
@ tusaa jimmaringulaur sima juq
@ tusaa jimmaringu laur sima juq
@ mali tsiariaqa laur tugut
@ tiki laur tugut

Compressed Data:
T= 32
1 2 3 4 5 6 1 2 7 8 9 6 1 10 11 4 12 1
10 13 8 4 12 1 14 15 8 16 1 17 8 16
```

```
Key:
4       sima
8       laur
9       tu
10      tusaa
5       giaqanir
7       maniu
16      tugut
1       @
14      mali
13      jimmaringu
11      jimmaringulaur
2       kati
3       rsui
6       mik
12      juq
17      tiki
15      tsiariaqa
```

Creutz 2005: Morfessor

- A Hidden Markov Model with 4 states
- Morphemes are the strings which transition between them and the probabilities of that transition



Morfessor Performance on Inuktitut

- 7% Precision
- 7% Recall
- Why?

Sample output by length(in number of sylables):

```
17 saimmatitsigasuarutaujaarnaturinalauqsimavuq
17 allasimajuliuqpaliagutiqarumaliqsungattauq
17 aatutiqangitsiammarittuuqutigilaursimajara
16 uqalimaagaksaliarijauqattalirajartuunnik
16 tusaumautiqattautitsiarunnarnirsaugajarnirmut
16 sivumugiallagutiqaqsimaliraluartillugit
16 piruqpalialaqititaugutigingunnatanginnik
16 pinasuagatuarijaunngusutsimagalaaganilu
16 katiqatigiinnirsaugutigijunnarsigajarmauk
16 isumaksarsiurutiqallattaaqattaraluarpita
15 taimannganimmarialunitauqataummigamilli
15 pivallirtitsijummariujunnarasugillugu
15 pigunnaniqsauqigigajannguataraluaminnit
15 kiinaujaqaqtitaugutigigunnatanginnilu
15 kiinaujaliurasuarutiksalirinasuarluta
15 isumagiqasiutigiaqangikkaluarpitigut
15 ikpigusugiallagutiginirsarilirakkit
15 avikturSimaugutigivalliatuinnartattinnik
```

Morphology 103

- Morphemes cannot appear in any order, the ordering is fixed
  - Within a language
  - For all human possible languages

FieldWorker 0.005

- Goals
  - Create a general system
  - Grow from an initial assumption of root+suffix (which is true for all human languages argument+head)
  - Expand to allow prefixing, multiple suffixes, compounding
  - Flexible enough to allow for allomorphs
  - Flexible enough to allow for nonlocal dependencies

Learning Grammar from morphological precedence relations

- Discover template
  - Take long words containing seed morphemes to discover full template
- Discover morphemes
  - Create dense corpora to find morphemes for each template category

Creating an Inuktitut Corpus

- Nunavut Hansard
  - Spelling is unsystematic, introduces too much noise for statistical learning

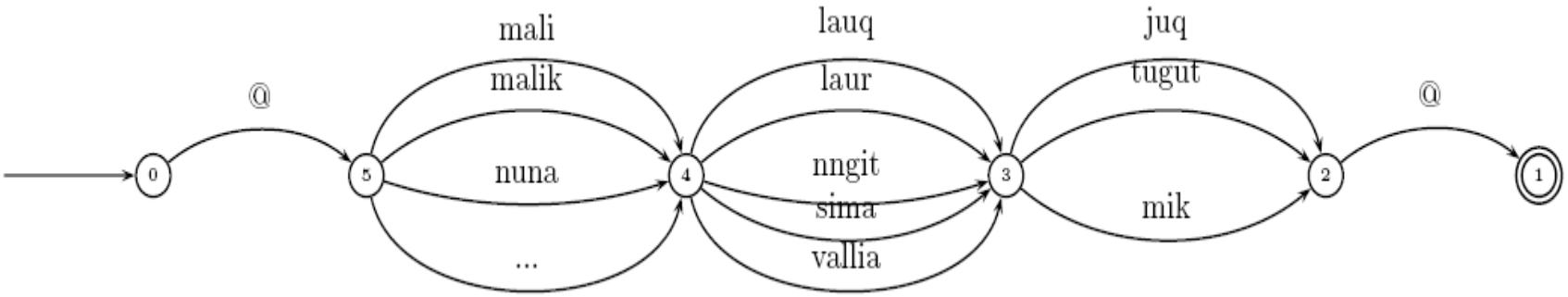
```
Examples of spelling variation:
10 aaggaaqtuqangittuq
6 aaggaaqtuqanngittuq
2 aanniaqannangittulirijikkuni
5 aanniaqannangittulirijikkunni
```

- Created a corpus from Inuktitut Magazine vol. 102-104
  - Parallel corpus in Inuktitut, English and French
  - 17,000 Inuktitut words for 32,000 English words
  - Consistent spelling

Overview

1. Corpus – word list
2. Word list – ranked possible morphs
3. Possible morphs – seed list
4. Seed list – precedence relations
5. Precedence relations – dense corpus
6. Dense corpus – precedence relations
7. Iterate

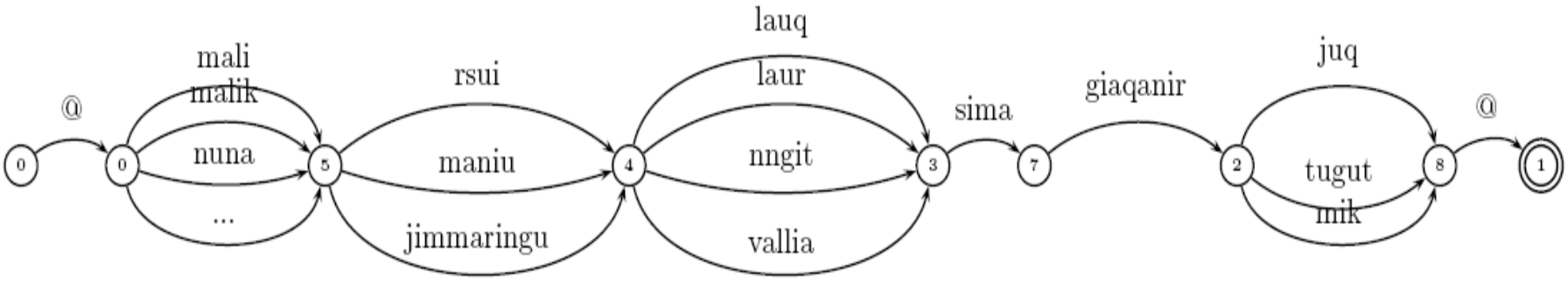
Sample Seed list



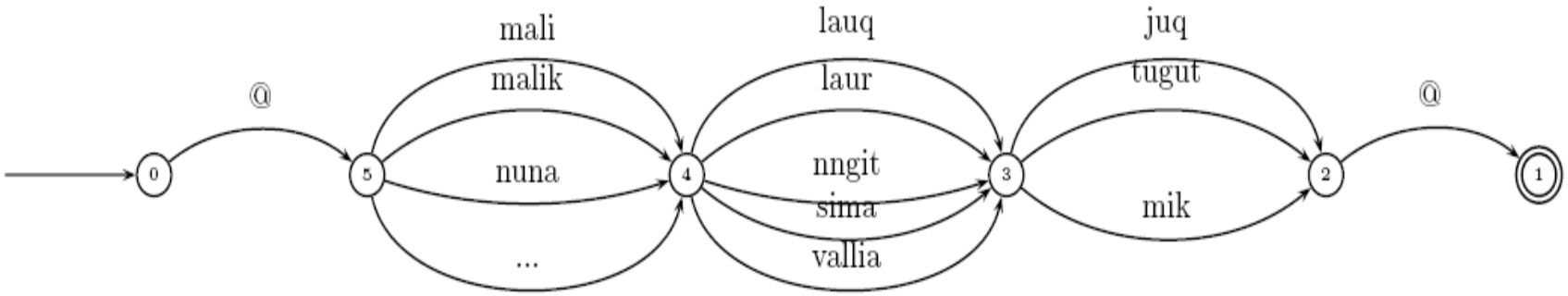
Gives a dense mini-corpus


```
1 + +kati+ +rsui+ +sima+ +giaqanir+ +mik+ +
1 + +kati+ +maniu+ +laur+ +tu+ +mik++
1 + +tusaa+ +jimmaringulaur+ +sima+ ++ +juq+ +
1 + +tusaa+ +jimmaringu+ +laur+ +sima+ +juq+ +
1 + +mali+ +tsiariaqa+ +laur+ ++ +tugut+ +
1 + +tiki+ ++ +laur+ ++ +tugut+ +
```

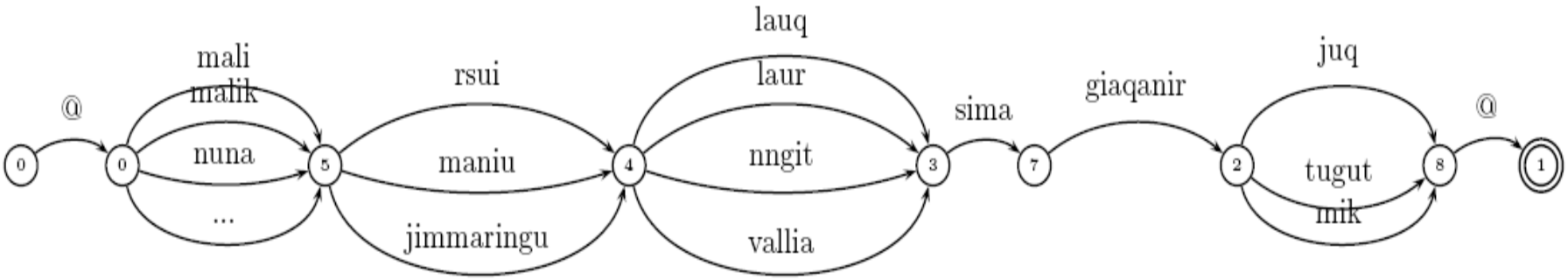
Which gives a new template



Progress




```
1 + +kati+ +rsui+ +sima+ +giaqanir+ +mik+ +
1 + +kati+ +maniu+ +laur+ +tu+ +mik++
1 + +tusaa+ +jimmaringulaur+ +sima+ ++ +juq+ +
1 + +tusaa+ +jimmaringu+ +laur+ +sima+ +juq+ +
1 + +mali+ +tsiariaqa+ +laur+ ++ +tugut+ +
1 + +tiki+ ++ +laur+ ++ +tugut+ +
```



Evaluation

1 + +kati+ +rsui+ +sima+ +gia!qanir+ +mik+ +

1 + +kati+ +maniu+ +laur+ +tu+ +mik++

1 + +tusaa+ +jimma!ringu!laur+ +sima+ ++ +juq+ +

1 + +tusaa+ +jimma!ringu+ +laur+ +sima+ +juq+ +

1 + +mali+ +tsia!ria!qa+ +laur+ ++ +tugut+ +

1 + +tiki+ ++ +laur+ ++ +tugut+ +

Recall = 22/32	69%
Precision = 18/22	81%

Evaluation

- Recall goes up as the model iterates
- Precision goes down as the model iterates

- Where to stop the model?

Morpho Challenge 2009 August

- Run my algorithm on English, German, Finish, Turkish and Arabic corpora of Morpho Challenge 2008
- If I am able to achieve respectable F-scores (~50%)
- Submit my algorithm to Morpho Challenge 2009

References

- Brent, Michael R and Xiaopeng Tao. 2001. Chinese text segmentation with mbdp-1: Making the most of training corpora. In 39th Annual Meeting of the ACL, pages 82–89.
- de Marcken, Carl. 1995. Acquiring a lexicon from unsegmented speech. In 33rd Annual Meeting of the ACL, pages 311–313.
- Goldsmith, J.A. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27:2 pp. 153-198.
- Johnson, Mark. 2008a. Unsupervised word segmentation for Sesotho using adaptor grammars. In Tenth Meeting of ACL SIGMORPHON, pages 20–27. ACL, Morristown, NJ.
- Johnson, Mark. 2008b. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In 46th Annual Meeting of the ACL, pages 398–406. ACL, Morristown, NJ.
- Kanungo, Tapas. 1999. "UMDHMM: Hidden Markov Model Toolkit," in "Extended Finite State Models of Language," A. Kornai (editor), Cambridge University Press. http://www.kanungo.com/software/software.htm http://www.umiacs.umd.edu/~resnik/nlstat_tutorial_summer1998/Lab_hmm.html
- Pirkola, Ari. 2001. Morphologcial Typology of Languages for Information Retrieva, *Journal of Documentation 57* (3), 330-348.
- Schone, P., & Jurafsky, D. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of CoNLL-2000 and LLL-2000*, pp. 67--72 Lisbon, Portugal.
- Venkataraman, Anand. 2001. A statistical model for word discovery in transcribed speech. Computational Linguistics, 27(3):352–372.