A Look inside the Distributionally Similar Terms

Kow Kuroda, Jun'ichi Kazama and Kentaro Torisawa
National Institute of Information and Communications Technology (NICT), Japan

The 2nd International Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)

Large-scale and sharable NLP infrastructures and beyond

August 28, 2010, Beijing International Convention Center

“Distributional” Hypothesis

- Extensive use of distributional similarity derived from the “distributional” hypothesis (Harris 1959) is one of the key concepts of NLP that made it successful.
  - Hindle (1990), Grefenstette (1993), Lee (1997), Lin (1998)
- Reason for its nearly unanimous acceptance is not so much positively motivated, however.
  - If the hypothesis is not accepted, then most of Web-derived data would be intractable.
- Yet ..

Three Questions We Address

- *Can distributional similarity really be equated with semantic similarity?*
  - No agreement seems to be reached as to what count as semantic similarity.
  - And there are several kinds of semantic similarity itself.
- *Even if distributional similarity can be equated with semantic similarity, to what extent is it so?*
- *Even if they can be equated to a large extent, is it valid on a large scale?*
- We address these questions in our study.

Outline

- Method
- Preparing data
- Classification task
- Results
- Summary

Method

General Framework

- Step 1. Select a set of "base" terms B = {b_1, b_1, ..., b_n}
- Step 2. Use a certain similarity measure M (such as Jensen-Shannon divergence) to construct a list of n terms T = [$t_{i,1}$, $t_{i,2}$, ..., $t_{i,j}$, ..., $t_{i,n}$]
  - where $t_{i,j}$ denotes the j^{th} most similar term in T against b_i in B.
- Step 3. Generate $P(k)$, a set of $t_{i,1}$, $t_{i,2}$, ..., $t_{i,k}$ with each paired with b_i. Human raters classify $P(k)$ with reference to a guideline.

Product of Steps 1 and 2

base	b_i's most similar term under M	b_i's 2nd most similar term under M		b_i's kth most similar term under M
b_1	$t_{1,1}$	$t_{1,2}$	…	$t_{1,k}$
b_2	$t_{2,1}$	$t_{2,2}$	…	$t_{2,k}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
b_n	$t_{n,1}$	$t_{n,2}$	…	$t_{n,k}$

Each row represents $T[b_i]$

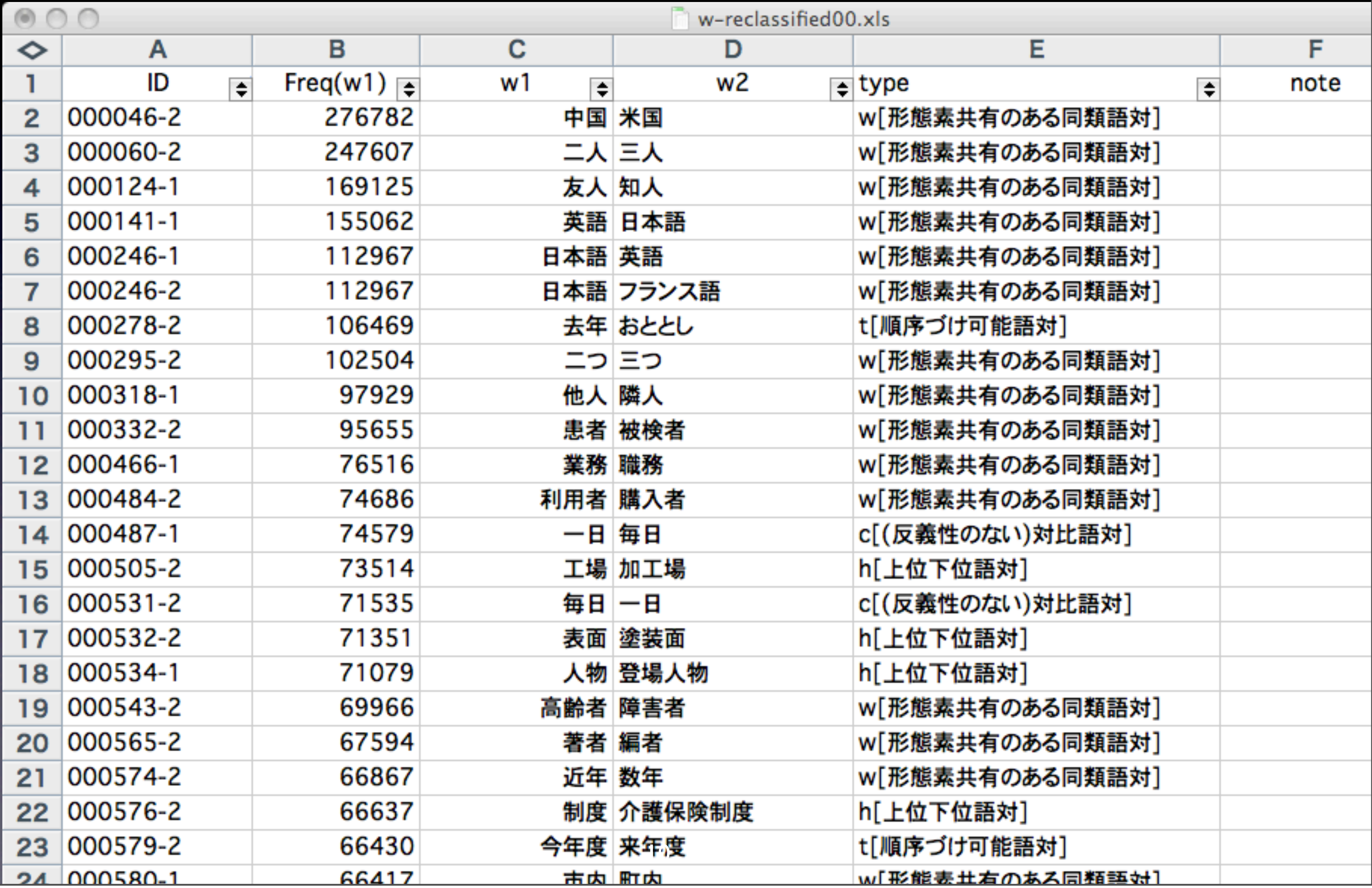
Parameters Considered

- How much for *n*? In other words, how many “bases” to evaluate?
  - In our case, *n* = 150,000
- How much for *k*? In other words, how many similar terms to evaluate?
  - In our case, *k* = 2.
- What similarity metric to use?
  - We used the Jensen-Shannon divergence for *M* under distributional probabilities of <*n*, *p*, *v*> (Kazama et al. 2009)

Characteristics of Step 3

- We classified 300,000 pairs into the 18 finer-grained classes of semantic relation (to be explained).
- But we also applied candidate filtering (to be explained).
- Note
  - In Kazama's clustering data, *n* corresponds to the count rank of dependency relation types. This should be an *indicator* of token frequencies of base terms.

Sample of Data Used in Step 3



	A	B	C	D	E	F
1	ID	Freq(w1)	w1	w2	type	note
2	000046-2	276782	中国	米国	w[形態素共有のある同類語対]	
3	000060-2	247607	二人	三人	w[形態素共有のある同類語対]	
4	000124-1	169125	友人	知人	w[形態素共有のある同類語対]	
5	000141-1	155062	英語	日本語	w[形態素共有のある同類語対]	
6	000246-1	112967	日本語	英語	w[形態素共有のある同類語対]	
7	000246-2	112967	日本語	フランス語	w[形態素共有のある同類語対]	
8	000278-2	106469	去年	おととし	t[順序づけ可能語対]	
9	000295-2	102504	二つ	三つ	w[形態素共有のある同類語対]	
10	000318-1	97929	他人	隣人	w[形態素共有のある同類語対]	
11	000332-2	95655	患者	被検者	w[形態素共有のある同類語対]	
12	000466-1	76516	業務	職務	w[形態素共有のある同類語対]	
13	000484-2	74686	利用者	購入者	w[形態素共有のある同類語対]	
14	000487-1	74579	一日	毎日	c[(反義性のない)対比語対]	
15	000505-2	73514	工場	加工場	h[上位下位語対]	
16	000531-2	71535	毎日	一日	c[(反義性のない)対比語対]	
17	000532-2	71351	表面	塗装面	h[上位下位語対]	
18	000534-1	71079	人物	登場人物	h[上位下位語対]	
19	000543-2	69966	高齢者	障害者	w[形態素共有のある同類語対]	
20	000565-2	67594	著者	編者	w[形態素共有のある同類語対]	
21	000574-2	66867	近年	数年	w[形態素共有のある同類語対]	
22	000576-2	66637	制度	介護保険制度	h[上位下位語対]	
23	000579-2	66430	今年度	来年度	t[順序づけ可能語対]	
24	000580-1	66417	市内	町内	w[形態素共有のある同類語対]	

Preparing Data

10 Most Similar Terms of “ピアノ” (piano)

rank	Japanese (original)	English translation	Score
1	エレクトーン	*Electone*, electric organ	–0.322
2	バイオリン	violin	–0.357
3	ヴァイオリン	violin	–0.358
3	チェロ	cello	–0.358
5	トランペット	trumpet	–0.377
6	三味線	*shamisen*, Japanese 3-string guitar	–0.383
7	サックス	saxophone	–0.390
8	オルガン	organ	–0.392
9	クラリネット	clarinet	–0.394
10	二胡	erh hu	–0.396

10 Most Similar Terms of “チャイコフスキー” (Tchaikovsky)

rank	Japanese (original)	English translation	Score
1	ブラームス	Brahms	–0.152
2	シューマン	Schumann	–0.163
3	メンデルスゾーン	Mendelssohn	–0.166
4	ショスタコーヴィッチ	Shostakovich	–0.178
5	シベリウス	Sibelius	–0.180
6	ハイドン	Haydn	–0.181
6	ヘンデル	Händel	–0.181
8	ラヴェル	Ravel	–0.182
9	シューベルト	Schubert	–0.197
10	ベートーヴェン	Beethoven	–0.190

Terms Excluded from Candidates

- Strings that were judged to fail to have meaning due to segmentation error.
  - An independent task was performed for this.
- Terms begin with Roman digits (i.e., “0”, “1”, ..., “9”)
- Terms ending with 88 derivational morphemes that lead to either POS-change or obscure semantics
- Terms containing more than one occurrence of “ ・ ”
  - “ ・ ” means either disjunction, conjunction or surrogate of “white space” in Japanese.

88 Derivational Morphemes for Candidate Filtering

- Hedge-deriver
  - -など, -等, -たち, -達, -ども, -ら, -以外, -ほか, -他, -くらい, -ぐらい, -まま, -ごと, -ついで, -づつ
- Modalizer
  - -とおり, -あたり, -ぶり, -振り, -あまり, -余り, -ほど, -かわり, -代わり
- Nominalizer
  - -たの, -いの, -うの, -くの, -すの, -つの, -ぬの, -ふの, -むの, -ゆの, -るの, -なの, -んか, -るか, -でか, -っか
- Epithet-deriver
  - -さん, -サン, -ちゃん, -チャン, -さま, -サマ, -様, -くん, -君, -どの, -殿
- Temporalizer or Locationalizer
  - -ばあい, 場合, -ため, -為, -せい, -コト, -こと, -事, -トコロ, -ところ, -所, -処, -とき, -時, -ころ, -ごろ, -頃, -際, -なか, -中, -うえ, -上, -下, -前, -後, -ちかく, -近く, -ほう, -方
- Deriver of other POS-terms
  - -的だ, -的に, -した, -った, -である, -では, -です, -ます

Classification Task

Its design and practice

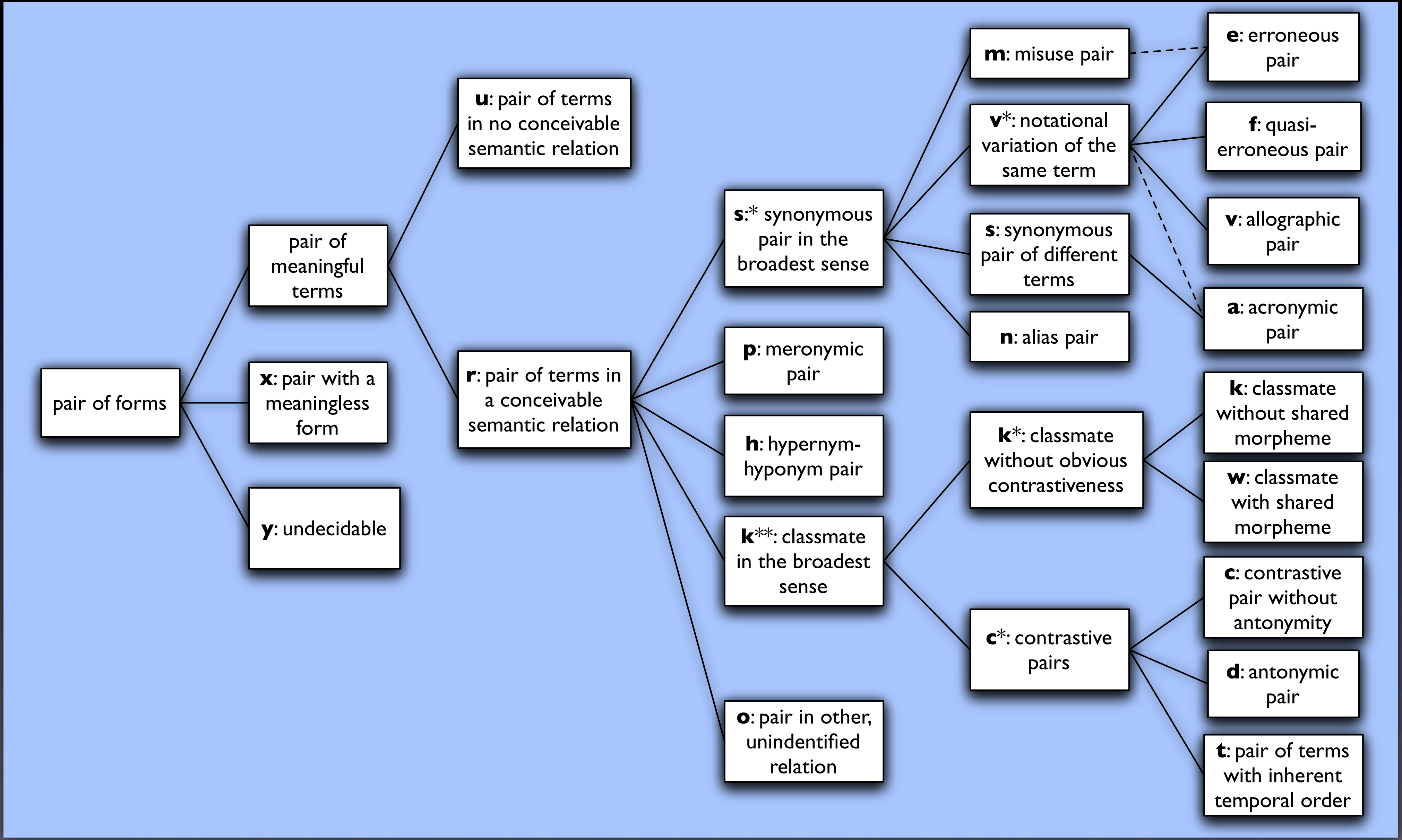
Factoring out "semantic similarity"

- We employed 18 finer-grained classes build on four basic "components" of semantic similarity

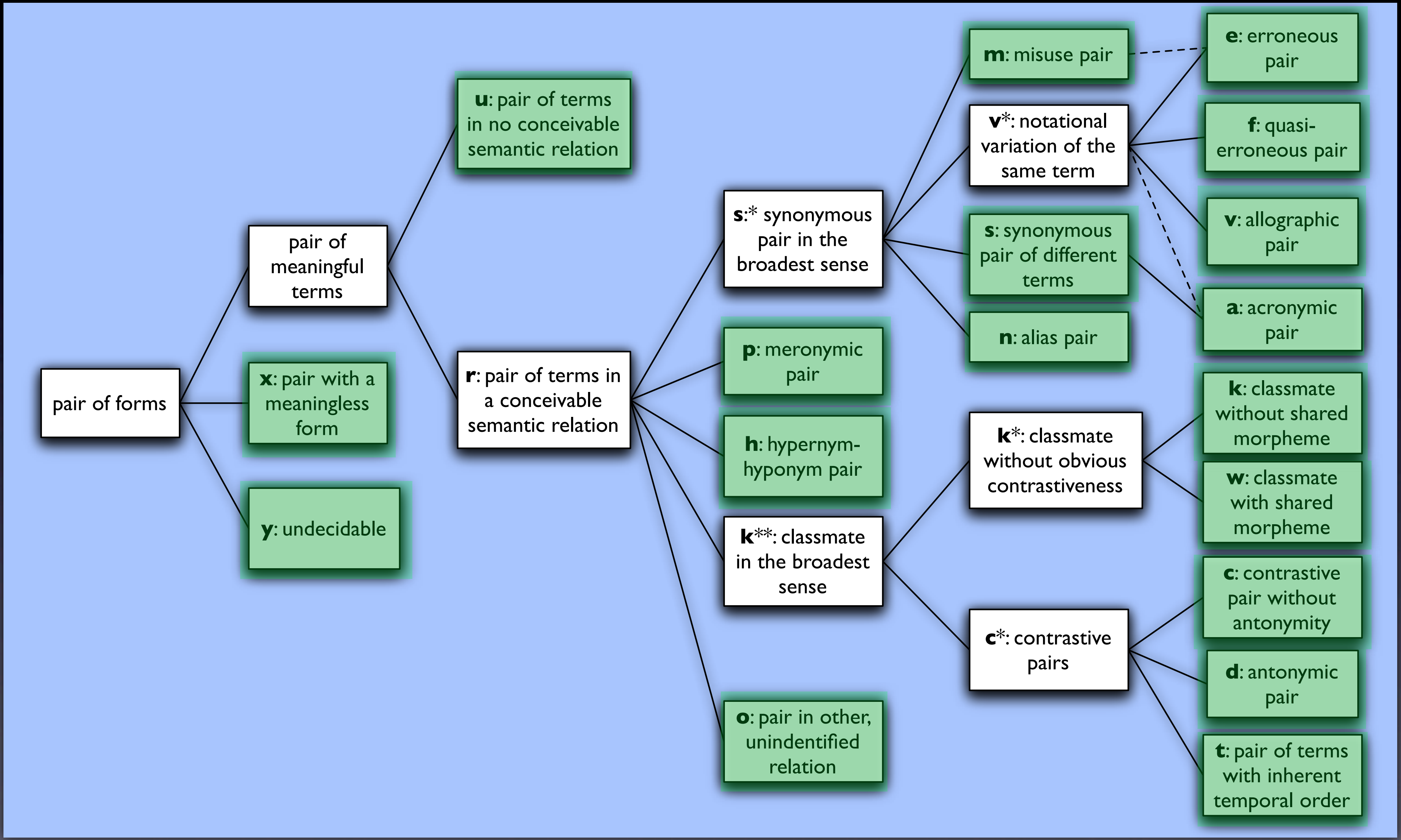
1. synonymic relation
2. hypernym-hyponym relation
3. meronymic relation
4. classmate relation

- They are designed based on research like Fellbaum, ed. (1998), Murphy (2003)

18 Subtypes in the Hierarchy


pair of forms
pair of meaningful terms
x: pair with a meaningless form
y: undecidable
u: pair of terms in no conceivable semantic relation
r: pair of terms in a conceivable semantic relation
s:* synonymous pair in the broadest sense
p: meronymic pair
h: hypernym-hyponym pair
k**: classmate in the broadest sense
o: pair in other, unindentified relation
m: misuse pair
v*: notational variation of the same term
s: synonymous pair of different terms
n: alias pair
k*: classmate without obvious contrastiveness
c*: contrastive pairs
e: erroneous pair
f: quasi-erroneous pair
v: allographic pair
a: acronymic pair
k: classmate without shared morpheme
w: classmate with shared morpheme
c: contrastive pair without antonymity
d: antonymic pair
t: pair of terms with inherent temporal order

18 Subtypes in the Hierarchy



pair of forms

pair of meaningful terms

x: pair with a meaningless form

y: undecidable

u: pair of terms in no conceivable semantic relation

r: pair of terms in a conceivable semantic relation

s:* synonymous pair in the broadest sense

p: meronymic pair

h: hypernym-hyponym pair

k**: classmate in the broadest sense

o: pair in other, unidentified relation

m: misuse pair

v*: notational variation of the same term

s: synonymous pair of different terms

n: alias pair

k*: classmate without obvious contrastiveness

c*: contrastive pairs

e: erroneous pair

f: quasi-erroneous pair

v: allographic pair

a: acronymic pair

k: classmate without shared morpheme

w: classmate with shared morpheme

c: contrastive pair without antonymity

d: antonymic pair

t: pair of terms with inherent temporal order

Characteristics of the Hierarchy

- **s***, **k****, **p**, **h**, and **o** are major divisions and are expected to be mutually exclusive.
  - **s*** has four subtypes: **s**, **m**, **v*** and **n**.
  - **k**** has two subtypes: **k*** and **c***.
  - **k*** has two subtypes: **s*** and **w** differing with presence of a common morpheme.
  - **c*** has three subtypes: **c**, **d** and **t**.
- In the most tolerant condition, {s*, k**, p, h} corresponds to the overall class of semantically similar terms.
- Note that {m, e} or {m, e, f} are only classes in which distributional and semantic similarities do not match up.

Dealing with Label Ambiguity

- But at least in practice, some labels are not mutually exclusive!
  - This does not guarantee the uniqueness of the labels to be assigned.
- To solve this, the following priority was set to choose the most appropriate one:
  - **e**, **f** < **v** < **a** < **n** < **p** < **h** < **s** < **t** < **d** < **c** < **w** < **k** < **m** < **o** < **u** < **x** < **y**
- the leftmost label is the most preferred one.

Examples

1. synonymous [s] pairs

1. (根元, 株元) [both mean *root*]
2. (サポート会員, 協力会員) [(*supporting member, cooperating, member*)]
3. (呼び出し元, 親プロセス) [(*invoker of the process, parent process*)]
4. (相手投手, 相手ピッチャー) (*opposing hurler, opposing pitcher*)
5. (病歴, 既往歴) [(*medical history, anamneses*)]

2. acronymic [a] pairs

1. (DEC, Digital Equipment)
2. (IBM, International Business Machine)
3. (MS 社, Microsoft 社) [(*MS, Inc., Microsoft, Inc.*)]
4. (難関大, 難関大学) [both mean *universities hard to enter*]
5. (配置転換, 配転) [both mean *job displacement*]

3. alias [n] pairs

1.(Steve Jobs, founder of Apple, Inc)

2.(Barak Obama, US President)

3.(侑一郎, うにっ子) [(*Yuichiro*, *Unikko*)]

- *Unikko* seems to be the nickname for a cartoon character.

4.(ノグチ, イサム・ノグチ) [(*Noguchi*, *Isamu Noguchi*)]

4. allographic [v] pairs

1. (Solo, solo) [with or without capitalization]
2. (center, centre), (colour, color) [difference between AmE and BE]
3. (アカスリ, あかすり) [both mean *skin-scrubbing*, pair of katakana notation and hiragana notation]
4. (がん, 癌) [both mean *cancer*, in different character types]
5. (廻り, 回り) [both mean *surrounding of*, in variation]
6. (コンピューター, コンピュータ) [both mean *computer*]

5. erroneous [e] pairs

1. (発砲スチロール, 発泡スチロール) [発砲 (shooting) is mistaken for 発泡 (foaming)]
2. (太宰府, 大宰府) [太 and 大 are mistaken]
3. (筋線維, 筋繊維) [線 and 繊 are mistaken]

6. quasi-erroneous [f] pairs

1. (スポイト, スポイド) [both mean *dropper*]
2. (ゴルフバッグ, ゴルフバック) [both mean *golf bag*]
3. (ビッグバン, ビックバン) [both mean *Big Bang*]

7. misuse [m] pairs

1. (氷漬け, 氷付け) [both mean *frozen*, but the former is not standard form]
2. (開講, 開校) [(*open a lecture*, *open a school*) yet susceptible for misuse]
3. (平行, 並行) [both mean *parallel* with difference in denotation]
4. (恋愛観, 恋愛感) [the latter is an apparently a new terms]

8. hypernym-hyponym [h] pairs

1. (検索ツール, 検索ソフト)

   [(*search tool, search software*)]

2. (失業対策, 雇用対策)

   [(*unemployment measures, employment measures*)]

3. (景況, 雇用情勢)

   [(business conditions, employment conditions)]

4. (フェスティバル, 音楽祭)

   [(festival, music festival)]

5. (シンビジウム, 洋ラン)

   [(*cymbidium, orchid*)]

6. (神秘体験, 臨死体験)

   [(*mystical experience, near-death experience*)]

9. meronymic [p] pairs

1. (ちきゅう, うみ) [(*earth, sea*)]
2. (確約, 了解) [(*affirmation, admission*)]
3. (知見, 研究成果) [(*findings, research results*)]
4. (ソーラーサーキット, 外断熱工法) [(*solar circuit system, exterior thermal insulation method*)]
5. (プロバンス, 南フランス) [(*Provence, South France*)]

10. classmates with shared morpheme [w]

1. (ガス設備, 電気設備) [(*gas facilities, electric facilities*)]
2. (系列局, 地方局) [(*affiliate station(s), local satation(s)*)]
3. (新潟市, 和歌山市) [(*Niigata City, Wakayama City*)]
4. (シナイ半島, マレー半島) [(*Sinai Peninsula, Malay Peninsula*)]

11. classmates without shared morpheme [k]

1. (Tom, Jerry)
2. (自分磨き, 体力作り) [(*self-culture, training*)]
3. (所属機関, 部局) [(*sub-organs, services*)]
4. (トンパ文字, ヒエログリフ) [(*Dongba alphabets, hieroglyphs*)]

12. contrastive pairs without antonymity [c]

1. (ロマン主義, 自然主義) [(*romanticism, naturalism*)]
2. (携帯ユーザー, インターネットユーザー) [(*mobile user(s), internet user(s)*)]
3. (海賊版, PS2版) [(*bootleg edition, PS2 edition*)]

13. antonymic [d] pairs

1. (接着, 分解) [(*bonding, disintegration*)]
2. (砂利道, 舗装路) [(*gravel road, pavement*)]
3. (西壁, 東壁) [(*west wall(s), east wall(s)*)]
4. (娘夫婦, 息子夫婦) [(*daugher and son-in-law, son and daughter-in-law*)]
5. (外税, 内税) [(*tax-exclusive prices, tax-inclusive prices*)]
6. (リアブレーキ, フロントブレーキ) [(*front break, rear brake*)]
7. (タッグマッチ, シングルマッチ) [(*tag-team match, single match*)]

14. pairs with inherent temporal order [t]

1. (稲刈り, 田植え) [(*harvesting of rice, planting of rice*)]
2. (ご出発日, ご到着日) [(*day of departure, day of arrival*)]
3. (進路決定, 進路選択) [(*career decision, career selection*)]
4. (居眠り, 夜更かし) [(*catnap, stay up*)]
5. (密猟, 密輸) [(*poaching, contraband trade*)]
6. (投降, 出兵) [(*surrender, dispatch*)]
7. (二回生, 三回生) [(*2nd-year student(s), 3rd-year student(s)*)]

15. pairs in other relation [o]

1. (下心, 独占欲) [(*ulterior motives, possessive feeling*)]
2. (理論的背景, 基本的概念) [(*theoretical background, basic concepts*)]
3. (アレクサンドリア, シラクサ) [(*Alexandria, Syracuse*)]

16. unrelated [u] pairs

1. (非接触, 高分解能) [(*noncontact, high resolution*)]
2. (模倣, 拡大解釈) [(*imitation, overinterpretation*)]

17. nonsensical [x] pairs

1. (わったん, まる赤)
2. (セルディィ, 瀬璃)
3. (チル, エルダ)
4. (ウーナ, 香螢)
5. (ma, ジョージア)

18. unclassified [y] pairs

1. (場所網, 無規準ゲーム)

2. (fj, スラド)

3. (反力, 断力)

Results

Details of the Classification Task

- 17 people were asked to perform the classification task using the guidelines specified by the first and second author.
  - The task took nearly 3 months (= regular 2 months + extra 1 month for rework).
- The quality of the product turned out to be very low in some cases.
  - Rework on **o**- and **w**-cases was requested.

Rank	Count	Ratio (%)	Cumulative (%)	Class	Label
1	108,149	36.04	36.04	classmates without common	**k**
2	67,089	22.35	58.39	classmates with common	**w**
3	26,113	8.70	67.09	synonymic pairs	**s**
4	24,599	8.20	75.29	hypernym-hyponym pairs	**h**
5	20,766	6.92	82.21	allographic pairs	**v**
6	18.950	6.31	88.52	pairs in “other” relation	**o**
7	12,383	4.13	92.65	unrelated pairs	**u**
8	8,092	2.70	95.34	contrastive pairs	**c**
9	3,793	1.26	96.61	pairs with temporal order	**t**
10	3,038	1.01	97.62	antonymic pairs	**d**
11	2,995	1.00	98.62	meronymic pairs	**p**
12	1,855	0.62	99.23	acronymic pairs	**a**
13	725	0.24	99.48	alias pairs	**n**
14	715	0.24	99.71	erroneous pairs	**e**
15	397	0.13	99.85	misuse pairs	**m**
16	250	0.08	99.93	nonsensical pairs	**x**
17	180	0.06	99.99	quasi-erroneous pairs	**f**
18	33	0.01	100.00	unclassified	**y**

Basic Results

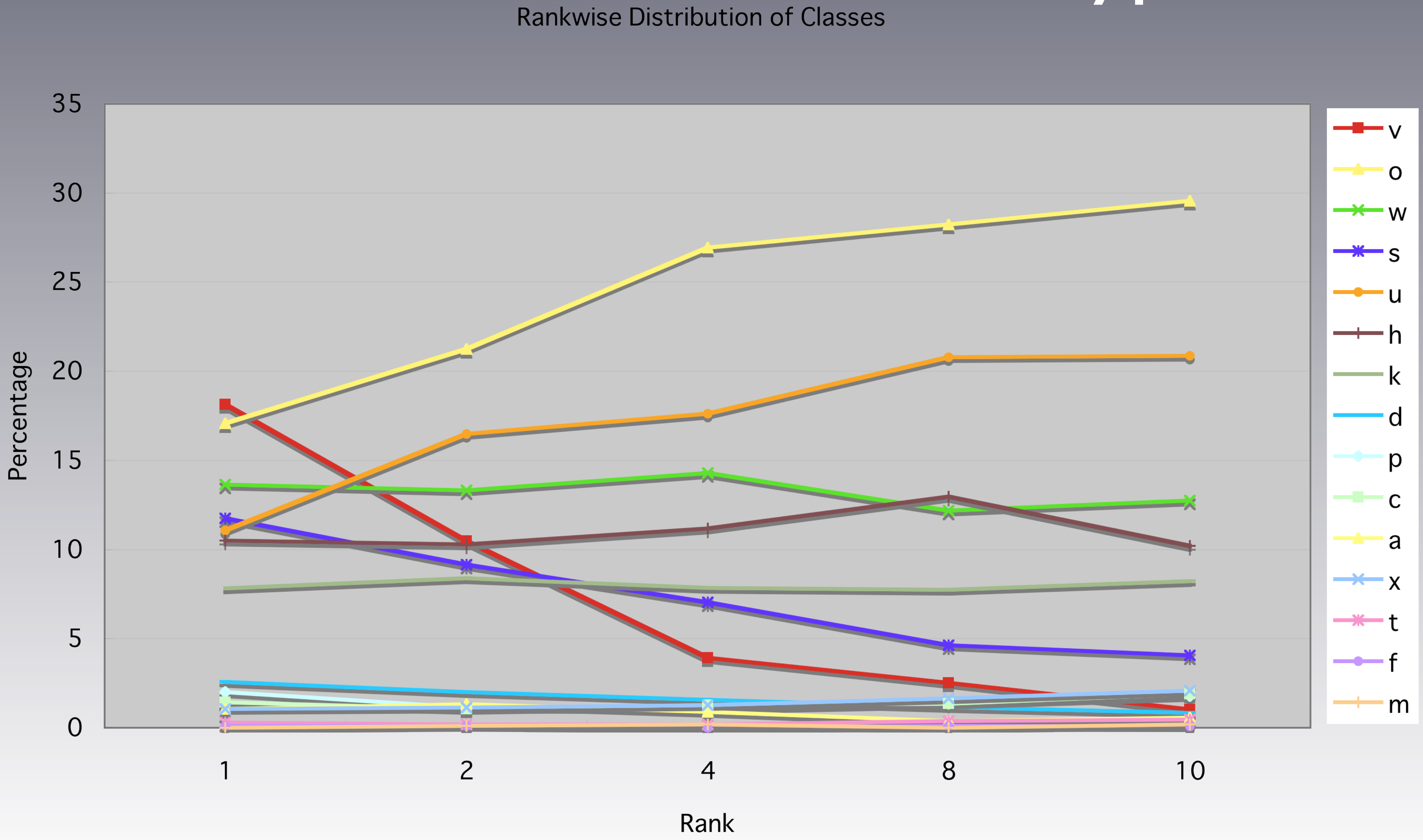
1. Union of **k** and **w** makes 58.39% (strict condition).
2. Union of **k**** and **s*** makes 79.01% (moderate condition).
   - **k**** = {**k**, **w**, **c**, **d**, **t**} is a generalized class of classmates to make 62.10%.
   - **s*** = {**s**, **a**, **n**, **v**, **e**, **f**, **m**} generalized class of synonymic pairs to make 16.91%
3. All classes except **o**, **u**, **m**, **x** and **y** make roughly 88% (loose condition).

- The second or third conditions can be understood as confirmations of the "distributional" hypothesis.

Further Question

- What is the (side)effect of k = 2? Did we get a representative result?
- An informal preliminary analysis of sample 1000 pairs (generated based on bases at ranks 2, 4, 8, 10) indicates
  - the rate of s* (especially v) decreases at lower ranks.
  - the rates of o and u increase at lower ranks.

Rankwise Distribution of Types



Summary

- Our aim was to see to what extent distributionally similar terms can be equated with semantically similar terms when semantic similarity is factored out.
- Loose condition with all labels except **o**, **u**, **m**, **x** and **y** make roughly 88%. Even moderate condition with k** and s* makes 79.01%. So, it would be safe to say that the "distributional" hypothesis is confirmed.
- Though our case is limited in that *n*=150,000 and *k*=2, rankwise distribution of class suggests that our results are with fair representativeness.

Thank you
for Your Attention

Appendix

Potential inconsistency

- The distinction among classes is sometimes obscure, especially the one between **p** and **h** is hard to make in Japanese.
  - For example, is the right label for (火星, 天体) **p** or **h**?
- This ambiguity is influenced by the ambiguity of 天体: If *heavenly body* is meant, then **h** is right. If *heavenly bodies* is meant, then **p** is right.