A Study of Reuse and Plagiarism in Speech and Natural Language Processing Papers

Joseph Mariani[1], Gil Francopoulo[2], Patrick Paroubek[1]

[1]LIMSI-CNRS, [2]Tagmatica

Objectives

- Study the practices of the NLP (Spoken, Written and Sign Language) community regarding reuse and plagiarism
  - Check whether there is a *meaningful* difference in taking the verbatim raw word strings compared with applying natural language processing methods to detect possible cases of reuse and plagiarism?

NLP4NLP Corpus

- Presently conduct large scholar analysis of NLP domain
  - Production, Collaboration, Citation, Innovation
- NLP4NLP: 34 sources over 50 years (1965-2015)
- Major conferences (ACL, IEEE-ICASSP, ISCA-Interspeech, ELRA-LREC, etc.) and Journals (IEEE-TASLP, CL, SpeechCom, CSAL, LRE, etc.)
- 558 Venues (conferences) / Issues (journals)
- 65,003 documents
- 48,894 Authors
- 270 MWords

NLP4NLP Corpus

short name	# docs	format	long name	language	access to content	period	# venues
acl	4264	conference	Association for Computational Linguistics Conference	English	open access *	1979-2015	37
acmtslp	82	journal	ACM Transaction on Speech and Language Processing	English	private access	2004-2013	10
alta	262	conference	Australasian Language Technology Association	English	open access *	2003-2014	12
anlp	278	conference	Applied Natural Language Processing	English	open access *	1983-2000	6
cath	932	journal	Computers and the Humanities	English	private access	1966-2004	39
cl	776	journal	American Journal of Computational Linguistics	English	open access *	1980-2014	35
coling	3813	conference	Conference on Computational Linguistics	English	open access *	1965-2014	21
conll	842	conference	Computational Natural Language Learning	English	open access *	1997-2015	18
csal	762	journal	Computer Speech and Language	English	private access	1986-2015	29
eacl	900	conference	European Chapter of the ACL	English	open access *	1983-2014	14
emnlp	2020	conference	Empirical methods in natural language processing	English	open access *	1996-2015	20
hlt	2219	conference	Human Language Technology	English	open access *	1986-2015	19
icassps	9819	conference	IEEE International Conference on Acoustics, Speech and Signal Processing - Speech Track	English	private access	1990-2015	26
ijcnlp	1188	conference	International Joint Conference on NLP	English	open access *	2005-2015	6
inlg	227	conference	International Conference on Natural Language Generation	English	open access *	1996-2014	7
isca	18369	conference	International Speech Communication Association	English	open access	1987-2015	28
jep	507	conference	Journées d'Etudes sur la Parole	French	open access *	2002-2014	5
lre	308	journal	Language Resources and Evaluation	English	private access	2005-2015	11
lrec	4552	conference	Language Resources and Evaluation Conference	English	open access *	1998-2014	9
ltc	656	conference	Language and Technology Conference	English	private access	1995-2015	7
modulad	232	journal	Le Monde des Utilisateurs de L'Analyse des Données	French	open access	1988-2010	23
mts	796	conference	Machine Translation Summit	English	open access	1987-2015	15
muc	149	conference	Message Understanding Conference	English	open access *	1991-1998	5
naacl	1186	conference	North American Chapter of ACL	English	open access *	2000-2015	11
paclic	1040	conference	Pacific Asia Conference on Language, Information and Computation	English	open access *	1995-2014	19
ranlp	363	conference	Recent Advances in Natural Language Processing	English	open access *	2009-2013	3
sem	950	conference	Lexical and Computational Semantics / Semantic Evaluation	English	open access *	2001-2015	8
speechc	593	journal	Speech Communication	English	private access	1982-2015	34
tacl	92	journal	Transactions of the Association for Computational Linguistics	English	open access *	2013-2015	3
tal	177	journal	Revue Traitement Automatique du Langage	French	open access	2006-2015	10
taln	1019	conference	Traitement Automatique du Langage Naturel	French	open access *	1997-2015	19
taslp	6612	journal	IEEE/ACM Transactions on Audio, Speech and Language Processing	English	private access	1975-2015	41
tipster	105	conference	Tipster DARPA text program	English	open access *	1993-1998	3
trec	1847	conference	Text Retrieval Conference	English	open access	1992-2015	24
Total incl. duplicates	67937					1965-2015	577
Total excl. duplicates	65,003					1965-2015	558

Definitions

- “**Self-reuse**”: copy & paste when the source of the copy has at least one author who belongs to the group of authors of the text of the paste and when the source is cited.
- “**Self-plagiarism**”: copy & paste when the source of the copy has at least one author who belongs to the group of authors of the text of the paste, but when the source is not cited.
- “**Reuse**”: copy & paste when the source of the copy has no author in the group of authors of the paste and when the source is cited.
- “**Plagiarism**”: copy & paste when the source of the copy has no author in the group of the paste and when the source is not cited.

Definitions

	Source paper is quoted	Source paper is not quoted
At least one author in common	Self-Reuse	Self-Plagiarism
No author in common	Reuse	Plagiarism

Each year: Papers of the focus borrowing papers of the search space (same year or previous years: Backward study)

Search Space Focus	NLP4NLP (Same year or previous years)			
	Self-Reusing	Self-Plagiarizing	Reusing	Plagiarizing
Year1				
Year2				
Year3				
…				

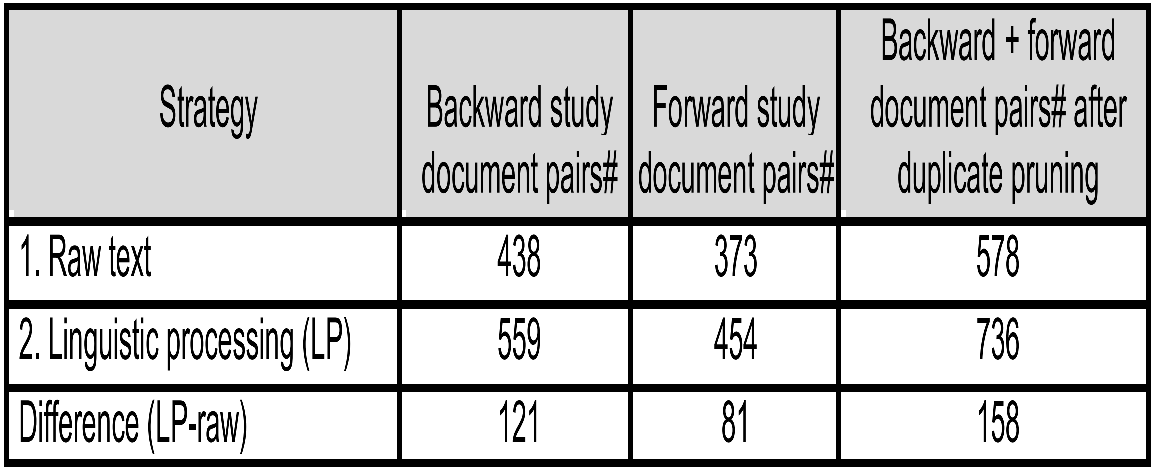
Each year: Papers of the focus being borrowed by papers of search space (same year or following years: Forward study)

Search Space Focus	NLP4NLP (Same year or following years)			
	Self-Reused	Self-Plagiarized	Reused	Plagiarized
Year1				
Year2				
Year3				
…				

Algorithm

- Based on comparison of word sequences, had to be optimized:
- For each pair of documents D1 of the focus (LREC) and D2 of the search space (NLP4NLP), consider
  1. either raw text
  2. or text after LP (Tagparser [Francopoulo 2007] with Global Atlas + LRE Map)
     - Hyphen variations
     - Caesura
     - Upper/lower cases
     - Plurals
     - Orthographic variations (British English versus American English)
     - Spelling errors
     - Abbreviations (BNC versus British National Corpus)
- Compare 2 texts D1 / D2 using sliding windows of (5-7) lemmas (excluding punctuations)
- Compute a similarity overlapping score [Lyon et al 2001] between documents D1 and D2, with (a variant of) the Jaccard similarity coefficient
  - **Score (D1,D2) = #shared windows / #union (D1 windows, D2 windows)**
- Filter the pairs of documents D1 / D2 according to a threshold of (0.03-0.04) (3-4% coverage) to retain only significantly similar pairs

Raw text versus LP



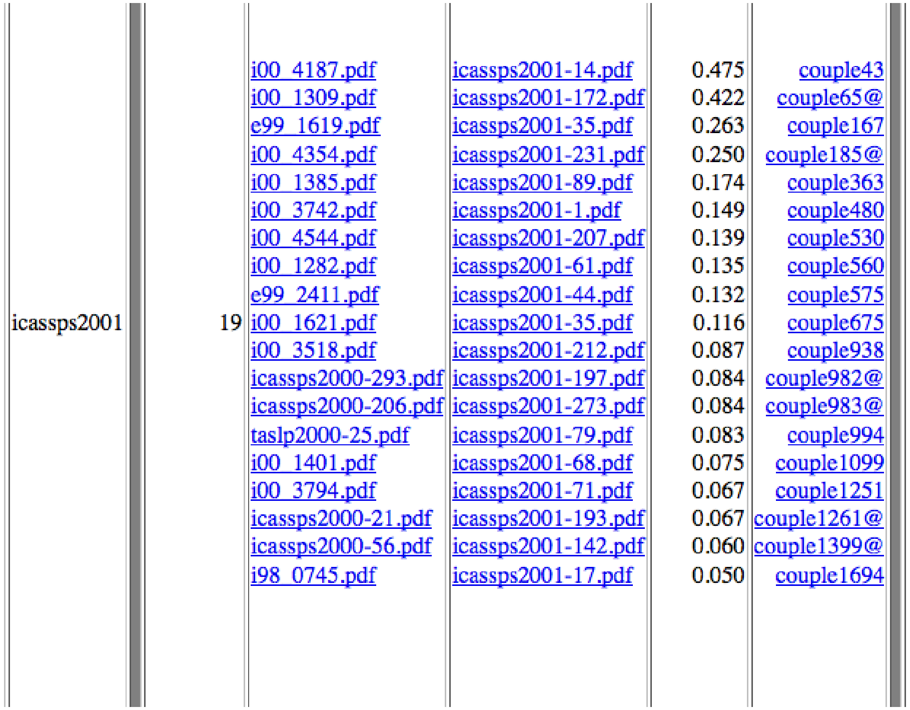
Strategy	Backward study document pairs#	Forward study document pairs#	Backward + forward document pairs# after duplicate pruning
1. Raw text	438	373	578
2. Linguistic processing (LP)	559	454	736
Difference (LP-raw)	121	81	158

Tuning Parameters

- Windows: 7 words
- Jaccard similarity coefficient
- Similarity threshold: 0.04 (4%)
- + Number of shared windows > 50

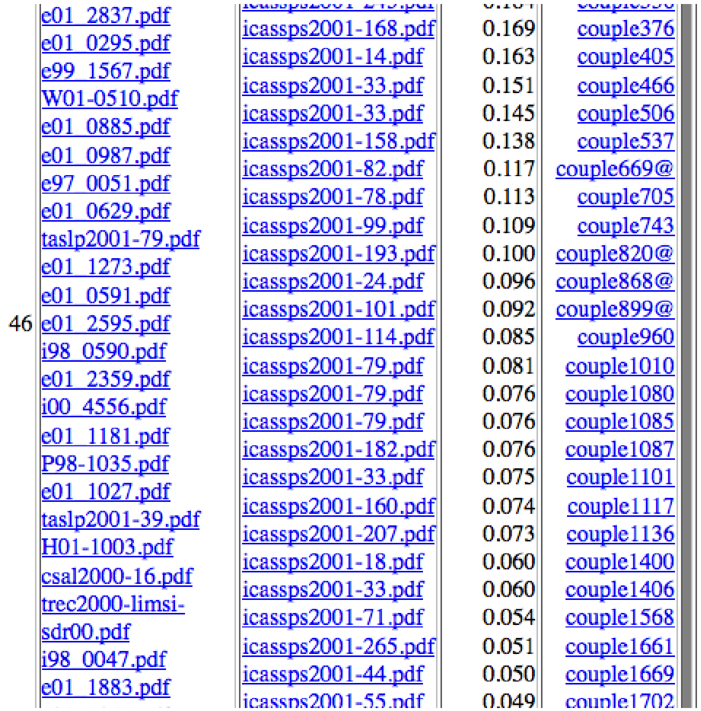
Example of IEEE ICASSP 2001

Self-Reusing



icassps2001	19	i00_4187.pdf	icassps2001-14.pdf	0.475	couple43
		i00_1309.pdf	icassps2001-172.pdf	0.422	couple65@
		e99_1619.pdf	icassps2001-35.pdf	0.263	couple167
		i00_4354.pdf	icassps2001-231.pdf	0.250	couple185@
		i00_1385.pdf	icassps2001-89.pdf	0.174	couple363
		i00_3742.pdf	icassps2001-1.pdf	0.149	couple480
		i00_4544.pdf	icassps2001-207.pdf	0.139	couple530
		i00_1282.pdf	icassps2001-61.pdf	0.135	couple560
		e99_2411.pdf	icassps2001-44.pdf	0.132	couple575
		i00_1621.pdf	icassps2001-35.pdf	0.116	couple675
		i00_3518.pdf	icassps2001-212.pdf	0.087	couple938
		icassps2000-293.pdf	icassps2001-197.pdf	0.084	couple982@
		icassps2000-206.pdf	icassps2001-273.pdf	0.084	couple983@
		taslp2000-25.pdf	icassps2001-79.pdf	0.083	couple994
		i00_1401.pdf	icassps2001-68.pdf	0.075	couple1099
		i00_3794.pdf	icassps2001-71.pdf	0.067	couple1251
		icassps2000-21.pdf	icassps2001-193.pdf	0.067	couple1261@
		icassps2000-56.pdf	icassps2001-142.pdf	0.060	couple1399@
		i98_0745.pdf	icassps2001-17.pdf	0.050	couple1694

Self-Plagiarizing

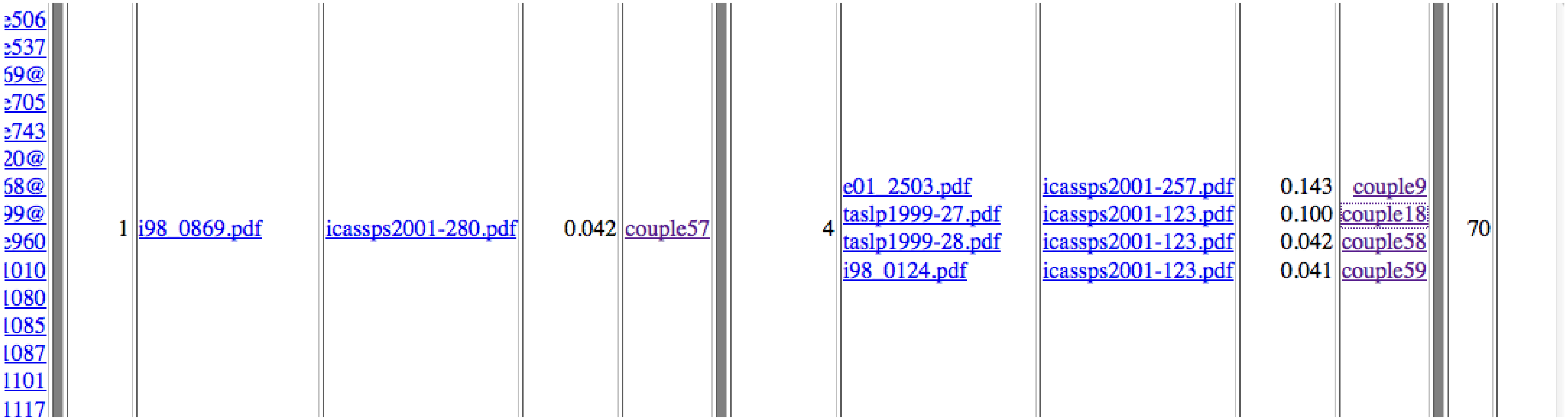


46	e01_2837.pdf e01_0295.pdf e99_1567.pdf W01-0510.pdf e01_0885.pdf e01_0987.pdf e97_0051.pdf e01_0629.pdf taslp2001-79.pdf e01_1273.pdf e01_0591.pdf e01_2595.pdf i98_0590.pdf e01_2359.pdf i00_4556.pdf e01_1181.pdf P98-1035.pdf e01_1027.pdf taslp2001-39.pdf H01-1003.pdf csal2000-16.pdf trec2000-limsi-sdr00.pdf i98_0047.pdf e01_1883.pdf	icassps2001-168.pdf	0.169	couple376
		icassps2001-14.pdf	0.163	couple405
		icassps2001-33.pdf	0.151	couple466
		icassps2001-33.pdf	0.145	couple506
		icassps2001-158.pdf	0.138	couple537
		icassps2001-82.pdf	0.117	couple669@
		icassps2001-78.pdf	0.113	couple705
		icassps2001-99.pdf	0.109	couple743
		icassps2001-193.pdf	0.100	couple820@
		icassps2001-24.pdf	0.096	couple868@
		icassps2001-101.pdf	0.092	couple899@
		icassps2001-114.pdf	0.085	couple960
		icassps2001-79.pdf	0.081	couple1010
		icassps2001-79.pdf	0.076	couple1080
		icassps2001-79.pdf	0.076	couple1085
		icassps2001-182.pdf	0.076	couple1087
		icassps2001-33.pdf	0.075	couple1101
		icassps2001-160.pdf	0.074	couple1117
		icassps2001-207.pdf	0.073	couple1136
		icassps2001-18.pdf	0.060	couple1400
		icassps2001-33.pdf	0.060	couple1406
		icassps2001-71.pdf	0.054	couple1568
		icassps2001-265.pdf	0.051	couple1661
		icassps2001-44.pdf	0.050	couple1669
		icassps2001-55.pdf	0.049	couple1702

Example of IEEE ICASSP 2001

Reusing

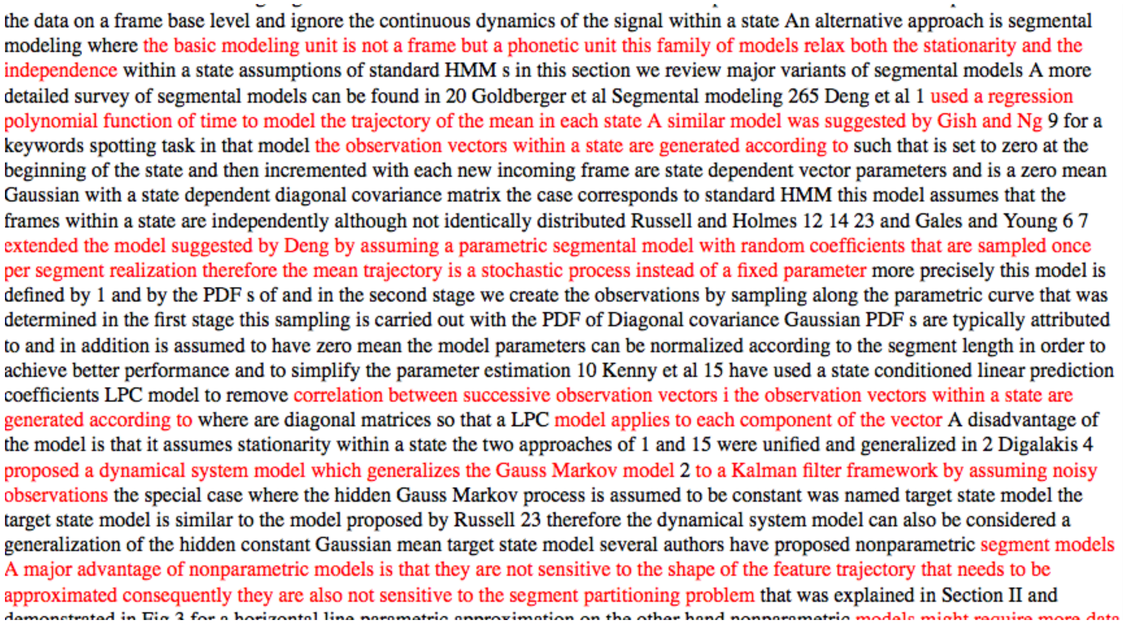
Plagiarizing



e506 e537 69@ e705 e743 20@ 68@ 99@ e960 1010 1080 1085 1087 1101 1117	1	i98_0869.pdf	icassps2001-280.pdf	0.042	couple57	4	e01_2503.pdf taslp1999-27.pdf taslp1999-28.pdf i98_0124.pdf	icassps2001-257.pdf icassps2001-123.pdf icassps2001-123.pdf icassps2001-123.pdf	0.143 0.100 0.042 0.041	couple9 couple18 couple58 couple59	70

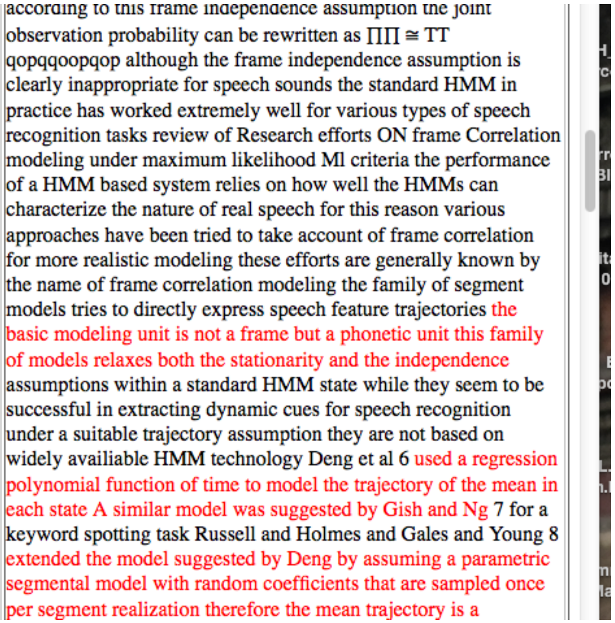
Example of similarities between 2 papers (couple 18)

taslp1999-27.pdf



the data on a frame base level and ignore the continuous dynamics of the signal within a state An alternative approach is segmental modeling where the basic modeling unit is not a frame but a phonetic unit this family of models relax both the stationarity and the independence within a state assumptions of standard HMM s in this section we review major variants of segmental models A more detailed survey of segmental models can be found in 20 Goldberger et al Segmental modeling 265 Deng et al 1 used a regression polynomial function of time to model the trajectory of the mean in each state A similar model was suggested by Gish and Ng 9 for a keywords spotting task in that model the observation vectors within a state are generated according to such that is set to zero at the beginning of the state and then incremented with each new incoming frame are state dependent vector parameters and is a zero mean Gaussian with a state dependent diagonal covariance matrix the case corresponds to standard HMM this model assumes that the frames within a state are independently although not identically distributed Russell and Holmes 12 14 23 and Gales and Young 6 7 extended the model suggested by Deng by assuming a parametric segmental model with random coefficients that are sampled once per segment realization therefore the mean trajectory is a stochastic process instead of a fixed parameter more precisely this model is defined by 1 and by the PDF s of and in the second stage we create the observations by sampling along the parametric curve that was determined in the first stage this sampling is carried out with the PDF of Diagonal covariance Gaussian PDF s are typically attributed to and in addition is assumed to have zero mean the model parameters can be normalized according to the segment length in order to achieve better performance and to simplify the parameter estimation 10 Kenny et al 15 have used a state conditioned linear prediction coefficients LPC model to remove correlation between successive observation vectors i the observation vectors within a state are generated according to where are diagonal matrices so that a LPC model applies to each component of the vector A disadvantage of the model is that it assumes stationarity within a state the two approaches of 1 and 15 were unified and generalized in 2 Digalakis 4 proposed a dynamical system model which generalizes the Gauss Markov model 2 to a Kalman filter framework by assuming noisy observations the special case where the hidden Gauss Markov process is assumed to be constant was named target state model the target state model is similar to the model proposed by Russell 23 therefore the dynamical system model can also be considered a generalization of the hidden constant Gaussian mean target state model several authors have proposed nonparametric segment models A major advantage of nonparametric models is that they are not sensitive to the shape of the feature trajectory that needs to be approximated consequently they are also not sensitive to the segment partitioning problem that was explained in Section II and demonstrated in Fig 3 for a horizontal line parametric approximation on the other hand nonparametric models might require more data

Icassps2001-123.pdf

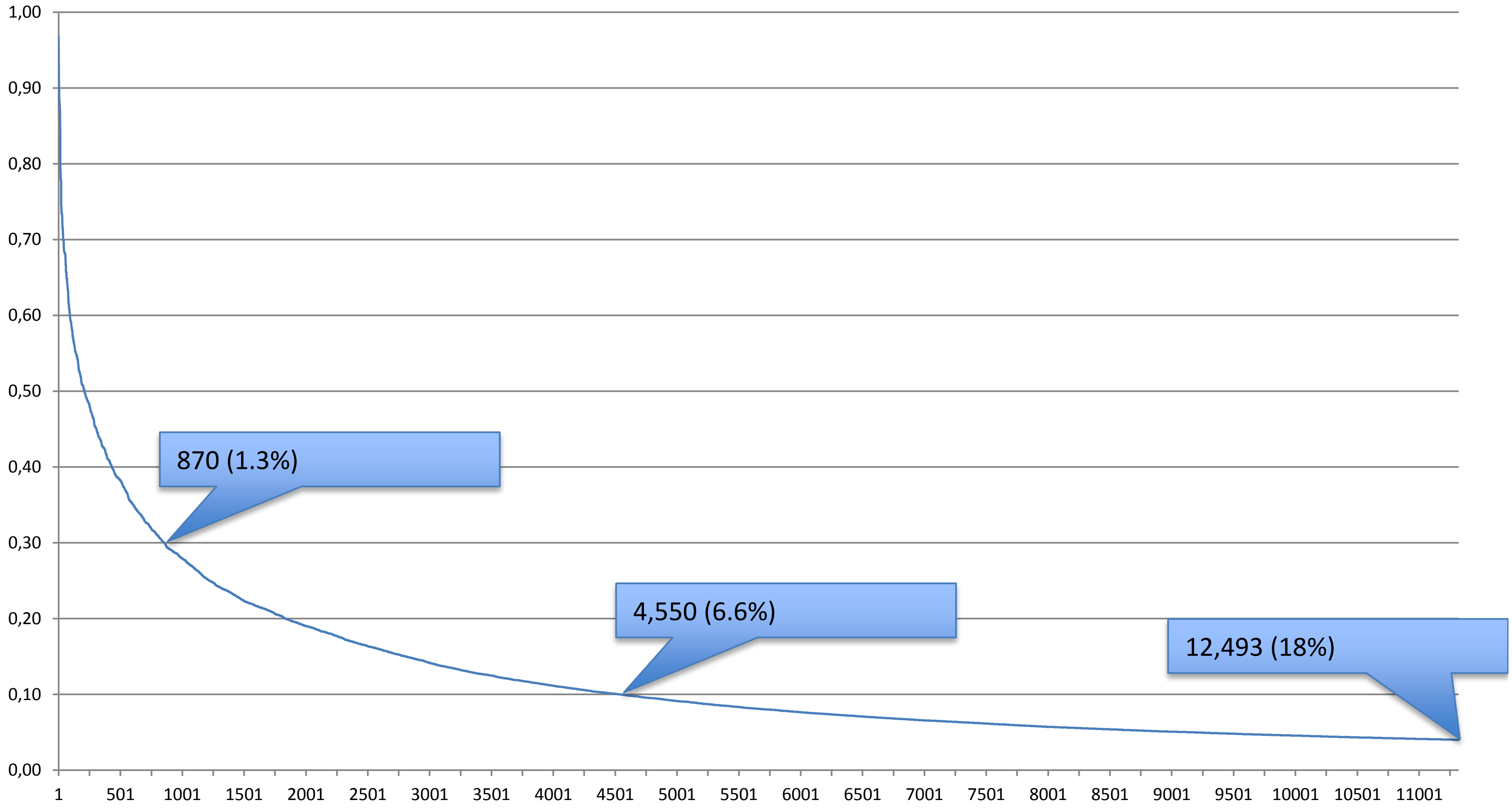


according to this frame independence assumption the joint observation probability can be rewritten as ∏∏ ≅ TT qopqqoopqop although the frame independence assumption is clearly inappropriate for speech sounds the standard HMM in practice has worked extremely well for various types of speech recognition tasks review of Research efforts ON frame Correlation modeling under maximum likelihood Ml criteria the performance of a HMM based system relies on how well the HMMs can characterize the nature of real speech for this reason various approaches have been tried to take account of frame correlation for more realistic modeling these efforts are generally known by the name of frame correlation modeling the family of segment models tries to directly express speech feature trajectories the basic modeling unit is not a frame but a phonetic unit this family of models relaxes both the stationarity and the independence assumptions within a standard HMM state while they seem to be successful in extracting dynamic cues for speech recognition under a suitable trajectory assumption they are not based on widely availiable HMM technology Deng et al 6 used a regression polynomial function of time to model the trajectory of the mean in each state A similar model was suggested by Gish and Ng 7 for a keyword spotting task Russell and Holmes and Gales and Young 8 extended the model suggested by Deng by assuming a parametric segmental model with random coefficients that are sampled once per segment realization therefore the mean trajectory is a

Self Reuse-Plagiarism

- 12,493 cases (18% papers) : no manual checking
  - 4% to 97% overlapping
  - In 61% of the cases, authors do not quote the source paper
  - 130 papers have both the same title and the same list of authors
  - 205 papers have the same title
- Some specific cases (largest similarities)
  - Republishing the corrigendum of a previously published paper
  - Republishing a paper with a small difference in the title and one missing author in the authors' list
  - Same research center described by the same author in two different conferences, with an overlapping of 90%
  - 2 papers presented by the same author in 2 successive conferences, the difference being primarily in the name of the 2 systems being presented, that have been funded by the same project agency in 2 different contracts, with an overlapping of 45%

Similarity Scores Self Reuse-Plagiarism

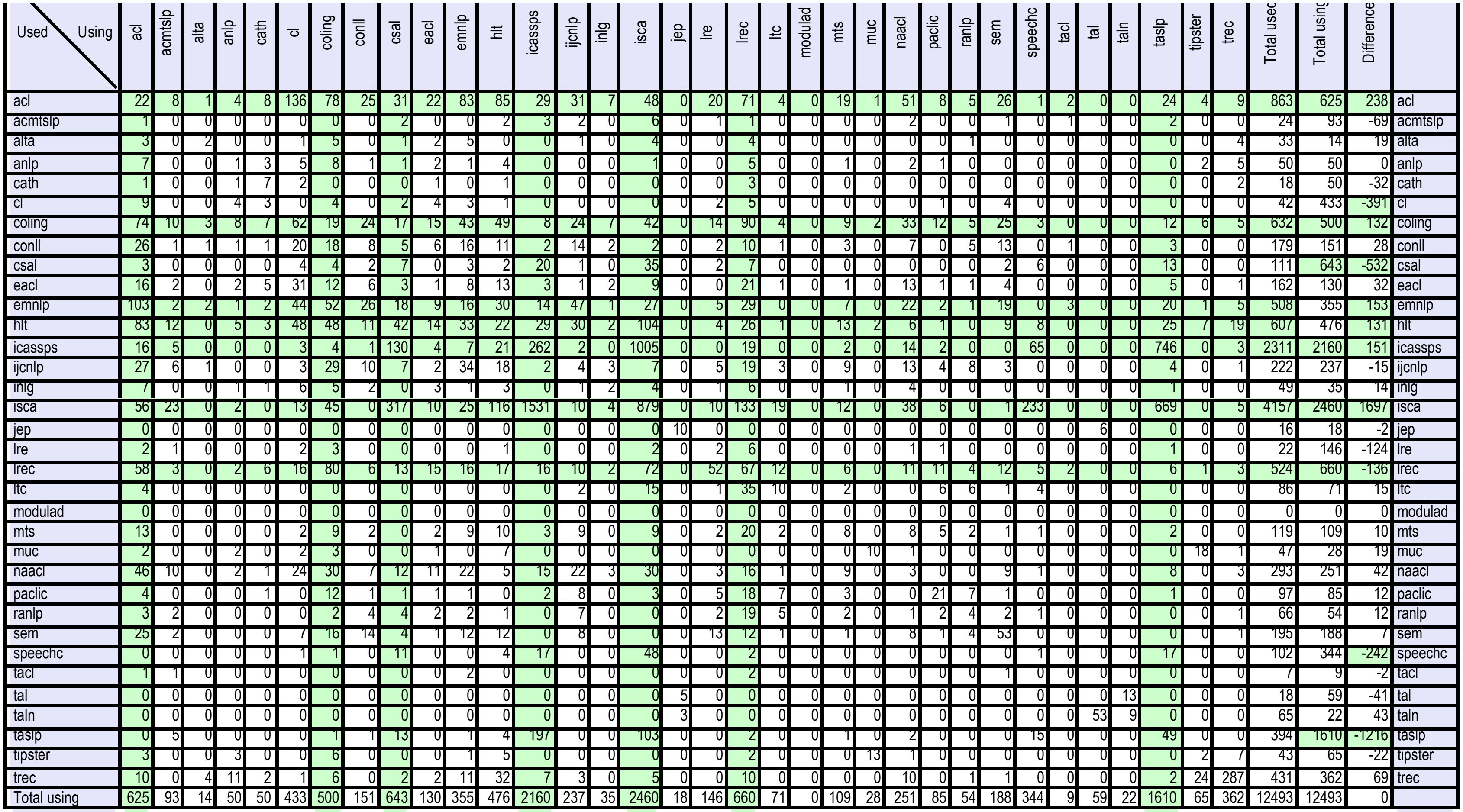


870 (1.3%)

4,550 (6.6%)

12,493 (18%)

Self Reuse-Plagiarism



Used \ Using	acl	acmtslp	alta	anlp	cath	cl	coling	conll	csal	eacl	emnlp	hlt	icassps	ijcnlp	inlg	isca	jep	lre	lrec	ltc	modulad	mts	muc	naacl	paclic	ranlp	sem	speechc	tacl	tal	taln	taslp	tipster	trec	Total used	Total using	Difference	
acl	22	8	1	4	8	136	78	25	31	22	83	85	29	31	7	48	0	20	71	4	0	19	1	51	8	5	26	1	2	0	0	24	4	9	863	625	238	acl
acmtslp	1	0	0	0	0	0	0	0	2	0	0	2	3	2	0	6	0	1	1	0	0	0	0	2	0	0	1	0	1	0	0	2	0	0	24	93	-69	acmtslp
alta	3	0	2	0	0	1	5	0	1	2	5	0	0	1	0	4	0	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4	33	14	19	alta
anlp	7	0	0	1	3	5	8	1	1	2	1	4	0	0	0	1	0	0	5	0	0	1	0	2	1	0	0	0	0	0	0	0	2	5	50	50	0	anlp
cath	1	0	0	1	7	2	0	0	0	1	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	18	50	-32	cath
cl	9	0	0	4	3	0	4	0	2	4	3	1	0	0	0	0	0	2	5	0	0	0	0	0	1	0	4	0	0	0	0	0	0	0	42	433	-391	cl
coling	74	10	3	8	7	62	19	24	17	15	43	49	8	24	7	42	0	14	90	4	0	9	2	33	12	5	25	3	0	0	0	12	6	5	632	500	132	coling
conll	26	1	1	1	1	20	18	8	5	6	16	11	2	14	2	2	0	2	10	1	0	3	0	7	0	5	13	0	1	0	0	3	0	0	179	151	28	conll
csal	3	0	0	0	0	4	4	2	7	0	3	2	20	1	0	35	0	2	7	0	0	0	0	0	0	0	2	6	0	0	0	13	0	0	111	643	-532	csal
eacl	16	2	0	2	5	31	12	6	3	1	8	13	3	1	2	9	0	0	21	1	0	1	0	13	1	1	4	0	0	0	0	5	0	1	162	130	32	eacl
emnlp	103	2	2	1	2	44	52	26	18	9	16	30	14	47	1	27	0	5	29	0	0	7	0	22	2	1	19	0	3	0	0	20	1	5	508	355	153	emnlp
hlt	83	12	0	5	3	48	48	11	42	14	33	22	29	30	2	104	0	4	26	1	0	13	2	6	1	0	9	8	0	0	0	25	7	19	607	476	131	hlt
icassps	16	5	0	0	0	3	4	1	130	4	7	21	262	2	0	1005	0	0	19	0	0	2	0	14	2	0	0	65	0	0	0	746	0	3	2311	2160	151	icassps
ijcnlp	27	6	1	0	0	3	29	10	7	2	34	18	2	4	3	7	0	5	19	3	0	9	0	13	4	8	3	0	0	0	0	4	0	1	222	237	-15	ijcnlp
inlg	7	0	0	1	1	6	5	2	0	3	1	3	0	1	2	4	0	1	6	0	0	1	0	4	0	0	0	0	0	0	0	1	0	0	49	35	14	inlg
isca	56	23	0	2	0	13	45	0	317	10	25	116	1531	10	4	879	0	10	133	19	0	12	0	38	6	0	1	233	0	0	0	669	0	5	4157	2460	1697	isca
jep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	16	18	-2	jep
lre	2	1	0	0	0	2	3	0	0	0	0	1	0	0	0	2	0	2	6	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	22	146	-124	lre
lrec	58	3	0	2	6	16	80	6	13	15	16	17	16	10	2	72	0	52	67	12	0	6	0	11	11	4	12	5	2	0	0	6	1	3	524	660	-136	lrec
ltc	4	0	0	0	0	0	0	0	0	0	0	0	0	2	0	15	0	1	35	10	0	2	0	0	6	6	1	4	0	0	0	0	0	0	86	71	15	ltc
modulad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	modulad
mts	13	0	0	0	0	2	9	2	0	2	9	10	3	9	0	9	0	2	20	2	0	8	0	8	5	2	1	1	0	0	0	2	0	0	119	109	10	mts
muc	2	0	0	2	0	2	3	0	0	1	0	7	0	0	0	0	0	0	0	0	0	0	10	1	0	0	0	0	0	0	0	0	18	1	47	28	19	muc
naacl	46	10	0	2	1	24	30	7	12	11	22	5	15	22	3	30	0	3	16	1	0	9	0	3	0	0	9	1	0	0	0	8	0	3	293	251	42	naacl
paclic	4	0	0	0	1	0	12	1	1	1	1	0	2	8	0	3	0	5	18	7	0	3	0	0	21	7	1	0	0	0	0	1	0	0	97	85	12	paclic
ranlp	3	2	0	0	0	0	2	4	4	2	2	1	0	7	0	0	0	2	19	5	0	2	0	1	2	4	2	1	0	0	0	0	0	1	66	54	12	ranlp
sem	25	2	0	0	0	7	16	14	4	1	12	12	0	8	0	0	0	13	12	1	0	1	0	8	1	4	53	0	0	0	0	0	0	1	195	188	7	sem
speechc	0	0	0	0	0	1	1	0	11	0	0	4	17	0	0	48	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	17	0	0	102	344	-242	speechc
tacl	1	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	7	9	-2	tacl
tal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	18	59	-41	tal
taln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	53	9	0	0	0	65	22	43	taln
taslp	0	5	0	0	0	0	1	1	13	0	1	4	197	0	0	103	0	0	2	0	0	1	0	2	0	0	0	15	0	0	0	49	0	0	394	1610	-1216	taslp
tipster	3	0	0	3	0	0	6	0	0	0	1	5	0	0	0	0	0	0	2	0	0	0	13	1	0	0	0	0	0	0	0	0	2	7	43	65	-22	tipster
trec	10	0	4	11	2	1	6	0	2	2	11	32	7	3	0	5	0	0	10	0	0	0	0	10	0	1	1	0	0	0	0	2	24	287	431	362	69	trec
Total using	625	93	14	50	50	433	500	151	643	130	355	476	2160	237	35	2460	18	146	660	71	0	109	28	251	85	54	188	344	9	59	22	1610	65	362	12493	12493	0	

Reuse and Plagiarism

- 261 cases : manual checking
- Reuse
  - 12 have a least one author in common, but with a somehow different spelling and should therefore be placed in the "Self-reuse" category.
- Plagiarism
  - 25 have a least one author in common, but with a somehow different spelling, and should therefore be placed in the "Self-plagiarism" category
  - 14 correctly quote the source paper, but with variants in the spelling of the authors' names, of the paper's title or of the conference or journal source, or correctly citing the source paper but forgetting to place it among the references, and should therefore be placed in the "Reuse" category.

Variants in Spelling Authors' Name

- Non-Linear Probability Estimation Method Used in HMM for Modeling Frame Correlation
  - Qing Guo, Fang Zheng, Jian Wu, and Wenhu Wu (ISCA-Interspeech 1998)
- An New Method Used in HMM for Modeling Frame Correlation
  - Guo Qing, Zheng Fang, Wu Jian and Wu Wenhu (IEEE-ICASSP 1999)

Variants in Spelling References

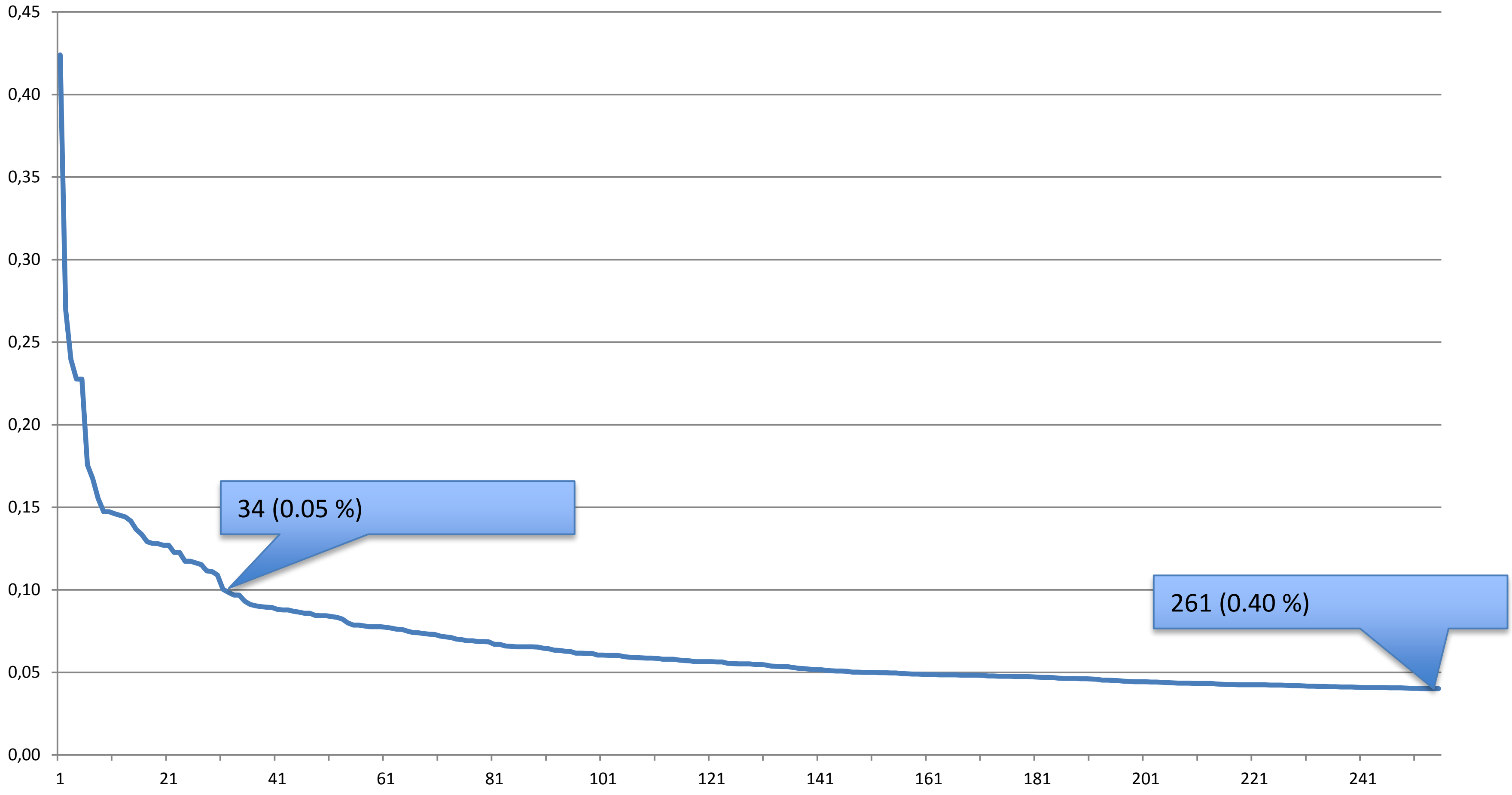
- Quoted Reference: Graham W. (2007) “an OWL Ontology for HPSG” proceeding of the ACL 2007 demo and poster sessions, 169-172.
- Correct Reference: Graham Wilcock (2007), “An OWL Ontology for HPSG”

- Quoted Reference: Li Liu, Jianglong He, “On the use of orthogonal GMM in speaker verification”
- Correct Reference: Li Liu and Jialong He, “On the use of orthogonal GMM in speaker recognition”

Reuse and Plagiarism

- After manual corrections: 224 cases (0.33% of papers)
  - 4% to 42% overlapping
  - In 52% of the cases, authors do not quote the source paper
  - This results in 117 possible cases of plagiarism (0.17%):
    - The copying paper cites another reference from the same authors of the source paper (typically a previous reference, or a paper published in a Journal) (46 cases)
    - Both papers use extracts of a third paper that they both cite (31 cases)
    - Authors of the two papers are different, but from same laboratory (typical in industrial laboratories or funding agencies) (11 cases)
    - Authors of the two papers previously co-authored papers (typically as supervisor and PhD student or postdoc) but are now in different laboratories (11 cases)
    - Authors of the papers are different, but collaborated in the same project which is presented in the two papers (2 cases)
    - The two papers present the same short example, result or definition coming from another source (13 cases)
    - Only 3 remaining cases of possible plagiarism: same paper as a patchwork of 3 other papers, while sharing several references with them.

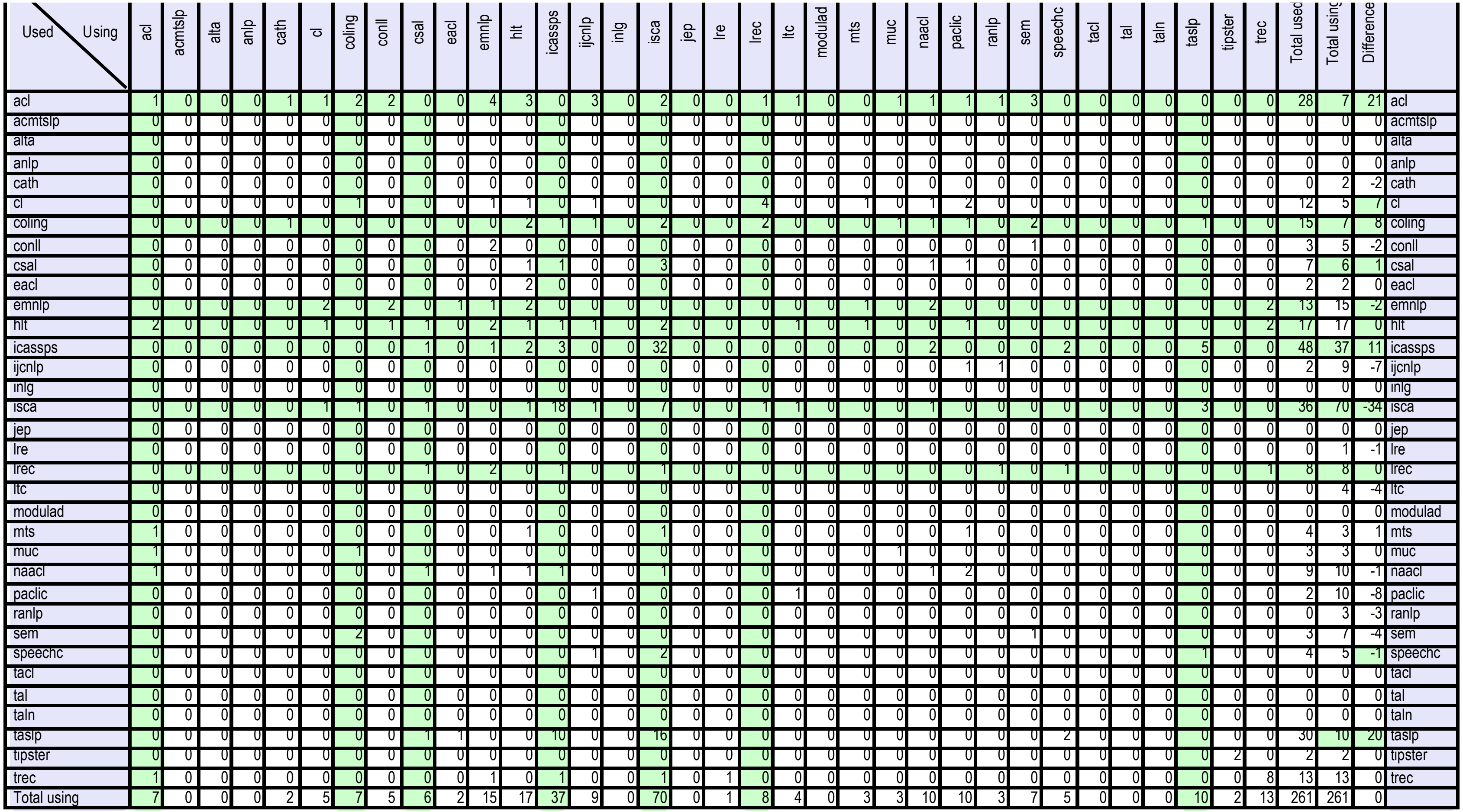
Similarity Scores Reuse/Plagiarism



34 (0.05 %)

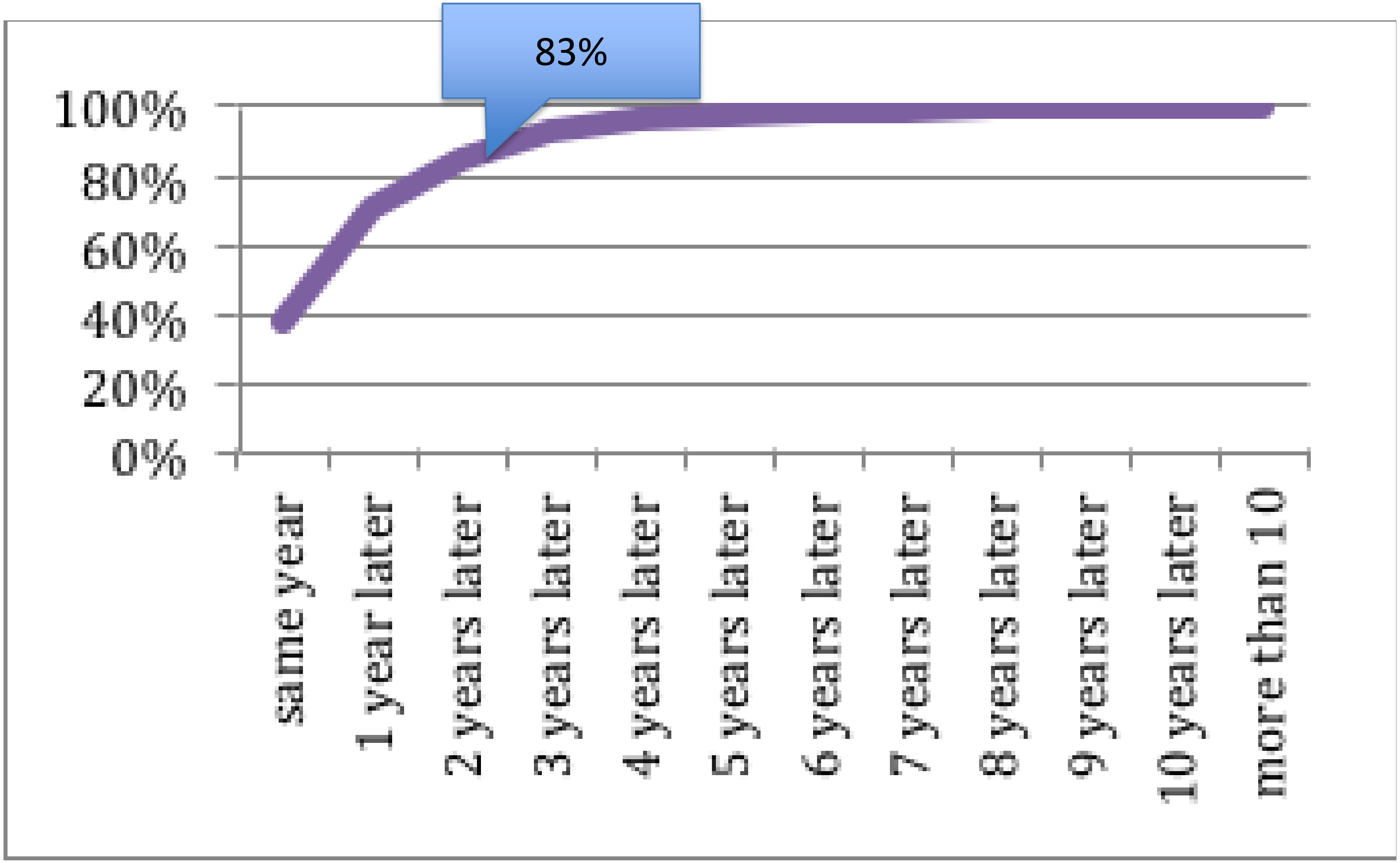
261 (0.40 %)

Reuse and Plagiarism

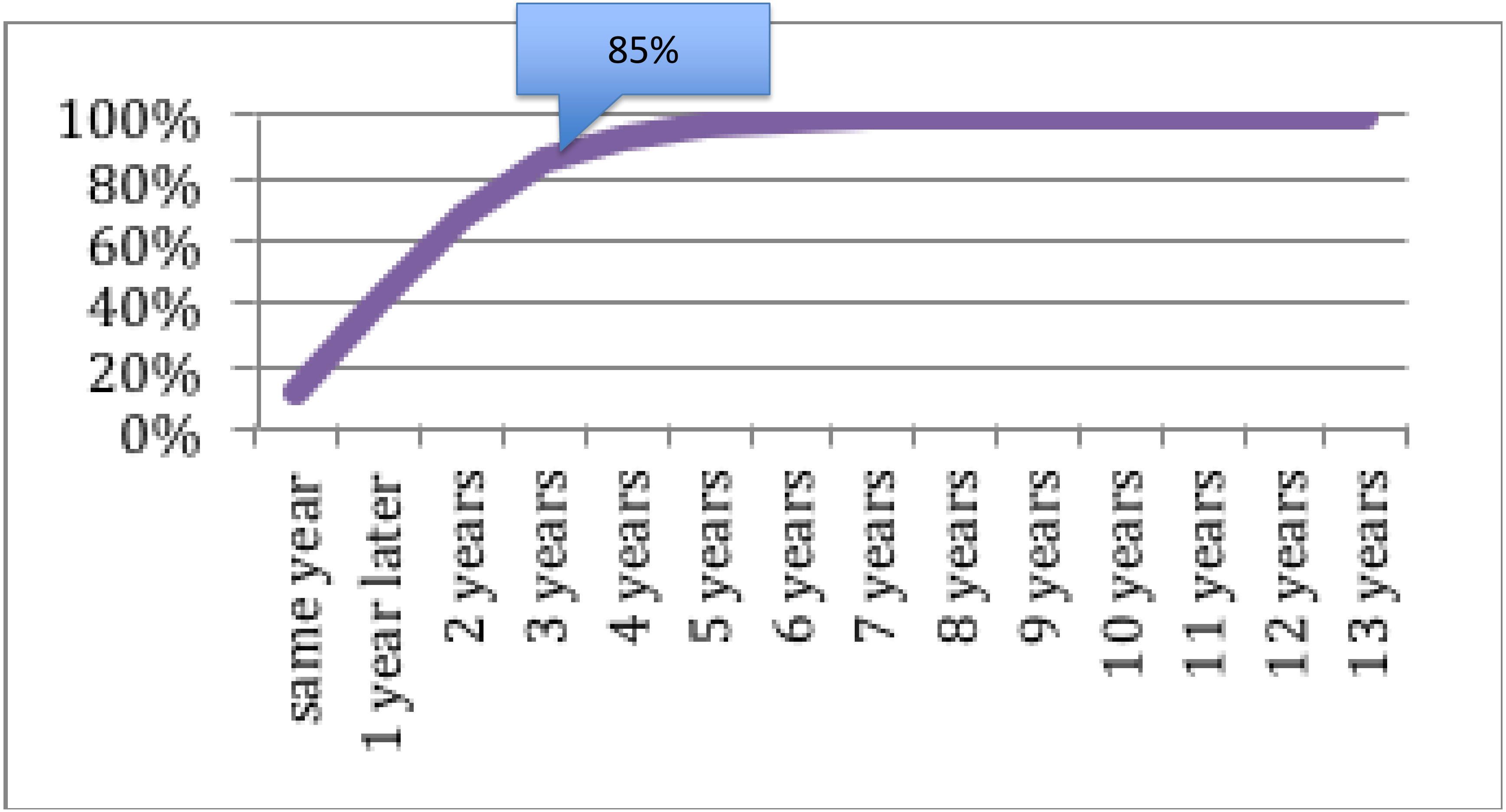


Used Using	acl	acmtslp	alta	anlp	cath	cl	coling	conll	csal	eacl	emnlp	hlt	icassps	ijcnlp	inlg	isca	jep	lre	lrec	ltc	modulad	mts	muc	naacl	paclic	ranlp	sem	speechc	tacl	tal	taln	taslp	tipster	trec	Total used	Total using	Difference	
acl	1	0	0	0	1	1	2	2	0	0	4	3	0	3	0	2	0	0	1	1	0	0	1	1	1	1	3	0	0	0	0	0	0	0	28	7	21	acl
acmtslp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	acmtslp
alta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	alta
anlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	anlp
cath	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	-2	cath
cl	0	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	4	0	0	1	0	1	2	0	0	0	0	0	0	0	0	0	12	5	7	cl
coling	0	0	0	0	1	0	0	0	0	0	0	2	1	1	0	2	0	0	2	0	0	0	1	1	1	0	2	0	0	0	0	1	0	0	15	7	8	coling
conll	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	5	-2	conll
csal	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	3	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	7	6	1	csal
eacl	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	eacl
emnlp	0	0	0	0	0	2	0	2	0	1	1	2	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	2	13	15	-2	emnlp
hlt	2	0	0	0	0	1	0	1	1	0	2	1	1	1	0	2	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	2	17	17	0	hlt
icassps	0	0	0	0	0	0	0	0	1	0	1	2	3	0	0	32	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	5	0	0	48	37	11	icassps
ijcnlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2	9	-7	ijcnlp
inlg	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	inlg
isca	0	0	0	0	0	1	1	0	1	0	0	1	18	1	0	7	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	3	0	0	36	70	-34	isca
jep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	jep
lre	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	lre
lrec	0	0	0	0	0	0	0	0	1	0	2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	8	8	0	lrec
ltc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	-4	ltc
modulad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	modulad
mts	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4	3	1	mts
muc	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	3	0	muc
naacl	1	0	0	0	0	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	9	10	-1	naacl
paclic	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	10	-8	paclic
ranlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	-3	ranlp
sem	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	7	-4	sem
speechc	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4	5	-1	speechc
tacl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	tacl
tal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	tal
taln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	taln
taslp	0	0	0	0	0	0	0	0	1	1	0	0	10	0	0	16	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	30	10	20	taslp
tipster	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	2	0	tipster
trec	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	13	13	0	trec
Total using	7	0	0	0	2	5	7	5	6	2	15	17	37	9	0	70	0	1	8	4	0	3	3	10	10	3	7	5	0	0	0	10	2	13	261	261	0	

Time Delay Publication / Reuse (1.22 years on average)



Time Delay Publication in Conferences / Reuse in Journals (2.07 years on average)



Self-Plagiarism or *Fair Use*?

(Pamela Samuelson, Comm. of ACM 1994)

- Acceptable if:
  - The previous work must be restated to lay the groundwork for a new contribution in the second work,
  - Portions of the previous work must be repeated to deal with new evidence or arguments,
  - The audience for each work is so different that publishing the same work in different places is necessary to get the message out,
  - The authors think they said it so well the first time that it makes no sense to say it differently a second time.
- 30% as an upper limit in the reuse of parts of a previously published paper.
    - Only 1.3% of NLP4NLP papers go beyond this limit

Plagiarism: *Right to Quote*

- "National legislations usually embody the *Berne convention limits* in one or more of the following requirements:
  - the cited paragraphs are within a reasonable limit,
    - <= 10% of the copied / copying papers in France / Canada
    - Only 0.05% of NLP4NLP papers go beyond this limit
  - the cited paragraphs are clearly marked as quotations and fully referenced,
  - the resulting new work is not just a collection of quotations, but constitutes a fully original work in itself".
- the copied paragraphs must have a function in the goal of the copying paper.

Conclusions

- Produce results on the study of copy & paste operations on corpora of NLP archives of very large size, using NLP methods
  - Large number of pairwise comparisons (65,000*65,000), which still represents a practical computing limitation.
- Self-reuse and self-plagiarism are common practices (18%)
  - 40% happen on same year (no way to detect beforehand)
  - No quote of source paper in 60% of the cases (75% if same year)
  - Natural flow from conferences to journals
  - Current tendency for “salami-slicing” publications caused by the publish-and-perish demand
- Plagiarism very uncommon in the NLP community (<0.05%)
- Ethically acceptable if principles are respected

Further developments

- Process "*rogeting*": replacing words with synonymous alternatives
- Study the position and rhetorical structure of the copy & paste in order to identify and justify their function.
- Explore whether copy & paste is more common for non native-English speakers
  - publish first in their native language, then in English in an international conference or an international journal, in order to broaden their audience

Thank you.