The SPECIALIST NLP Tools
Suffix Derivations

Dr. Chris J. Lu

The Lexical Systems Group

NLM. LHNCBC. CGSB

Dec., 2012
Table of Contents

• Introduction
  ▪ NLP Tools
  ▪ Lexical Tools – derivational flows
  ▪ Derivational variants

• Suffix derivations
  ▪ Current state
  ▪ Systematic approaches
  ▪ SD-Rules
    o new SD-Rules
    o parents – child SD-Rules
  ▪ Results

• Questions
Introduction - NLP

• Natural Language (English)
  ▪ is ordinary language that humans use naturally
  ▪ may be spoken, signed, or written

• Natural Language Processing
  ▪ NLP is to process human language to make their information accessible to computer applications
  ▪ The goal is to design and build software that will analyze, understand, and generate human language
  ▪ Most NLP applications require knowledge from linguistics, computer science, and statistics
NLP Example

Features:
- Information retrieval
- Filter
- Summarize
- Alert & suggestion
- Questions answering
- …

Questions
Symptoms

NLP System
NLP System

- EMR (Electronic Medical Records)
- MEDLINE Article/Abstract
- ...

Features:
- Information retrieval
- Filter
- Summarize
- Alert & suggestion
- Questions answering
- ...

- The SPECIALIST Lexicon
- UMLS semantic network
- ...

Domain Knowledge

NLP algorithm programs

Structured Data
NLP Core Tasks

Example: Information retrieval (search engine)
- Tokenize & tagging (entity recognition)
  - break inputs into words <Text Tools, wordInd>
  - POS tagging <dTagger>
  - Other annotation <Visual Tagging Tool, VTT>
- spelling check
  - suggest correct spelling for misspelled words <gSpell>
- lexical variants (query expansion)
  - spelling variants, inflectional/uninflectional variants, synonyms, acronyms/abbreviations, expansions, derivational variants, etc. <Lexical Tools, LexAccess, LexCheck, STMT>
- semantic knowledge (concept mapping)
  - map text to Metathesaurus concepts <MetaMap, MMTX, STMT>
  - Word Sense Disambiguation <TC – StWSD>
Introduction - NLP Tools

The SPECIALIST NLP Tools

LexBuild

The SPECIALIST LEXICON

The SPECIALIST NLP Tools

Lexical Tools - 2013

• Lexical Tools include 7 tools:
  ▪ lvg (Lexical Variants Generation)
    o 62 flow components
    o 39 options
  ▪ lgt (Lexical GUI Tool)
  ▪ norm/luiNorm
  ▪ toAscii
  ▪ wordInd
  ▪ fields
Derivational Related

- 7 flow components:
  - -f:d
  - -f:dc
  - -f:R
  - -f:G
  - -f:Ge
  - -f:Gn
  - -f:v

- 3 flow specific options
  - -kd: 1|2|3 (default: 1)
  - -kdn: B|N|O (default: O)
  - -kdt: Z|S|P (default: ZSP)
LVG - Derivation Examples

- Please input a term (type "Ctl-d" to quit) >

hyperuricemic

hyperuricemic|hyperuricemic|<noun>|<base>|d|1|
hyperuricemic|hyperuricemia|<noun>|<base>|d|1|
hyperuricemic|hyperuricemic|<adj>|<base>|d|1|

hyperuricemic|uricemia|<noun>|<base>|R|2|
hyperuricemic|hyperuricemia|<noun>|<base>|R|2|
Derivations in NLP Application

- hyperuricemic|adj, E0317343, no CUI
- hyperuricemia|noun, E0032862, is a UMLS Metathesaurus term (C0740394)
Derivational Variants

• **Words** related by a derivational process
  ▪ Used to create new words based on existing words
  ▪ Meaning change (related)
  ▪ Category change
  ▪ Derivational process: suffix, prefix, and conversion

• Focus on relatedness (no direction)
Derivation Types (-kdt)

- Example (kind|adj):
  - zeroD: kind|adj|kind|noun
  - prefixD: kind|adj|unkind|adj
  - suffixD: kind|adj|kindly|adv
Derivational Pair

- Each link and the associated two nodes in derivational network define a derivational pair
- Includes base forms and syntactic category information
- Bi-directional
- Only involves one or none derivational affix
- Lvg format: base 1|category 1|base 2|category 2
- Examples:
  - kind|adj|kindness|noun
  - kind|adj|kindly|adv
  - kind|adj|unkind|adj
  - kind|adj|kind|noun
PrefixD and ZeroD

- Added to Lexicon/Lexical Tools with a systematic method in 2012 release

- A Systematic Approach for Automatically Generating Derivational Variants in Lexical Tools Based on the SPECIALIST Lexicon
  Lu, Chris J.; McCreedy, Lynn; Tormey, Destinee; and Browne, Allen C.
  IEEE IT Professional Magazine, May/June, 2012, p. 36-42

- It also includes nomD (nominalization derivations)
SuffixD - Process

• Also called a postfix or ending
• Placed after the stem of a word to form another word
• Several hundreds of derivational suffixes
• SD-Pairs:
  ▪ kind|adj|kindness|noun
  ▪ kind|adj|kindly|adv
Derivational Flow – SD

• Facts
  ▪ 4,559 derivational pairs in DB (2011-)

<table>
<thead>
<tr>
<th>Base 1</th>
<th>Category 1</th>
<th>Base 2</th>
<th>Category 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>treatment</td>
<td>noun</td>
<td>treat</td>
<td>noun</td>
</tr>
</tbody>
</table>

• Rules
  ▪ 97 SD-Rules
  ▪ Use exceptions to increase precision

EXAMPLE: retirement | noun | retire | verb
RULE: ment$ | noun | $ | verb
EXCEPTION: apartment | apart;
EXCEPTION: basement | base;
EXCEPTION: department | depart;
...
SD-Rules (Trie)

• retirement noun => retire verb

EXAMPLE: retire verb retirement noun
RULE: $ verb ment$ noun
EXCEPTION: apart apartment;
...

EXAMPLE: conformant adj conformance noun
RULE: ant$ adj $ verb
EXCEPTION: important import;
...

EXAMPLE: relaxant adj relax verb
RULE: ant$ adj $ verb
EXCEPTION: important import;
...

EXAMPLE: conformant adj conformance noun
RULE: ant$ adj $ verb
EXCEPTION: important import;
...

EXAMPLE: fluent adj fluency noun
RULE: ent$ adj ency$ noun
EXCEPTION: emergency emergent;
...

EXAMPLE: retirement noun retire verb
RULE: ment$ noun $ verb
EXCEPTION: apartment apart;
...
SD-Rules Filters

- **Exception filter**
  - Exclude exceptions for the rules
  - Implemented in the Trie
  - depart|verb|department|noun
- **Word length filter**
  - Exclude short word
  - Default (min.) value is 3
  - moment|noun|mo|verb
- **Stem length filter**
  - stem length = word length – suffix length
  - Default (min.) value is 3
  - lament|noun|la|verb
- **Domain filter**
  - Exclude words not in Lexicon
  - color|verb|colorment|noun
Derivational Flow - Evaluation

• Facts
  ▪ 4,559 derivational pairs (2011-)
  ▪ Coverage is low
  ▪ Static data: not grown with Lexicon …

• Rules
  ▪ 97 SD-Rules
  ▪ Accuracy, how good are these rules?
  ▪ Coverage & frequency?
Goal - Challenges

- To establish a systematic approach and maintainable system for suffix derivations to reach overall accuracy rate of 95% with higher coverage.

- Facts (virtually 100% accurate)
  - focus on higher coverage
  - include more derivational pairs known to Lexicon
  - grow proportionally with Lexicon annually

- Rules: establish a systematic approach to
  - evaluate and refine existing SD-Rules
  - add new SD-Rules
  - handle issues of parents-child SD-Rules
  - higher coverage and accuracy (95%)
SD - Facts

• Known:
  ▪ Lexicon
  ▪ Nominalization (nomD)
  ▪ Existing 97 SD-Rules, used as SD-Rule candidates

• Process:
SD – Facts

SD – Rules (97):
... asia$|noun|astic$|adj
... ate$|verb|ation$|noun
... ate$|verb|ative$|noun
...

Lexicon (616,328):
... locate|verb|E0037939
location|noun|1|E0037940
...
... state|verb|E0057695
station|noun|E0057711
...
Lexicon (616,328):
...
locate|verb|E0037939
location|noun|1|E0037940
...
state|verb|E0057695
station|noun|E0057711
...

SD – Rules (97):
...
asia$|noun|astic$|adj
ate$|verb|ation$|noun
ate$|verb|ative$|noun
...

Raw SD-Pairs (2,025):
...
compensate|verb|E0018113|compensation|noun|E0018118
...
locate|verb|E0037939|location|noun|1|E0037940
...
state|verb|E0057695|station|noun|E0057711
...

Lexicon

SD – Rules

Raw SD-pairs

Nominalization

Tag

Negation
SD Facts - Nominalization

- The process of producing a noun from a verb or an adjective via the derivational suffix
- Coded in Lexicon
- A type of suffixD (zeroD)
- Bi-directional

```
{base=locate
text=E0037939
cat=verb
variants=reg
tran=np
link=advbl
cplxtran=np,advbl

nominalization=location|noun|E0037940
}

{base=location
text=E0037940
cat=noun
variants=reg
variants=uncount
compl=pphr(of,np)
compl=pphr(by,np)

nominalization_of=locate|verb|E0037939
}
```
NomD Process

• Raw nomD pairs: retrieve all nominalization information from Lexicon
• Filters:
  ● Pattern filter: exclude invalid suffixD for verb particle nomD
    Pattern-1: baseParticle|noun|base|verb => backup|noun|back|verb
    Pattern-2: base-Particle|noun|base|verb => cut-through|noun|cut|verb
    Pattern-3: inflParticle|noun|base|verb => grownup|nou|grow|verb
    Pattern-4: infl-Particle|noun|base|verb => salting-in|noun|salt|verb
    Particle Exception: “per” => shopper|noun|shop|verb
  ● Exception filter: exclude other known nomD pairs
    Examples:
    face-saving|noun|save|verb
    decision-making|noun|make|verb
    merry-making|noun|make|verb
    lovemaking|noun|make|verb
    ...


SD – Facts

- Automatically tag valid nomD as valid suffixD

Raw SD-Pairs (1,586/2,025, 78%):

... compensate|verb|E0018113|compensation|noun|E0018118
...
... locate|verb|E0037939|location|noun|1|E0037940|yes
...
... state|verb|E0057695|station|noun|E0057711
...
SD – Facts

- Manually tag the rest by linguists

**Raw SD-Pairs (439/2,025, 22%)**:

... compensate|verb|E0018113|compensation|noun|E0018118|yes
...
... state|verb|E0057695|station|noun|E0057711|no
...
SD – Facts

- Automatically tag negation on valid SD-pairs

**Valid SD-Pairs (2,020/2,025, 99.75%):**

... compensate|verb|E0018113|compensation|noun|E0018118|yes|O
...
... state|verb|E0057695|station|noun|E0057711|no
...
Derivation – Negations (-kdn)

• Derivational variants are used to find related variants in a wider coverage in NLP. Negative derivations should be filtered out because the big meaning drift, such as convulsive and anti-convulsive; able and unable; use and useless, etc.

• Example (kind|adj):
  ▪ prefixD:
    o Class N (10): anti-, contra-, counter-, dis-, il-, im-, ir-, mis-, non-, un-
    o Class O (129): abs-, af-, Afro-, ambi-, etc.
    o Class B (6): a-, an-, de-, dys-, in-, under-
  ▪ suffixD:
    o -less: care|careless
  ▪ zeroD: no negations
SD – Facts

SD-Rules: ate$|verb|ation$|noun

• Raw SD-Pairs: 2,025
  ▪ Valid (yes): 2,020 -> add to SD Facts
  ▪ Invalid (no): 5
    o delimitate|verb|E0021381|delimitation|noun|E0021382|no
    o legate|verb|E0540056|legation|noun|E0593456|no
    o rate|verb|E0052016|ration|noun|E0052025|no
    o predate|verb|E0068010|predation|noun|E0068011|no
    o state|verb|E0057695|station|noun|E0057711|no

▪ Accuracy rate: 99.75% (= 2020/2025)
SD Facts - Results

• Apply all candidate SD-Rules on Lexicon (2013)

• Total raw SD-Pairs: 51,599
  ▪ nomD: 14,368 (27.85%, auto-tagged)
    o Not covered in SD-Rules: 2,281 (4.42%)
    o Covered by SD-Rules: 12,087
  ▪ Covered by SD-Rules: 49,318 (95.58%)
    o From nomD: 12,087
    o Manual Tag: 37,231 (72.15%, manual tagged)

• Total raw SD-Pairs: 51,599 (Tagged stats)
  ▪ Valid: 44,832 (86.89%)
    o Class N: 564 (1.26%)
    o Class O: 44,268 (98.74%)
  ▪ Invalid: 6,767 (13.11%)
SD Facts

- 28% are auto-tagged
- SD-Rules covers 96% of SD-Pairs known in Lexicon

Raw SD-Pairs (51,599)

- SD-Rules-Tag
  - 37,231
  - 72%

- SD-Rules-nomD
  - 12,087
  - 24%

- nomD
  - 2,281
  - 4%
SD Facts – Valid/Invalid SD-Pairs

Raw SD-Pairs (51,599)

Valid
44,832
87%

Invalid
6,767
13%
SD Facts – Negation

Valid SD-Pairs (44,832)

Class O
44,268
99%

Class N
564
1%
# 97 SD Rules

- **Baseline:** 97 SD-Rules on Lexicon (2013)

<table>
<thead>
<tr>
<th>No.</th>
<th>Accuracy</th>
<th>Total</th>
<th>yes</th>
<th>no</th>
<th>SD-Rule</th>
<th>Accum Total</th>
<th>Accum Yes</th>
<th>System Accuracy</th>
<th>System Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100.00%</td>
<td>2723</td>
<td>2723</td>
<td>0</td>
<td>$</td>
<td>adj</td>
<td>ness</td>
<td>noun</td>
<td>2723</td>
</tr>
<tr>
<td>2</td>
<td>100.00%</td>
<td>1278</td>
<td>1278</td>
<td>0</td>
<td>ability</td>
<td>noun</td>
<td>able</td>
<td>adj</td>
<td>4001</td>
</tr>
<tr>
<td>3</td>
<td>100.00%</td>
<td>1215</td>
<td>1215</td>
<td>0</td>
<td>ization</td>
<td>noun</td>
<td>ize</td>
<td>verb</td>
<td>5216</td>
</tr>
<tr>
<td>4</td>
<td>100.00%</td>
<td>366</td>
<td>366</td>
<td>0</td>
<td>osis</td>
<td>noun</td>
<td>otic</td>
<td>5582</td>
<td>5582</td>
</tr>
<tr>
<td>5</td>
<td>100.00%</td>
<td>326</td>
<td>326</td>
<td>0</td>
<td>le</td>
<td>adj</td>
<td>ly</td>
<td>adv</td>
<td>5908</td>
</tr>
<tr>
<td>71</td>
<td>72.09%</td>
<td>43</td>
<td>31</td>
<td>12</td>
<td>ious</td>
<td>adj</td>
<td>ly</td>
<td>noun</td>
<td>40124</td>
</tr>
<tr>
<td>72</td>
<td>64.22%</td>
<td>109</td>
<td>70</td>
<td>39</td>
<td>ant</td>
<td>adj</td>
<td>ate</td>
<td>verb</td>
<td>40233</td>
</tr>
<tr>
<td>73</td>
<td>62.65%</td>
<td>332</td>
<td>208</td>
<td>124</td>
<td>noun</td>
<td>list</td>
<td>noun</td>
<td>40565</td>
<td>38483</td>
</tr>
<tr>
<td>93</td>
<td>0.74%</td>
<td>136</td>
<td>1</td>
<td>135</td>
<td>ia</td>
<td>noun</td>
<td>ian</td>
<td>noun</td>
<td>45881</td>
</tr>
<tr>
<td>94</td>
<td>0.37%</td>
<td>273</td>
<td>1</td>
<td>272</td>
<td>a</td>
<td>noun</td>
<td>an</td>
<td>noun</td>
<td>46154</td>
</tr>
<tr>
<td>95</td>
<td>0.00%</td>
<td>358</td>
<td>0</td>
<td>358</td>
<td>gram</td>
<td>noun</td>
<td>graphy</td>
<td>noun</td>
<td>46512</td>
</tr>
<tr>
<td>96</td>
<td>0.00%</td>
<td>228</td>
<td>0</td>
<td>228</td>
<td>gram</td>
<td>noun</td>
<td>graphic</td>
<td>adj</td>
<td>46740</td>
</tr>
<tr>
<td>97</td>
<td>0.00%</td>
<td>57</td>
<td>0</td>
<td>57</td>
<td>verb</td>
<td>ably</td>
<td>adv</td>
<td>46797</td>
<td>40289</td>
</tr>
</tbody>
</table>
Top 72/97 SD-Rules

- Accuracy rate: 95.13%
- Coverage rate: 95.00%
- Used to predict derivations in general English
New SD-Rules (nomD)

- Derive SD-Rules from known SD-pairs:
  - nomD (14,638 SD)
  - location|noun|locate|verb => ion$|noun|e$|verb
  - Identified 513 possible SD-Rules

<table>
<thead>
<tr>
<th>Identified Rules</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>adj</td>
</tr>
<tr>
<td>e$</td>
<td>verb</td>
</tr>
<tr>
<td>$</td>
<td>adj</td>
</tr>
<tr>
<td>ility$</td>
<td>noun</td>
</tr>
<tr>
<td>ation$</td>
<td>noun</td>
</tr>
<tr>
<td>e$</td>
<td>adj</td>
</tr>
<tr>
<td>ce$</td>
<td>noin</td>
</tr>
<tr>
<td>iness$</td>
<td>noun</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ious$</td>
<td>adj</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sm$</td>
<td>noun</td>
</tr>
<tr>
<td>ty$</td>
<td>noun</td>
</tr>
<tr>
<td>ty$</td>
<td>noun</td>
</tr>
</tbody>
</table>
New SD-Rules (nomD)

- Derive SD-Rules from known SD-pairs:
  - nomD (14,638 SD)
  - location|noun|locate|verb => ion$|noun|e$|verb
  - Identified 513 possible rules

<table>
<thead>
<tr>
<th>Identified Rules</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>adj</td>
</tr>
<tr>
<td>e$</td>
<td>verb</td>
</tr>
<tr>
<td>$</td>
<td>adj</td>
</tr>
<tr>
<td>ility$</td>
<td>noun</td>
</tr>
<tr>
<td>a$</td>
<td>tion$</td>
</tr>
<tr>
<td>e$</td>
<td>adj</td>
</tr>
<tr>
<td>ce$</td>
<td>noun</td>
</tr>
<tr>
<td>iness$</td>
<td>noun</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ious$</td>
<td>adj</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sm$</td>
<td>noun</td>
</tr>
<tr>
<td>ty$</td>
<td>noun</td>
</tr>
<tr>
<td>ty$</td>
<td>noun</td>
</tr>
</tbody>
</table>

1. Duplicates: remove
New SD-Rules (nomD)

- Derive SD-Rules from known SD-pairs:
  - nomD (14,638 SD)
  - location|noun|locate|verb => ion$|noun|e$|verb
  - Identified 513 possible rules

<table>
<thead>
<tr>
<th>Identified Rules</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>adj</td>
</tr>
<tr>
<td>e$</td>
<td>verb</td>
</tr>
<tr>
<td>$</td>
<td>adj</td>
</tr>
<tr>
<td>ility$</td>
<td>noun</td>
</tr>
<tr>
<td>ation$</td>
<td>noun</td>
</tr>
<tr>
<td>e$</td>
<td>adj</td>
</tr>
<tr>
<td>ce$</td>
<td>noin</td>
</tr>
<tr>
<td>iness$</td>
<td>noun</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ious$</td>
<td>adj</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sm$</td>
<td>noun</td>
</tr>
<tr>
<td>ty$</td>
<td>noun</td>
</tr>
<tr>
<td>ty$</td>
<td>noun</td>
</tr>
</tbody>
</table>

1. Duplicates: remove
2. Low frequency: remove
New SD-Rules (nomD)

- Derive SD-Rules from known SD-pairs:
  - nomD (14,638 SD)
  - location | noun | locate | verb => ion$ | noun | e$ | verb
  - Identified 513 possible rules

<table>
<thead>
<tr>
<th>Identified Rules</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>adj</td>
</tr>
<tr>
<td>e$</td>
<td>verb</td>
</tr>
<tr>
<td>$</td>
<td>adj</td>
</tr>
<tr>
<td>ility$</td>
<td>noun</td>
</tr>
<tr>
<td>ation$</td>
<td>noun</td>
</tr>
<tr>
<td>e$</td>
<td>adj</td>
</tr>
<tr>
<td>ce$</td>
<td>noun</td>
</tr>
<tr>
<td>iness$</td>
<td>noun</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ious$</td>
<td>adj</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sm$</td>
<td>noun</td>
</tr>
<tr>
<td>ty$</td>
<td>noun</td>
</tr>
<tr>
<td>ty$</td>
<td>noun</td>
</tr>
</tbody>
</table>

- 1. Duplicates: remove
- 2. Low frequency: remove
- 3. Candidates: further analysis
### New SD-Rules (nomD)

- Decompose SD-Rule: e$|verb|ion$|noun (1,740)

<table>
<thead>
<tr>
<th>Child SD-Rules</th>
<th>Example</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>ate$</td>
<td>verb</td>
<td>ation$</td>
</tr>
<tr>
<td>se$</td>
<td>verb</td>
<td>sion$</td>
</tr>
<tr>
<td>ute$</td>
<td>verb</td>
<td>ution$</td>
</tr>
<tr>
<td>ete$</td>
<td>verb</td>
<td>etion$</td>
</tr>
<tr>
<td>ote$</td>
<td>verb</td>
<td>otion$</td>
</tr>
<tr>
<td>ite$</td>
<td>verb</td>
<td>ition$</td>
</tr>
<tr>
<td>ce$</td>
<td>verb</td>
<td>cion$</td>
</tr>
</tbody>
</table>
Process: Add a New SD-Rules

1. Propose a new SD-Rule candidate (nomD, Facts, or linguistics)
2. Check if duplicated
3. Check frequency in Lexicon
4. Further analysis Parents-Child SD-Rules
5. Above optimization of accuracy & coverage
6. Add to the SD-Rules
# 10 New SD Rules

- **nomD**: high frequency candidates:

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Total</th>
<th>Yes</th>
<th>No</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.81%</td>
<td>536</td>
<td>535</td>
<td>1</td>
<td>iness$</td>
</tr>
<tr>
<td>97.70%</td>
<td>651</td>
<td>636</td>
<td>15</td>
<td>ed$</td>
</tr>
<tr>
<td>93.31%</td>
<td>553</td>
<td>516</td>
<td>37</td>
<td>$</td>
</tr>
<tr>
<td>91.57%</td>
<td>510</td>
<td>467</td>
<td>43</td>
<td>$</td>
</tr>
</tbody>
</table>

- **Facts**: high frequency candidates:

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Total</th>
<th>Yes</th>
<th>No</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.95%</td>
<td>1931</td>
<td>1930</td>
<td>1</td>
<td>ic$</td>
</tr>
<tr>
<td>99.64%</td>
<td>559</td>
<td>557</td>
<td>2</td>
<td>$</td>
</tr>
<tr>
<td>95.63%</td>
<td>504</td>
<td>482</td>
<td>22</td>
<td>ist$</td>
</tr>
<tr>
<td>91.70%</td>
<td>277</td>
<td>254</td>
<td>23</td>
<td>ic$</td>
</tr>
<tr>
<td>89.93%</td>
<td>139</td>
<td>125</td>
<td>14</td>
<td>$</td>
</tr>
<tr>
<td>1.84%</td>
<td>381</td>
<td>7</td>
<td>374</td>
<td>$</td>
</tr>
</tbody>
</table>
# 107 SD Rules

- 107 SD-Rules on Lexicon (2013)

<table>
<thead>
<tr>
<th>No.</th>
<th>Accuracy</th>
<th>Total</th>
<th>yes</th>
<th>no</th>
<th>SD-Rule</th>
<th>Accum Total</th>
<th>Accum Yes</th>
<th>System Accuracy</th>
<th>System Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100.00%</td>
<td>2723</td>
<td>2723</td>
<td>0</td>
<td>$</td>
<td>adj</td>
<td>ness$</td>
<td>noun</td>
<td>2723</td>
</tr>
<tr>
<td>2</td>
<td>100.00%</td>
<td>1278</td>
<td>1278</td>
<td>0</td>
<td>ability$</td>
<td>noun</td>
<td>able$</td>
<td>adj</td>
<td>4001</td>
</tr>
<tr>
<td>3</td>
<td>100.00%</td>
<td>1215</td>
<td>1215</td>
<td>0</td>
<td>ization$</td>
<td>noun</td>
<td>ize$</td>
<td>verb</td>
<td>5216</td>
</tr>
<tr>
<td>4</td>
<td>100.00%</td>
<td>366</td>
<td>366</td>
<td>0</td>
<td>osis$</td>
<td>noun</td>
<td>otic$</td>
<td>5582</td>
<td>5582</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>79</td>
<td>72.09%</td>
<td>43</td>
<td>31</td>
<td>12</td>
<td>ious$</td>
<td>adj</td>
<td>y$</td>
<td>noun</td>
<td>45784</td>
</tr>
<tr>
<td>80</td>
<td>64.22%</td>
<td>109</td>
<td>70</td>
<td>39</td>
<td>ant$</td>
<td>adj</td>
<td>ate$</td>
<td>verb</td>
<td>45893</td>
</tr>
<tr>
<td>81</td>
<td>62.65%</td>
<td>332</td>
<td>208</td>
<td>124</td>
<td>$</td>
<td>noun</td>
<td>ist$</td>
<td>noun</td>
<td>46225</td>
</tr>
<tr>
<td>82</td>
<td>60.66%</td>
<td>183</td>
<td>111</td>
<td>72</td>
<td>ar$</td>
<td>adj</td>
<td>e$</td>
<td>noun</td>
<td>46408</td>
</tr>
<tr>
<td>83</td>
<td>58.08%</td>
<td>582</td>
<td>338</td>
<td>244</td>
<td>al$</td>
<td>adj</td>
<td>e$</td>
<td>noun</td>
<td>46990</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>104</td>
<td>0.37%</td>
<td>273</td>
<td>1</td>
<td>272</td>
<td>a$</td>
<td>noun</td>
<td>an$</td>
<td>noun</td>
<td>52195</td>
</tr>
<tr>
<td>105</td>
<td>0.00%</td>
<td>358</td>
<td>0</td>
<td>358</td>
<td>gram$</td>
<td>noun</td>
<td>graphy$</td>
<td>noun</td>
<td>52553</td>
</tr>
<tr>
<td>106</td>
<td>0.00%</td>
<td>228</td>
<td>0</td>
<td>228</td>
<td>gram$</td>
<td>noun</td>
<td>graphic$</td>
<td>adj</td>
<td>52781</td>
</tr>
<tr>
<td>107</td>
<td>0.00%</td>
<td>57</td>
<td>0</td>
<td>57</td>
<td>$</td>
<td>verb</td>
<td>ably$</td>
<td>adv</td>
<td>52838</td>
</tr>
</tbody>
</table>
Top 80/107 SD Rules

- Accuracy rate: 95.39%
- Coverage rate: 95.59%
- Used to predict derivations in general English
Lvg – Derivations Enhancement

• 2011-
  ▪ SuffixD
    o Facts: 4,559 derivational pairs
    o Rules: 97 SD-Rules
      Use exceptions & heuristic rules to increase accuracy

• 2012:
  ▪ Facts: Added zeroD, prefixD and nomD (89,950)

• 2013:
  ▪ Facts: Added suffixD (121,078)
  ▪ Algorithm:
    o Update source restriction (-kd)
    o Added negation option (-kdn)
    o Added type option (-kdt)
Conclusion

• Better coverage:
  ▪ Facts: cover all SD-pairs known to Lexicon
    | 2011 Lvg | 2012 Lvg | 2013 Lvg |
    |---------|---------|---------|
    | 4,559   | 89,950  | 121,078 |
  ▪ SD-Rules: covers 95.59% of SD-Pairs (in Lexicon)

• Better accuracy rate:
  ▪ Mainly rely on facts: virtually 100% accurate
  ▪ SD-Rules (not in Lexicon): above 95%
Future Work

• Annual routine update with lexicon release
• Enhancement:
  ▪ prefixD: work on more prefixes
  ▪ suffixD: work on more candidate SD-Rules

• More research on SuffixD:
  ▪ Parents-Child Rules:
    o Meet the requirements of accuracy rate and coverage
    o Less SD-Rules
  ▪ Lexicon is representable subset of general English?
Questions