The SPECIALIST NLP Tools Suffix Derivations

Dr. Chris J. Lu

The Lexical Systems Group

NLM. LHNCBC. CGSB

Dec., 2012

Table of Contents

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

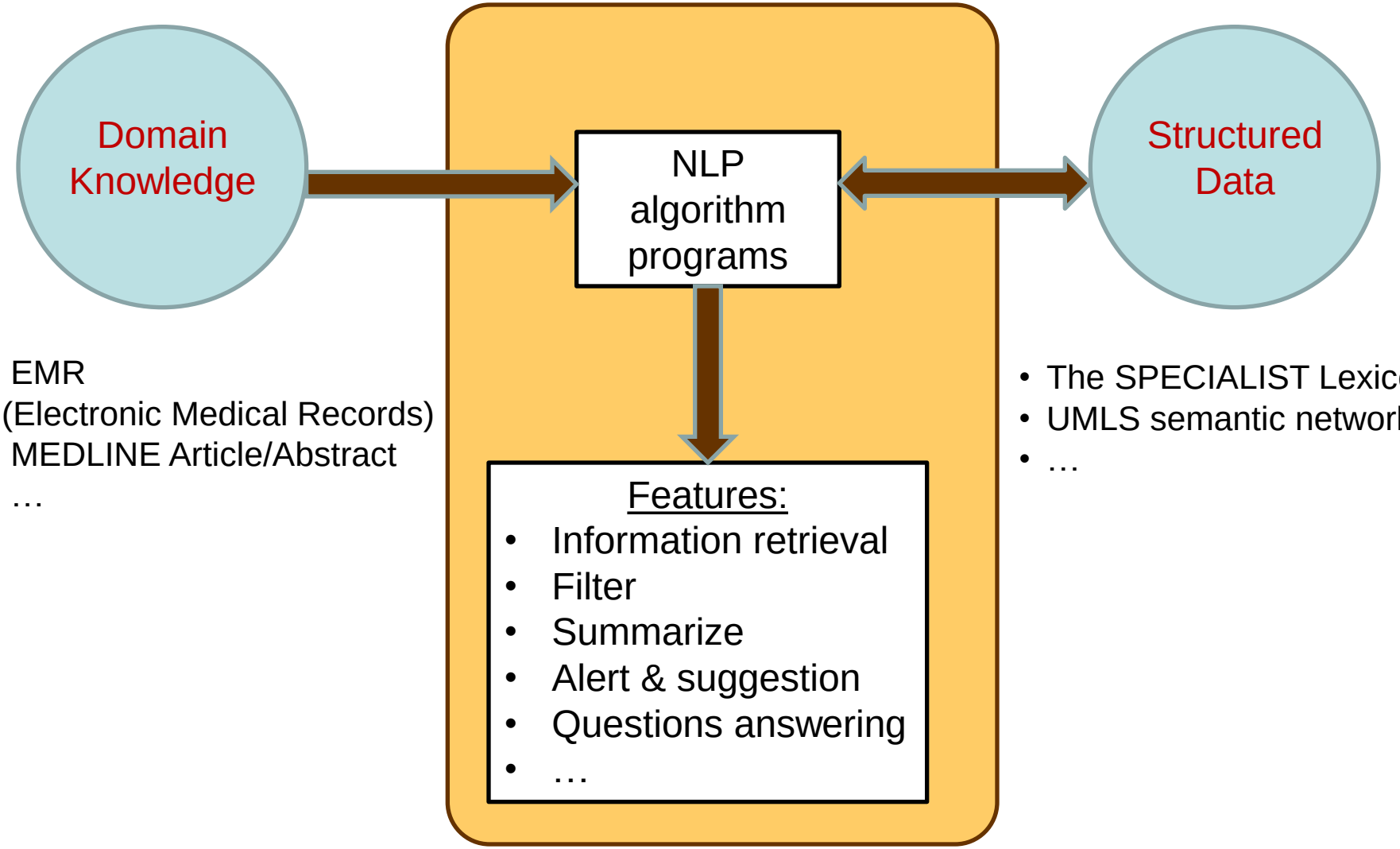
Questions
Symptoms

NLP System

Features:

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

NLP System



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

- The SPECIALIST Lexicon
- UMLS semantic network
- …

NLP Core Tasks

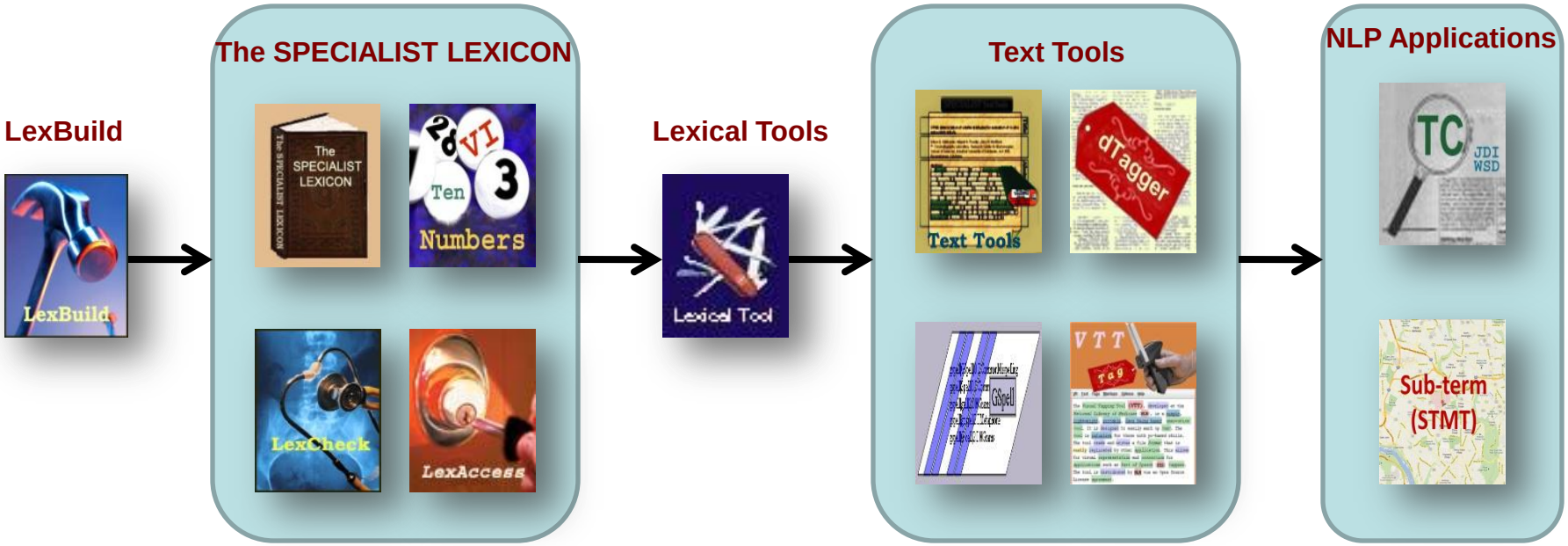
Example: Information retrieval (search engine)

- 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

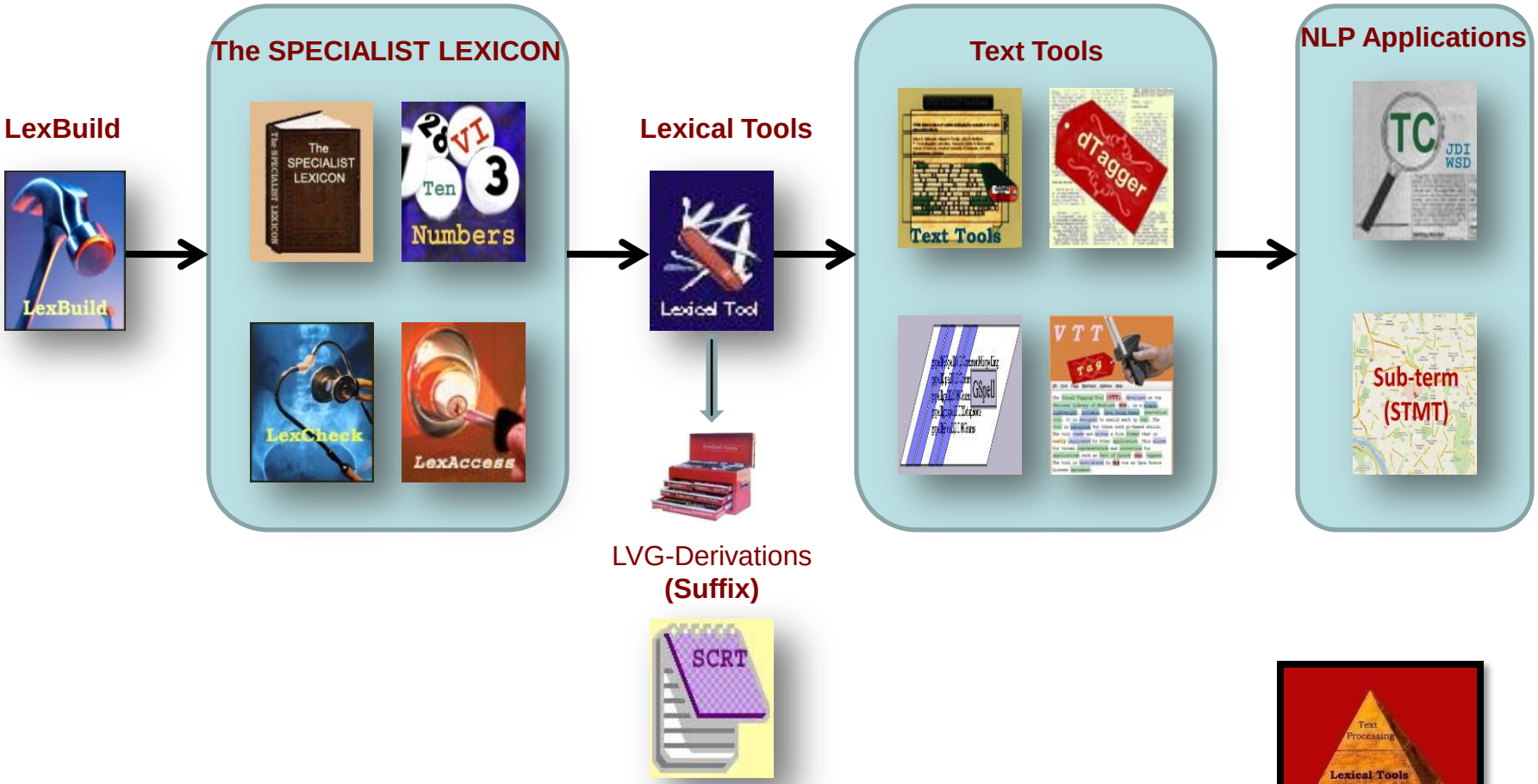
- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

The SPECIALIST NLP Tools



- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

The SPECIALIST NLP Tools



- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov

Lexical Tools - 2013

- Lexical Tools include 7 tools:
  - lvg (Lexical Variants Generation)
    - 62 flow components
    - 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

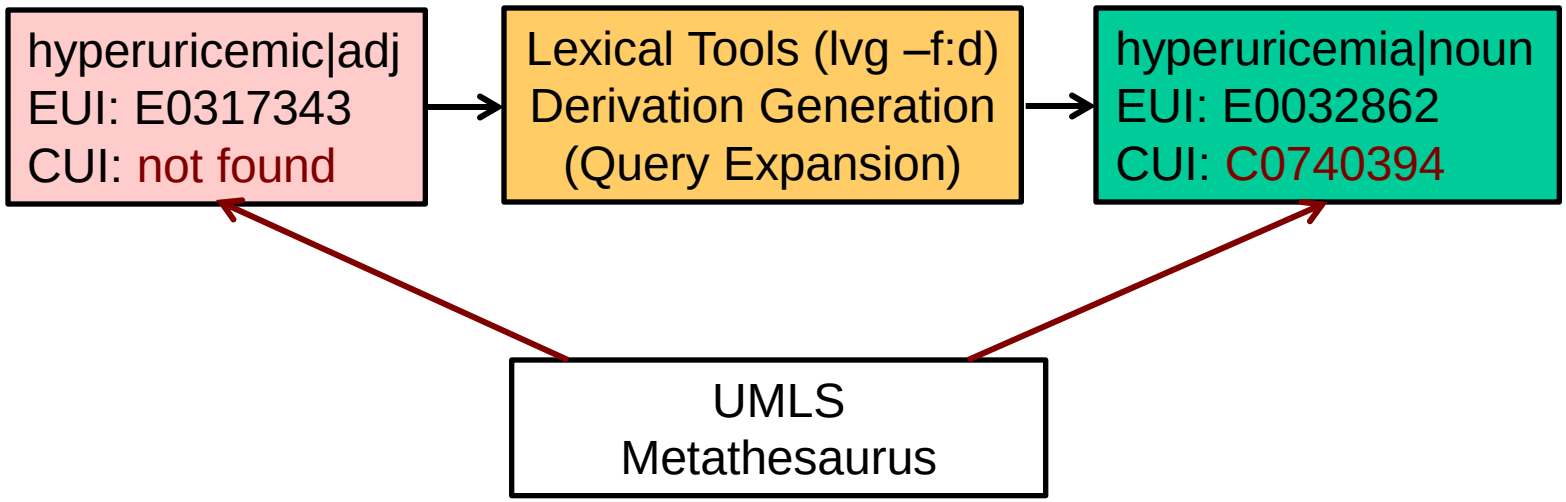
- **shell> lvg –f:d –f:R –SC –SI –p**
  - Please input a term (type "Ctl-d" to quit) >
  hyperuricemic

  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

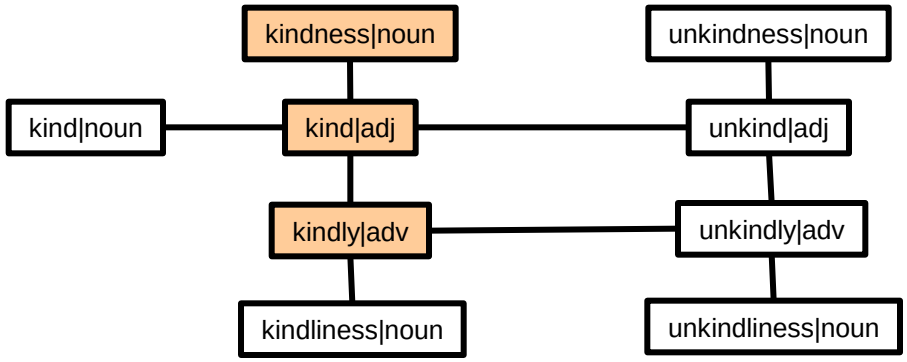
- Added to Lexicon/Lexical Tools with a systematic method in 2012 release

- 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|kindnly|adv



Derivational Flow – SD

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

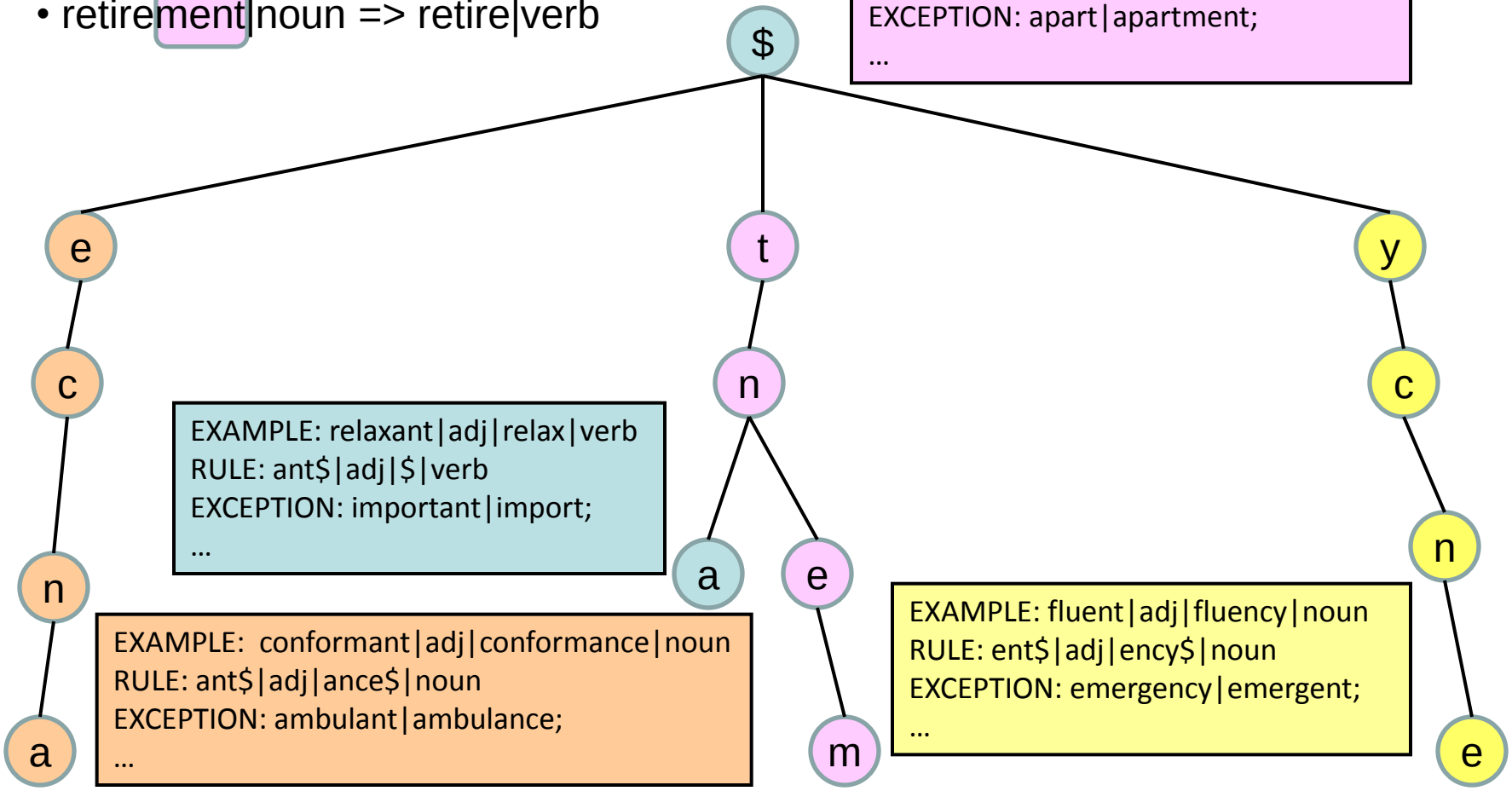
Base 1	Category 1	Base 2	Category 2
…	…	…	…
treatment	noun	treat	noun
…	…	…	…

- 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



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.

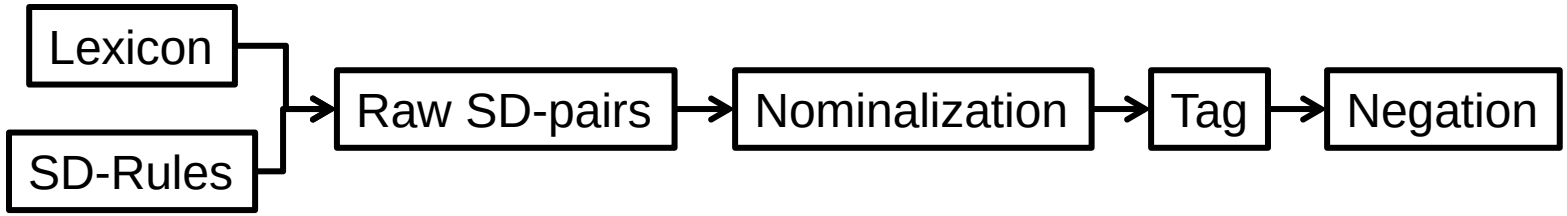
  - Facts (virtually 100% accurate)
    - focus on higher coverage
    - include more derivational pairs known to Lexicon
    - grow proportionally with Lexicon annually

  - 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%)

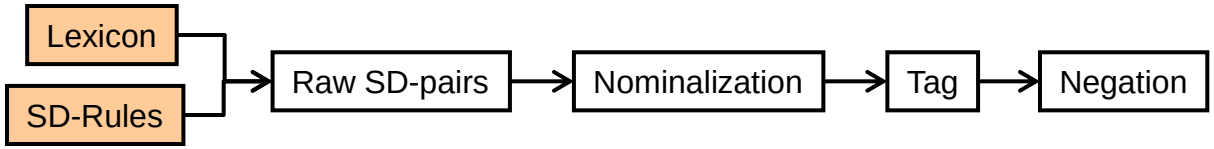
SD - Facts

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

- Process:



SD – Facts



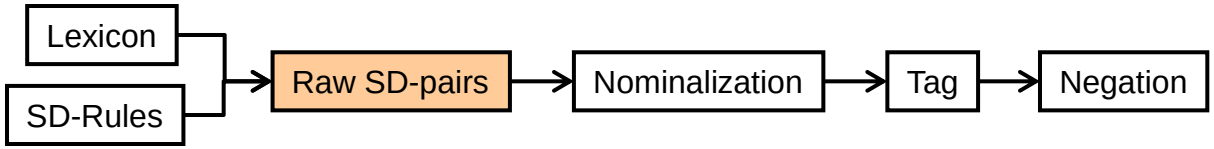
SD –Rules (97):

…
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
…

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
…

Raw SD-Pairs (2,025):

…
compensate|verb|E0018113|compensation|noun|E0018118
…
locate|verb|E0037939location|noun|1|E0037940
…
state|verb|E0057695|station|noun|E0057711
…

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

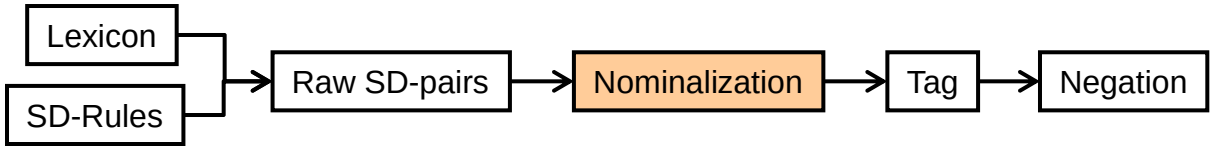
```
{base=locate
entry=E0037939
        cat=verb
        variants=reg
        tran=np
        link=advbl
        cplxtran=np,advbl
        nominalization=location|noun|E0037940
}
```

```
{base=location
entry=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

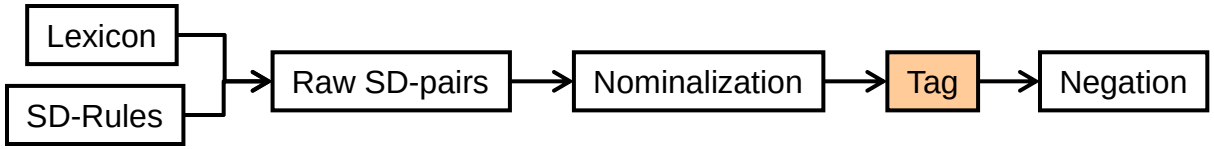


- Automatically tag valid nomD as valid suffixD

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

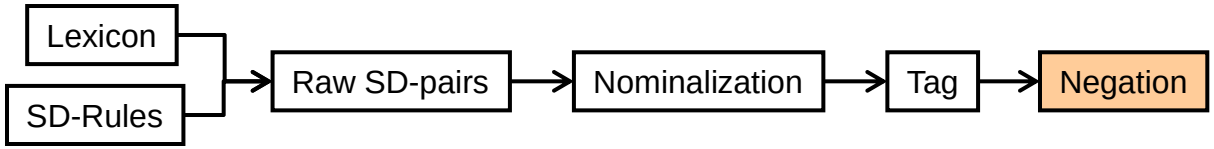


- Manually tag the rest by linguists

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:
    - Class N (10): anti-, contra-, counter-, dis-, il-, im-, ir-, mis-, non-, un-
    - Class O (129): abs-, af-, Afro-, ambi-, etc.
    - Class B (6): a-, an-, de-, dys-, in-, under-
  - suffixD:
    - -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
    - delimitate|verb|E0021381|delimitation|noun|E0021382|no
    - legate|verb|E0540056|legation|noun|E0593456|no
    - rate|verb|E0052016|ration|noun|E0052025|no
    - predate|verb|E0068010|predation|noun|E0068011|no
    - state|verb|E0057695|station|noun|E0057711|no
  - Accuracy rate: 99.75% (= 2020/2025)

SD Facts - Results

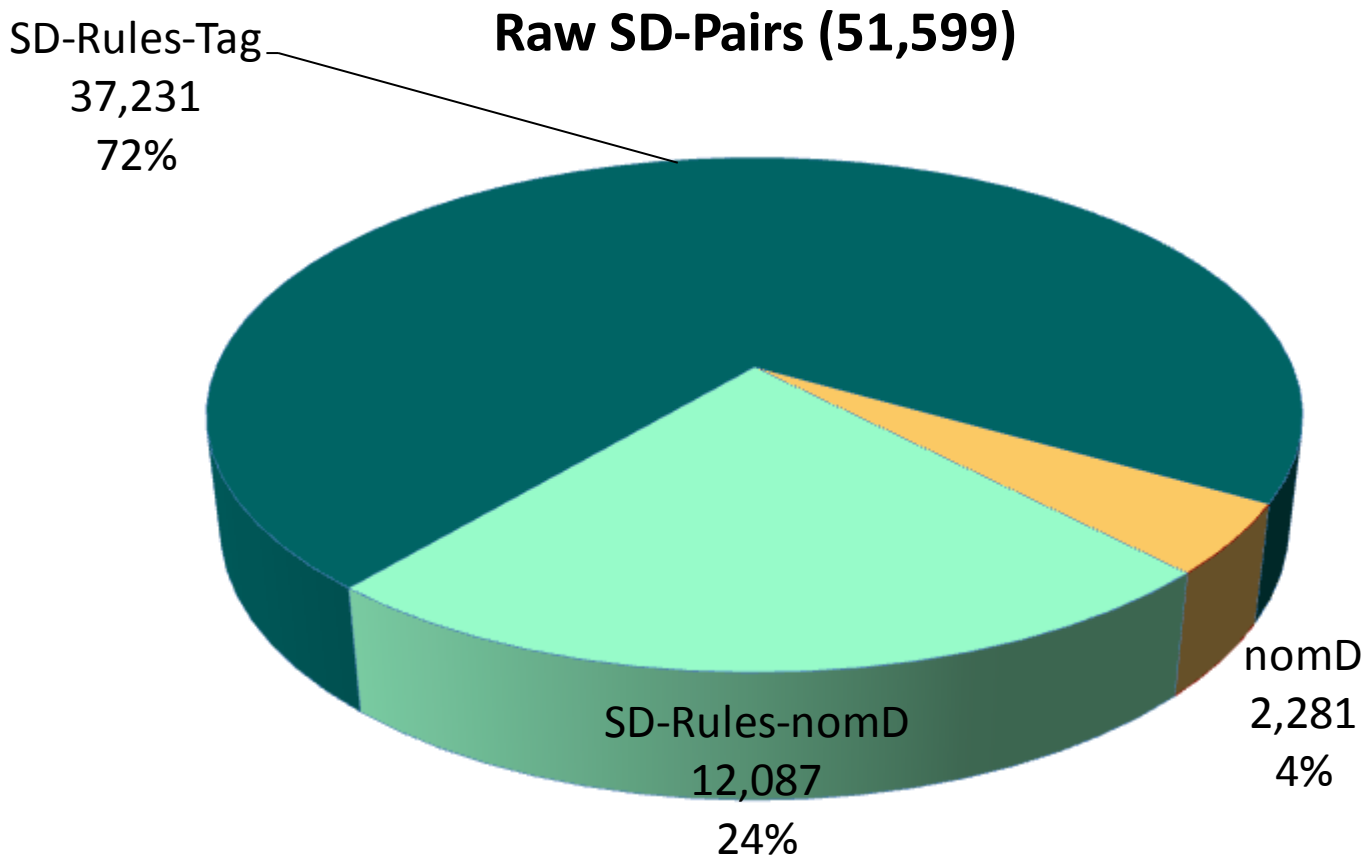
- Apply all candidate SD-Rules on Lexicon (2013)

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

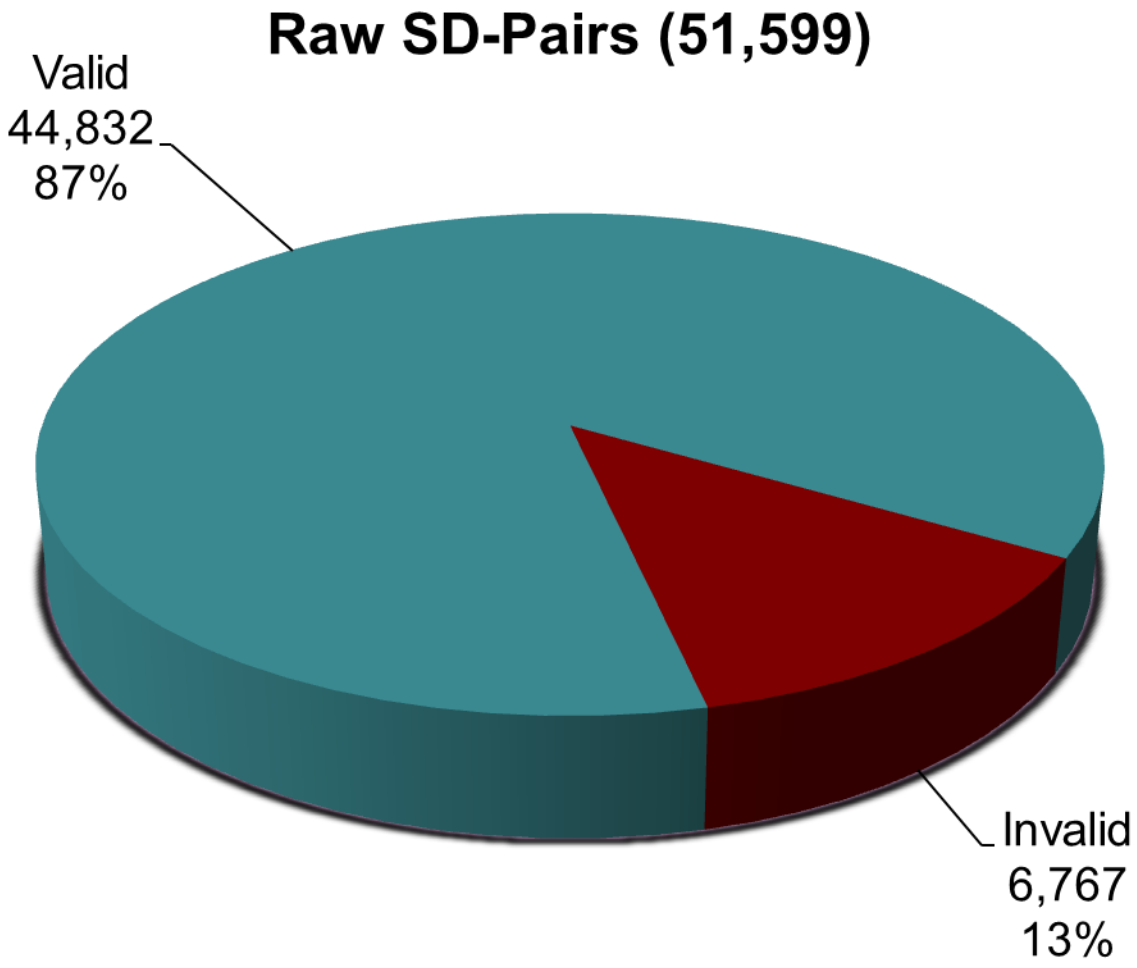
- Total raw SD-Pairs: 51,599 (Tagged stats)
  - Valid: 44,832 (86.89%)
    - Class N: 564 (1.26%)
    - Class O: 44,268 (98.74%)
  - Invalid: 6,767 (13.11%)

SD Facts

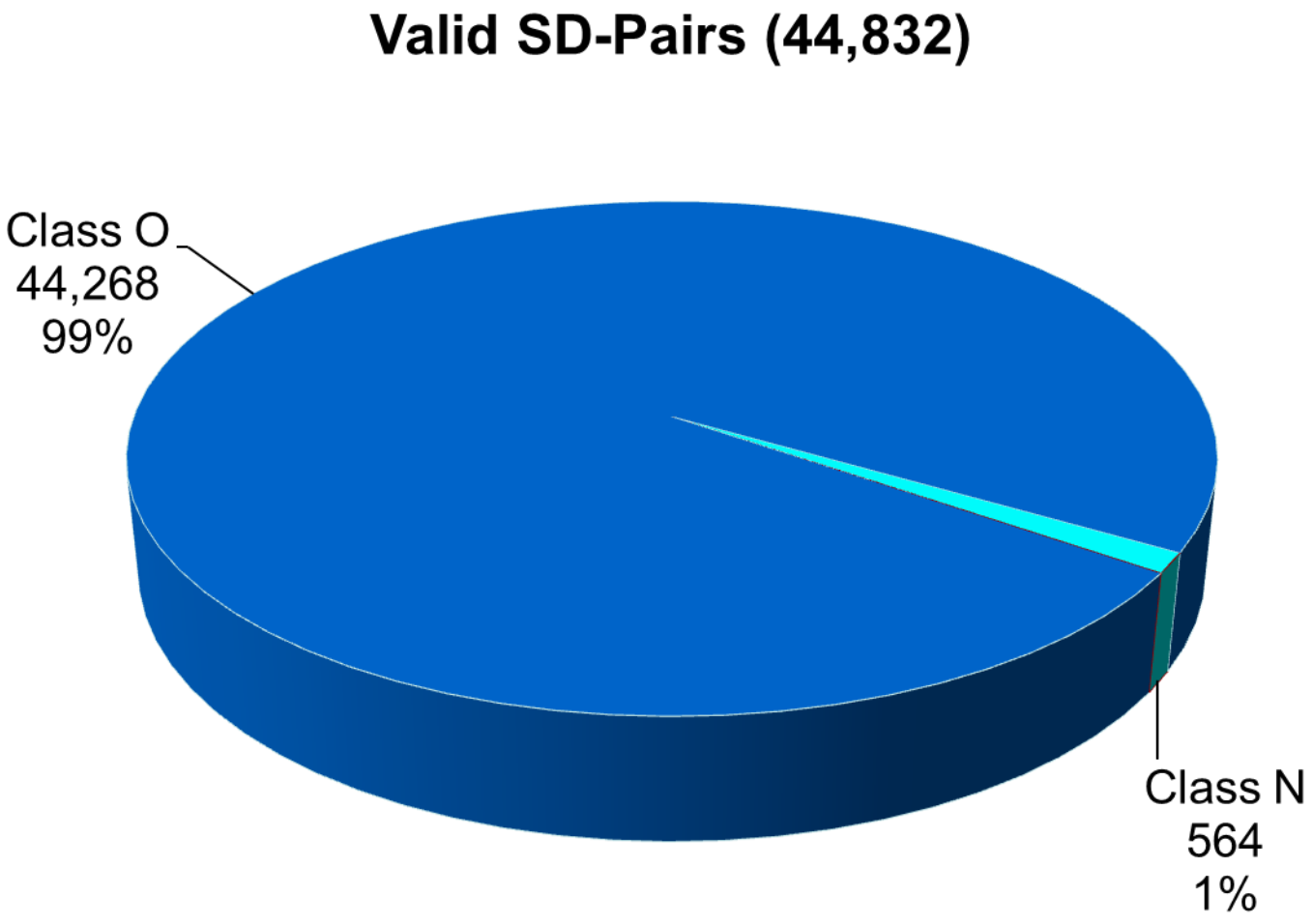
- 28% are auto-tagged
- SD-Rules covers 96% of SD-Pairs known in Lexicon



SD Facts – Valid/Invalid SD-Pairs



SD Facts – Negation



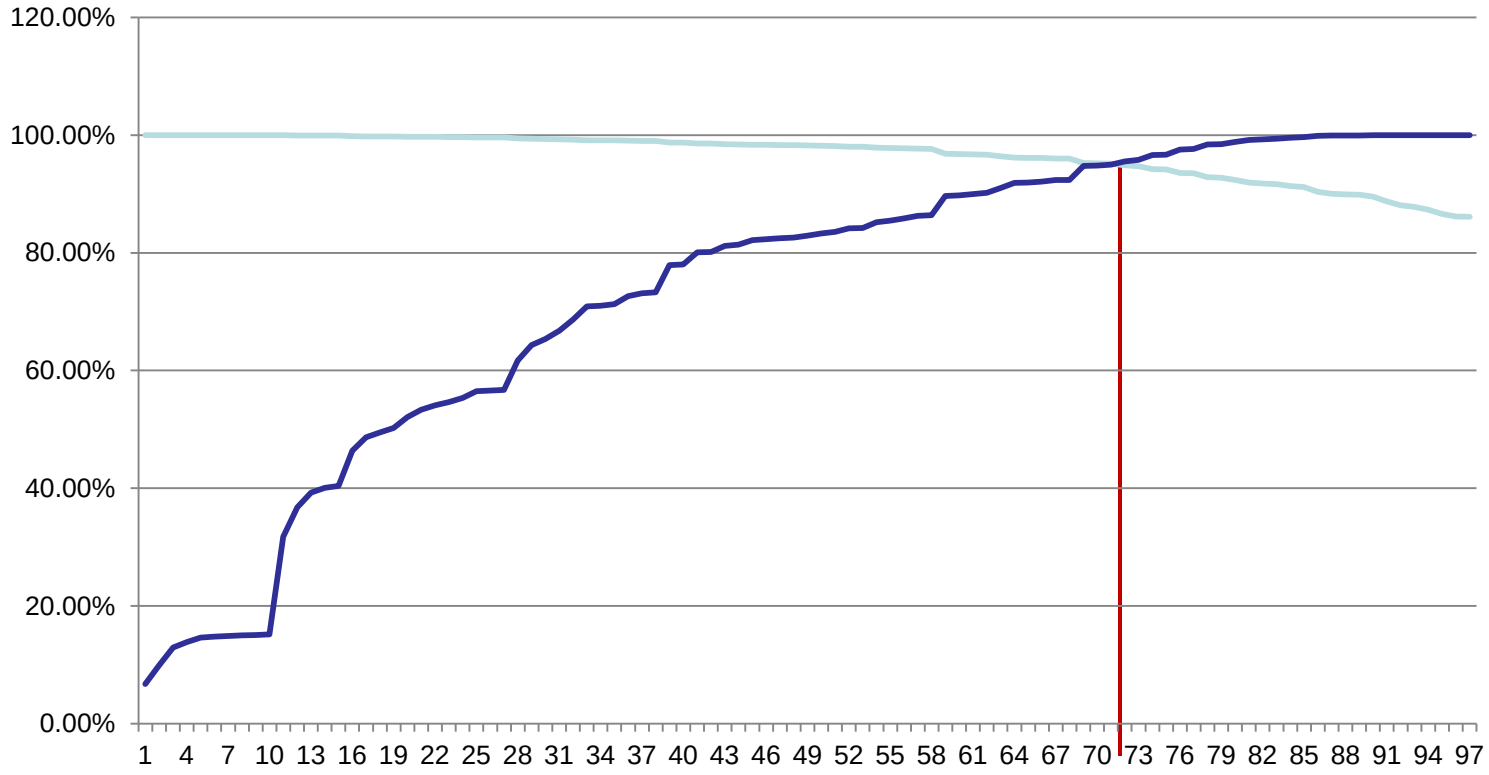
97 SD Rules

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

No.	Accuracy	Total	yes	no	SD-Rule	Accum Total	Accum Yes	System Accuracy	System Coverage
1	100.00%	2723	2723	0	$\|adj\|ness$|noun	2723	2723	100.00%	6.76%
2	100.00%	1278	1278	0	ability$\|noun\|able$|adj	4001	4001	100.00%	9.93%
3	100.00%	1215	1215	0	ization$\|noun\|ize$|verb	5216	5216	100.00%	12.95%
4	100.00%	366	366	0	osis$\|noun\|otic$	5582	5582	100.00%	13.85%
5	100.00%	326	326	0	le$\|adj\|ly$|adv	5908	5908	100.00%	14.66%
	…	…	…	…	…	…	…	…	…
71	72.09%	43	31	12	ious$\|adj\|y$|noun	40124	38205	95.22%	94.83%
72	64.22%	109	70	39	ant$\|adj\|ate$|verb	40233	38275	**95.13%**	**95.00%**
73	62.65%	332	208	124	$\|noun\|ist$|noun	40565	38483	94.87%	95.52%
	…	…	…	…	…	…	…	…	…
93	0.74%	136	1	135	ia$\|noun\|ian$|noun	45881	40288	87.81%	100.00%
94	0.37%	273	1	272	a$\|noun\|an$|noun	46154	40289	87.29%	100.00%
95	0.00%	358	0	358	gram$\|noun\|graphy$|noun	46512	40289	86.62%	100.00%
96	0.00%	228	0	228	gram$\|noun\|graphic$|adj	46740	40289	86.20%	100.00%
97	0.00%	57	0	57	$\|verb\|ably$|adv	46797	40289	86.09%	100.00%

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)
  - **locat**ion|noun|**locat**e|verb => **ion$|noun|e$|verb**
  - Identified 513 possible SD-Rules

Identified Rules	Counts
$\|adj\|ness$|noun	2,489
e$\|verb\|ion$|noun	1,740
$\|adj\|ity$|noun	1,635
ility$\|noun\|le$|adj	1,295
ation$\|noun\|e$|verb	1,164
e$\|adj\|ity$|noun	604
ce$\|noin\|t$|adj	522
iness$\|noun\|y$|adj	501
…	…
ious$\|adj\|y$|noun	10
…	…
sm$\|noun\|ve$|adj	1
ty$\|noun\|ve$|adj	1
ty$|noun|ze|verb	1

New SD-Rules (nomD)

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

Identified Rules	Counts
$\|adj\|ness$|noun	2,489
e$\|verb\|ion$|noun	1,740
$\|adj\|ity$|noun	1,635
ility$\|noun\|le$|adj	1,295
ation$\|noun\|e$|verb	1,164
e$\|adj\|ity$|noun	604
ce$\|noin\|t$|adj	522
iness$\|noun\|y$|adj	501
…	…
ious$\|adj\|y$|noun	10
…	…
sm$\|noun\|ve$|adj	1
ty$\|noun\|ve$|adj	1
ty$|noun|ze|verb	1

→ 1. Duplicates: remove

New SD-Rules (nomD)

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

Identified Rules	Counts
$\|adj\|ness$|noun	2,489
e$\|verb\|ion$|noun	1,740
$\|adj\|ity$|noun	1,635
ility$\|noun\|le$|adj	1,295
ation$\|noun\|e$|verb	1,164
e$\|adj\|ity$|noun	604
ce$\|noin\|t$|adj	522
iness$\|noun\|y$|adj	501
…	…
ious$\|adj\|y$|noun	10
…	…
sm$\|noun\|ve$|adj	1
ty$\|noun\|ve$|adj	1
ty$|noun|ze|verb	1

→ 1. Duplicates: remove

→ 2. Low frequency: remove

New SD-Rules (nomD)

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

Identified Rules	Counts
$\|adj\|ness$|noun	2,489
e$\|verb\|ion$|noun	1,740
$\|adj\|ity$|noun	1,635
ility$\|noun\|le$|adj	1,295
ation$\|noun\|e$|verb	1,164
e$\|adj\|ity$|noun	604
ce$\|noin\|t$|adj	522
iness$\|noun\|y$|adj	501
…	…
ious$\|adj\|y$|noun	10
…	…
sm$\|noun\|ve$|adj	1
ty$\|noun\|ve$|adj	1
ty$|noun|ze|verb	1

→ 3. Candidates: further analysis

→ 1. Duplicates: remove

→ 2. Low frequency: remove

New SD-Rules (nomD)

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

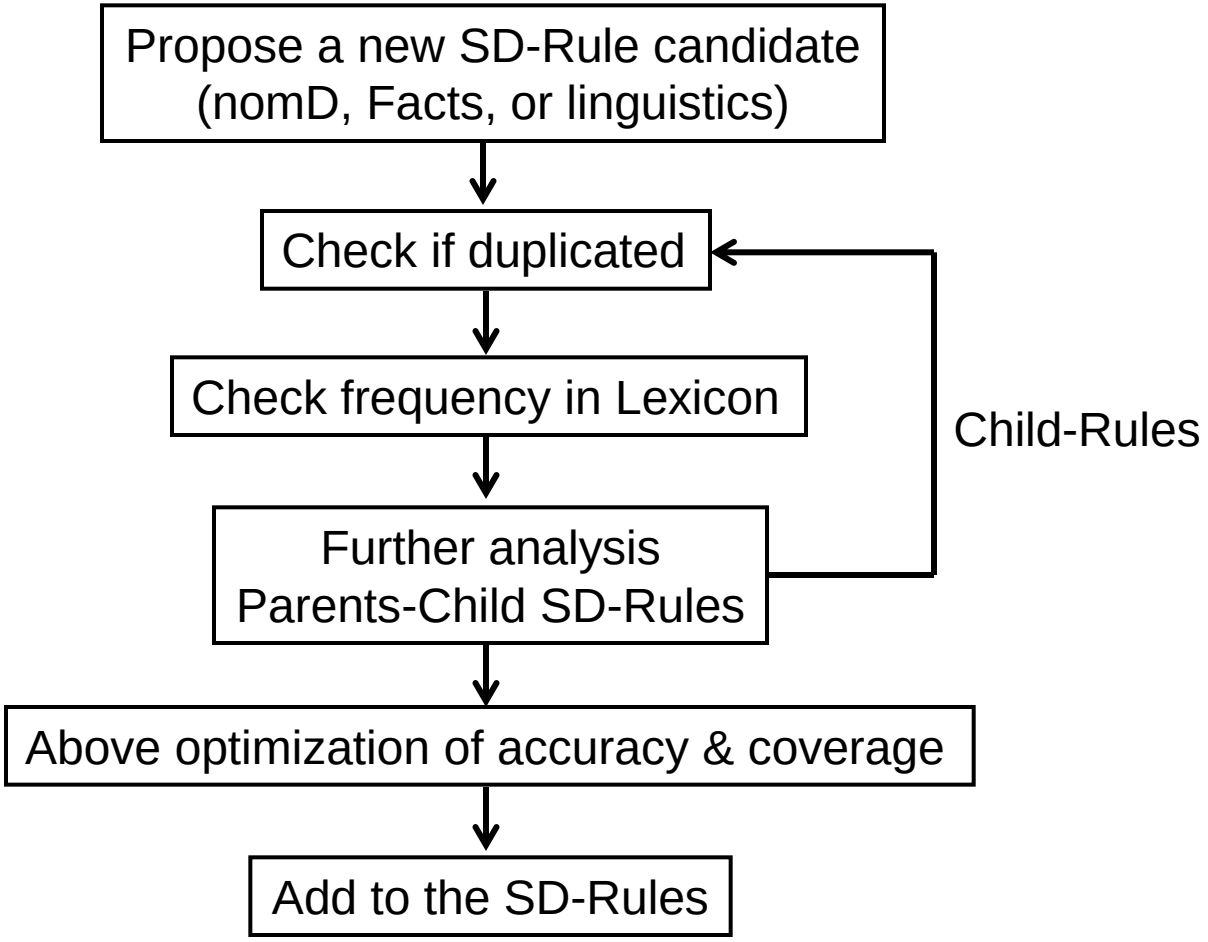
Child SD-Rules	Example	Counts
ate$\|verb\|ation$|noun	locate|verb|location|noun	1,587
se$\|verb\|sion$|noun	tense|verb|tension|noun	80
ute$\|verb\|ution$|noun	delute|verb|delution|noun	39
ete$\|verb\|etion$|noun	complete|verb|completion|noun	22
ote$\|verb\|otion$|noun	devote|verb|devotion|noun	6
ite$\|verb\|ition$|noun	ignite|verb|ignition|noun	5
ce$\|verb\|cion$|noun	coerce|verb|coercion|noun	1

→ duplicates

→ candidates

→ low frequency

Process: Add a New SD-Rules



10 New SD Rules

- nomD: high frequency candidates:

Accuracy	Total	Yes	No	Rules
99.81%	536	535	1	iness$\|noun\|y$|adj
97.70%	651	636	15	ed$\|adj\|ion$|noun
93.31%	553	516	37	$\|verb\|ion$|noun
91.57%	510	467	43	$\|verb\|ing$|noun

- Facts: high frequency candidates:

Accuracy	Total	Yes	No	Rules
99.95%	1931	1930	1	ic$\|adj\|ically$|adv
99.64%	559	557	2	$\|noun\|less$|adj
95.63%	504	482	22	ist$\|noun\|y$|noun
91.70%	277	254	23	ic$\|adj\|is$|noun
89.93%	139	125	14	$\|noun\|ful$|adj
1.84%	381	7	374	$\|verb\|less$|adj

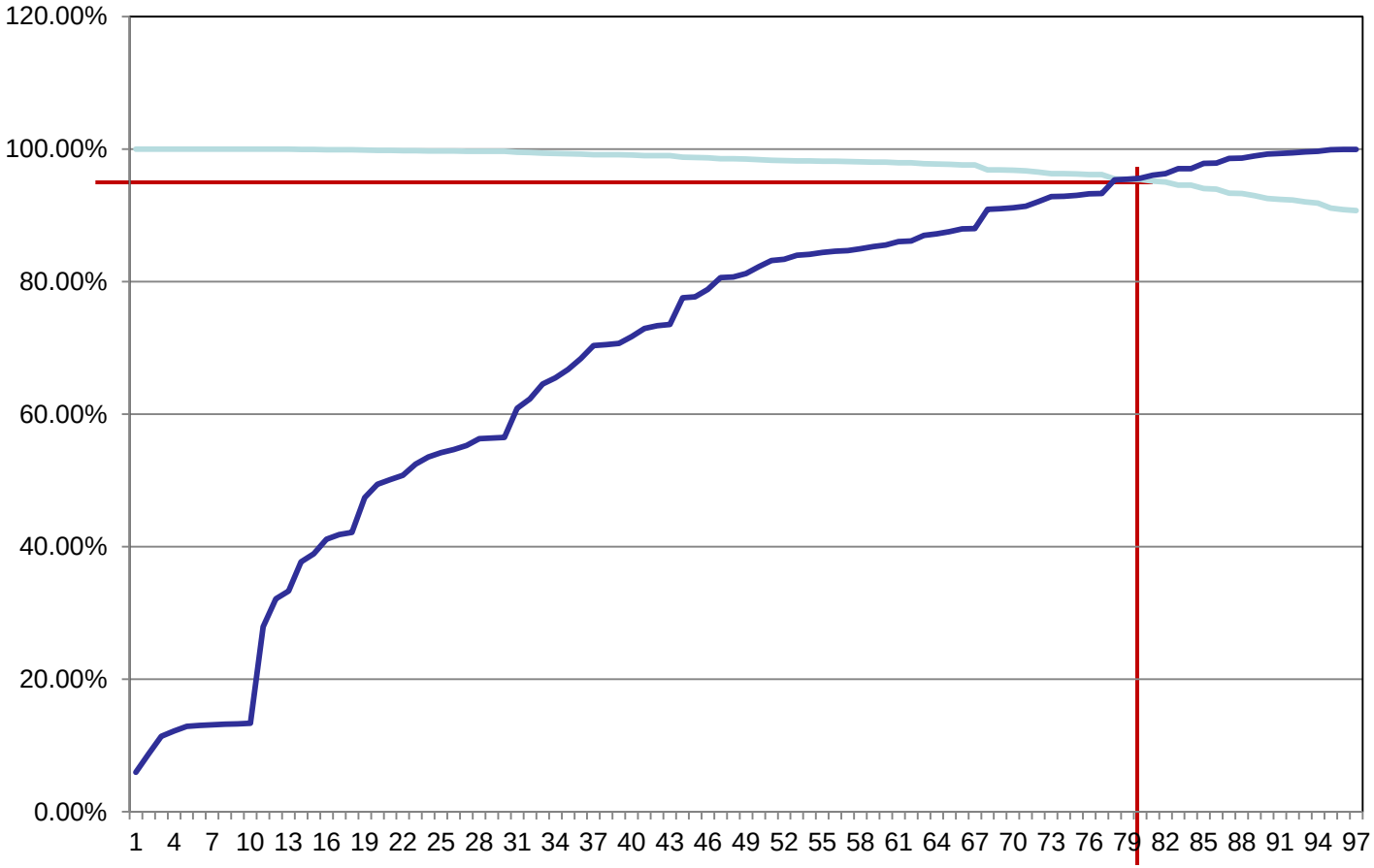
107 SD Rules

- 107 SD-Rules on Lexicon (2013)

No.	Accuracy	Total	yes	no	SD-Rule	Accum Total	Accum Yes	System Accuracy	System Coverage
1	100.00%	2723	2723	0	$\|adj\|ness$|noun	2723	2723	100.00%	5.95%
2	100.00%	1278	1278	0	ability$\|noun\|able$|adj	4001	4001	100.00%	8.74%
3	100.00%	1215	1215	0	ization$\|noun\|ize$|verb	5216	5216	100.00%	11.39%
4	100.00%	366	366	0	osis$\|noun\|otic$	5582	5582	100.00%	12.19%
	…	…	…	…	…	…	…	…	
79	72.09%	43	31	12	ious$\|adj\|y$|noun	45784	43707	95.46%	95.43%
80	**64.22%**	**109**	**70**	**39**	**ant$\|adj\|ate$|verb**	**45893**	**43777**	**95.39%**	**95.59%**
81	62.65%	332	208	124	$\|noun\|ist$|noun	46225	43985	95.15%	96.04%
82	60.66%	183	111	72	ar$\|adj\|e$|noun	46408	44096	95.02%	96.28%
83	58.08%	582	338	244	al$\|adj\|e$|noun	46990	44434	94.56%	97.02%
	…	…	…	…	…	…	…	…	
104	0.37%	273	1	272	a$\|noun\|an$|noun	52195	45798	87.74%	100.00%
105	0.00%	358	0	358	gram$\|noun\|graphy$|noun	52553	45798	87.15%	100.00%
106	0.00%	228	0	228	gram$\|noun\|graphic$|adj	52781	45798	86.77%	100.00%
107	0.00%	57	0	57	$\|verb\|ably$|adv	52838	45798	86.68%	100.00%

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
    - Facts: 4,559 derivational pairs
    - Rules: 97 SD-Rules
      Use exceptions & heuristic rules to increase accuracy

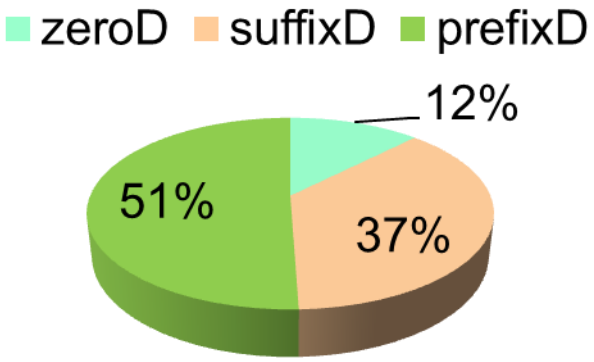
- 2012:
  - Facts: Added zeroD, prefixD and nomD (89,950)

- 2013:
  - Facts: Added suffixD (121,078)
  - Algorithm:
    - Update source restriction (-kd)
    - Added negation option (-kdn)
    - 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:
    - Meet the requirement s of accuracy rate and coverage
    - Less SD-Rules
  - Lexicon is representable subset of general English?

Questions

- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov