Jak psát vědecké články a nepsat

Pavel Pecina

pecina@ufal.mff.cuni.cz

Seminář ÚFAL

Horní Mísečky, 9.2.2009

Úvod

- Vědecké publikace jsou stěžejním prvkem výzkumu
- Pokud vaše práce neprodukuje publikace, je „zbytečná"

- Cílem vědecké práce je:
  1) Formulovat a ověřovat (testovat) hypotézy
  2) Vyvozovat závěry z těchto testů
  3) Seznámit s těmito závěry ostatní

Úvod (pokračování)

- publikovat výsledky vědeckého výzkumu není „jen tak"
- viz poměr přijatých příspěvků na CL konference:

	ACL	COLING	NAACL	EACL	EMNLP
2007	22	–	24	–	27
2006	23	23	24	20	31
2005	18	–	32	–	32
2004	25	39	26	–	24
2003	20	–	23	–	–

zdroj: aclweb.org

→ šance na přijetí konferenčního příspěvku: 15–35% !

Proč autoři publikují?

1) zpřístupnění výsledků práce (54%)
2) kariérní vyhlídky (20%)
3) financování výzkumu (13%)
4) vlastní ego (9%)
5) patentová ochrana (4%)
6) jiné důvody (5%)

Zdroj: Bryan Coles (ed.) The STM Information System in the UK, BL Report 6123, Royal Society, BL, ALPSP, 1993

Priority autorů vs. priority čtenářů

- Dostatečná reputace časopisu/konference
- rychlé recenzní řízení
- kvalitní recenze
- další publikační služby

- chtějí <u>publikovat více</u>

- kvalitní a aktuální informace
- pohodlný přístup (snadný, rychlý, levný)

- chtějí <u>číst méně</u>

Zdroj: Elsevier study of 36,000 authors (1999–2002) presented by Michael Mabe at ALPSP Seminar on "Learning from users", 2003

Základní aspekty publikování

1) O čem psát

2) Etické záležitosti

3) Jazyk a styl

4) Struktura článku

5) Výběr časopisu/konference

6) Odeslání článku, recenzní řízení a publikační proces

O čem psát?

1) Slibné téma
   - nevyřešená úloha na aktuální téma (dostatečná motivace)
   - nepracuje na ní mnoho lidí (malá konkurence)
2) Nápad, jak úlohu řešit
   - Realizovatelný (lidské, datové a výpočetní zdroje)
3) Příprava experimentů
   - implementace metod, dostupnost dat, výpočetní kapacita
4) Provedení experimentů a ověření hypotézy
5) Příprava rukopisu
   - Pro konkrétní konferenci/časopis (k danému termínu)

Etické záležitosti

- publikované výsledky musí být původní
- dostatečné citování použité literatury (převzatých výsledků)
- pozor na plagiátorství (včetně sebe-plagiátorství)
- použití převzatých obrázků a grafiky (a jejich úprava)
- oprávněnost použití dat a programů (licence)
- koho uvést jako autora/spoluautora?
- uvedení zdrojů financování (grantů)
- poděkování těm, kteří si to zaslouží

  (a nevešli se mezi autory :=)

Jazyk

- **Jazyk:** angličtina :=)
- **Osoba:** první, plurál (i singulár, jen pokud je jeden autor)
- **Čas:** minulý nebo přítomný
  - Věc osobní preference, minulý čas spíše pro popis provedených experimentů, obecné závěry spíše v přítomném
  - Přítomný čas naznačuje stále probíhající výzkum (nedokončený)
  - Použití musí být konzistentní (alespoň v rámci odstavce)
- **Rod:** činný (zbytečné použití pasiva ztěžuje pochopení)
- **Věty:** krátké (max 15–20 slov), smysluplné, jednoznačné
- **Odstavce:** krátké, jeden odstavec – jedna myšlenka
- **Zkratky, akronymy:** při prvním použití rozepsat

Styl

- cílem je stručné, jasné a výstižné sdělení
- požadovaný formát se liší dle časopisu/konference
- text lze psát:
  a) dle stylu pro konkrétní časopis/konferenci
  b) bez stylu a následně jej upravit dle spec. požadavků
- před odesláním:
  a) vhodná kontrola jazyka rodilým mluvčím
  b) nutná kontrola obsahu např. kolegou, školitelem, ...

Struktura článku

- Víceméně povinná
- Ustálená, ověřená, vyvíjená po stovky let

1) Název
2) Autoři a jejich afilace
3) Abstrakt
4) Klíčová slova*
5) Úvod
6) Metody a data
7) Experimenty a výsledky
8) Diskuse
9) Závěr
10) Oznámení*
11) Reference
12) Přílohy*

*volitelné

Pohled čtenáře: tři/čtyři úrovně čtení

1) Název článku
   - v programu konference/obsahu sborníku, časopisu

2) Název + abstrakt
   - rozhoduje o tom, zda čtenář přečte článek celý

3) Název + abstrakt + celý text
   - téma článku čtenáře zajímá
   - cílem je hlubší porozumění

(4) Název + abstrakt + celý text <u>několikrát</u>
   - kompletní porozumění, např. pro replikaci experimentů

Název článku

- Jasně a přesně vystihuje obsah článku
- Délka max 1–2 řádky
- Důležitý pro indexovací systémy
- Nedoporučuje se použití zkratek, akronymů, žargonu
- Je „reklamou“ na celý článek
- Měl by upoutat (přilákat účastníky konference na prezentaci, čtenáře časopisu k přečtení článku)
- Může být gramaticky pestřejší (např. otázka)

Seznam autorů a jejich afilace

- Obsahuje

  buď: ty, kteří intelektuálně přispěli k výzkumu

  nebo: ty, kteří budou veřejně obhajovat jeho výsledky

- Pořadí autorů

  buď: dle zásluh (často se rozlišuje jen první autor)

  nebo: dle abecedy (ev. označen korespondenční autor)

- Jména autorů by měla být konzistentní

  (tvary jmen, uvedení druhého jména, apod.)

- Afilace dle požadavků (např. včetně kompletní adresy)

- Uvedení emailových adres již standardem

Abstrakt

- Stručné (150–300 slov) shrnutí celého článku: (problém, metody, výsledky, závěr)
- Slouží čtenáři k rozhodnutí, zda číst celý článek
- Na jeho základě se recenzenti rozhodují, zda přijmou článek k recenzi
- Většinou se tvoří až po dokončení hlavního textu (lépe potom reflektuje obsah článku)
- Bez zkratek, akronymů, citací, tabulek, obrázků, odrážek
- Na konferencích v tištěné podobě často pouze abstrakt (celý text jen elektronicky na CD)

Klíčová slova

- 2–4 hesla z tezauru, případně volně vytvořené
- Specifikují předmět výzkumu a hlavní téma článku
- Důležité pro indexaci a vyhledávání
- V některých časopisech/sbornících nevyžadována

Úvod

- Jasně specifikujte cíle práce:
  1) řešenou úlohu (problém)
  2) její kontext, využití, důležitost
  3) motivaci pro celou práci
- Uveďte odkazy na ostatní relevantní práce
- Srovnejte metody s již publikovanými přístupy
- Jaké jsou otázky, na které hledáme odpověď, a hypotézy, které se snažíme ověřit
- Závěrem stručný popis navrhovaného přístupu a provedených experimentů

Metody

- Popis metod musí být natolik detailní, aby čtenáři umožňoval replikovat všechny experimenty
- Předpokládejte úroveň čtenářových základních znalostí stejnou jako je vaše
- Dílčí kroky a procedury nejlépe v chronologickém pořadí
- Detailní popis metodologie a způsob evaluace, nebo odkaz na již publikovanou práci, preferovány standardní (obecně uznávané) přístupy
- Všechna rozhodnutí nutno zdůvodnit (inženýrský vs. vědecký přístup)

Data

a) Standardní datové sady

- Odkaz na publikaci/organizaci/akci, při které data vznikla (LDC/ELDA, NIST/CLEF/ostatní „shared task" evaluace)

b) Vlastní data

- Motivace pro jejich vytvoření/použití
- Přesný popis, jak data vznikla, jejich statistiky (velikost)
- Zdroj dat, předzpracování, detaily anotace, zdůvodnění
- Vyjádření ke kvalitě dat (např. chyby zpracování)
- Dostupnost dat pro ostatní (licence, kde získat)
- Odhad mezianotátorské shody !

Výsledky

- Objektivně prezentované výsledky experimentů
- Zatím bez jejich interpretace
- Vhodné použít přehledné tabulky a obrázky

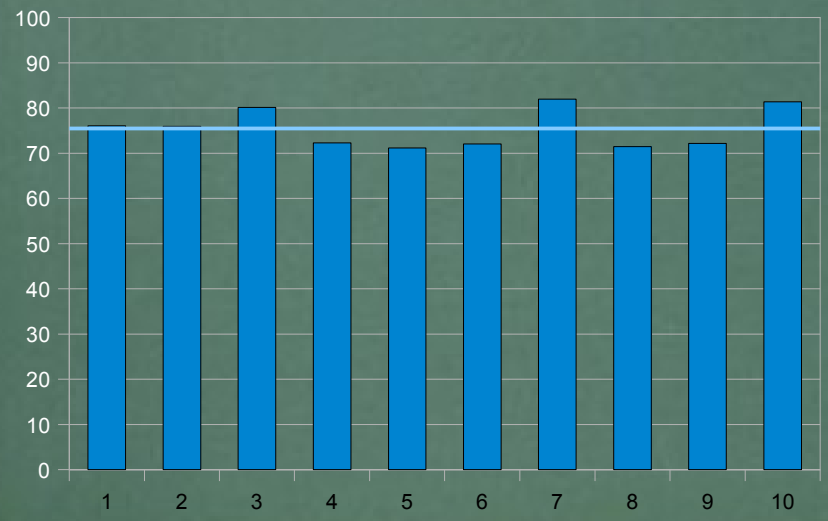
- Specifikujte „baseline" experimentů
  (triviální i tu, kterou chcete pokořit :=)
- Uveďte také horní hranici výsledku
  (kam a až to lze dotáhnout, např. dle mezianotátorské shody)
- Ověřte statistickou signifikanci výsledků

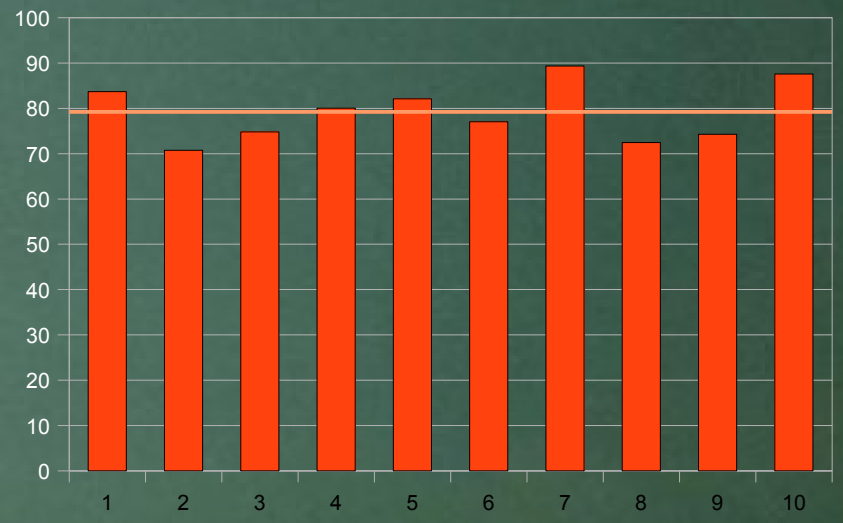
Tabulky a obrázky

- Tabulka prezentuje čísla, obrázek vztahy a tendence
- Neprezentujte stejná data tabulkou i obrázkem
- Nezapomínejte na jednotky, stupnice, měřítka
- Neopakujte informace z popisků v textu článku
- Na každou tabulku/obrázek nutno v textu odkázat

Testy signifikance

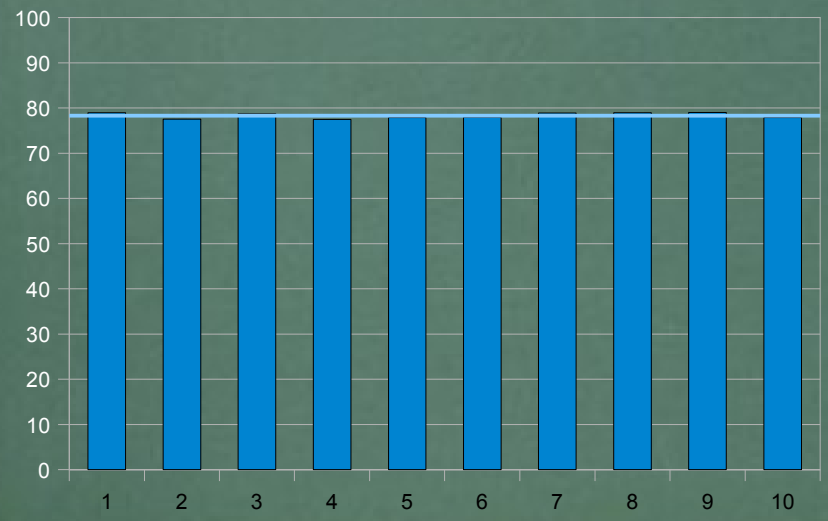
- Každou evaluační míru (accuracy, precision/recall, BLEU) nutno chápat jako odhad získaný na nějakém vzorku dat
- Jako takový má nějaký rozptyl/chybu, kterou je nutno uvažovat při porovnávání výsledků dvou metod/systémů
- **Příklad:** A = 75% B = 80% A < B nesignifikantní rozdíl

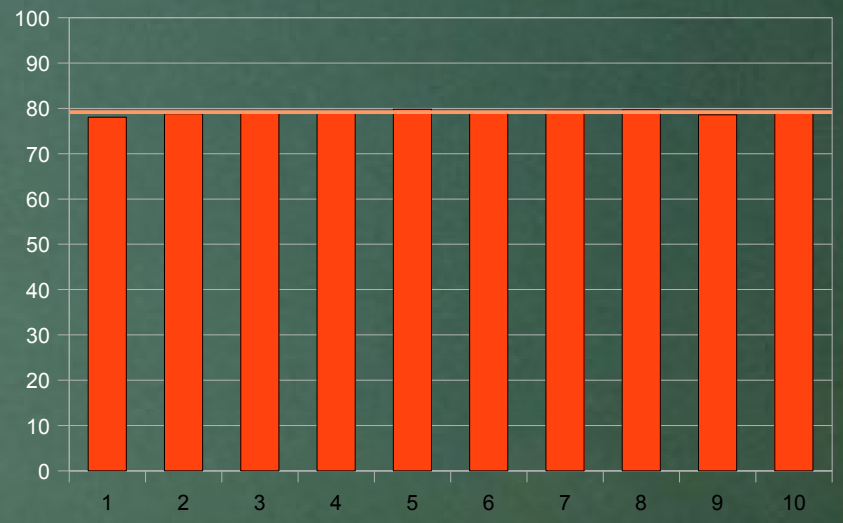




Testy signifikance

- Každou evaluační míru (accuracy, precision/recall, BLEU) nutno chápat jako odhad získaný na nějakém vzorku dat
- Jako takový má nějaký rozptyl/chybu, kterou je nutno uvažovat při porovnávání výsledků dvou metod/systémů
- **Příklad:** A = 77% B = 78% A<B signifikantní rozdíl





Diskuse

- Interpretace získaných výsledků
- Vždy v kontextu ostatních publikovaných prací
- Odpovězte na otázky položené v úvodu článku
- Vyjádřete se k hypotézám (potvrzení, zamítnutí)
- Veškeré závěry musí přímo plynout z dosažených výsledků
- Nedějte příliš odvážné ani příliš obecné závěry
- Diskutujte případná omezení vašich experimentů
- Můžete spekulovat o překvapivých výsledcích, obzvlášť pokud se liší od již publikovaných zjištění

Závěr

- Shrnutí hlavních výsledků práce v kontextu celé problematiky (oblasti) a dříve dosažených výsledků
- Splnily výsledky vaše očekávání, potvrdily se hypotézy, odpověděli jste na položené otázky?
- V jakém vztahu jsou výsledky k již publikovaným zjištěním?
- Jak přispěla vaše práce ke zkoumání problematiky a poznání v dané oblasti?
- Jaké budou/by měly být další kroky výzkumu?

Oznámení

- Uvedení zdroje financování výzkumu
  (instituce, číslo projektu/grantu)
- Poděkování ostatním, kdo přispěl k práci/článku apod.
  (anotátoři, programátoři, recenzenti, manželky, rodiče :=)
- Dostupnost dat, kódu (pod jakou licencí?)

Citovaná literatura

- Citujte příslušný zdroj, vždy když:
  a) se zmíníte o již publikované skutečnosti
  b) uvedete nějakou informaci, která nevyplývá z vašich experimentu nebo nepatří do obecných znalostí
- Doslovné citace (v uvozovkách) včetně stránek
- Vyhněte
  - obtížně dostupným referencím
  - „řetězovým citacím“ (citujte originál)
  - zbytečným citacím
    (těm, které nejsou pro práci relevantní nebo důležité)

Jak správně citovat

a)Harvardský styl

- V textu autor a rok publikace: (Pecina,2008), Pecina(2008)
- Bibliografie setříděná dle abecedy

  Pecina Pavel (2008): Lexical Association Measures: Collocation Extraction, Ph.D. thesis, Charles University in Prague, Prague, Czech Republic

b)Vancouverský styl

- V textu posloupnost čísel: [1], Pecina [1]
- Bibliografie setříděná dle pořadí výskytu v textu

  [1] Pecina Pavel: Lexical Association Measures: Collocation Extraction, Ph.D. thesis, Charles University in Prague, Prague, Czech Republic, 2008

Přílohy

- Místo pro uvedení dodatečných informací
  - rozsáhlejší tabulky
  - detailnější/kompletnější výsledky
  - ukázky dat a výstupů
  - apod.
- Do příloh přesuňte vše, co by jinak narušovalo strukturu článku a zhoršovalo tak jeho přehlednost

Výběr časopisu/konference

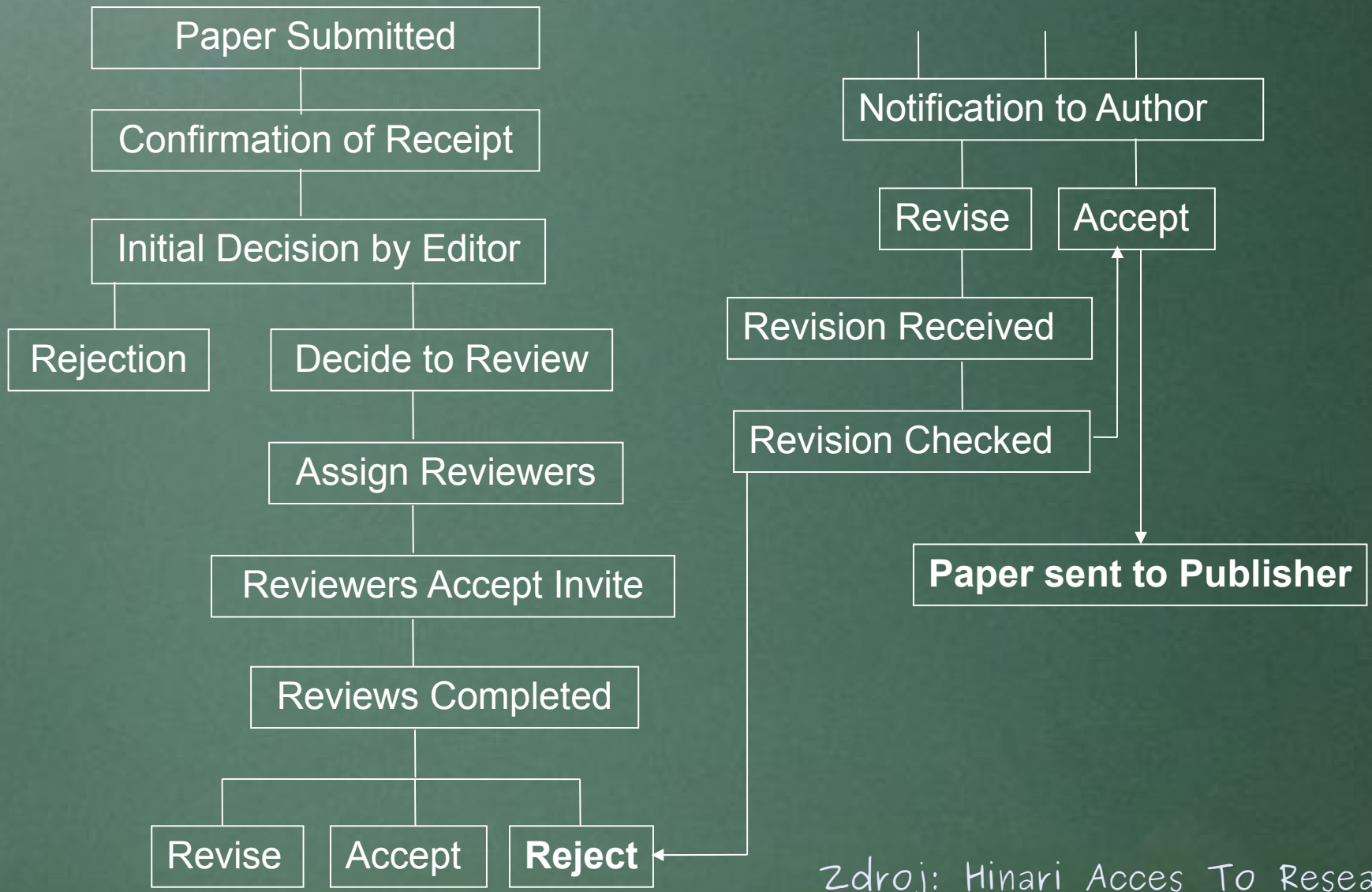
- Zaměření („aims and scope")
- Cílové publikum
- Způsob přístupu k článkům (on-line, zdrarma)
- Reputace/impact factor
- Šance na přijetí
- Délka recenzního a publikačního řízení
- Maximální povolená délka článku
- Barevný tisk :=)

Odeslání článku

- Pečlivě vyberte časopis/konferenci
- Ujistěte se, že článek splňuje požadavky na styl
- Zkontrolujte správnost číselných výsledků, tabulky, grafy
- Konferenční příspěvky anonymizované
- Pečlivě vybírejte tématickou oblast
  - důležité pro výběr recenzentů
  - uveďte 1–2 oblasti, pro které má článek největší přínos
- Uveďte jiné konference, kam byl příspěvek zaslán

  (u článků pro časopis se multiple-submissions nepovolují)
- Nenechávejte odeslání na poslední chvíli :-)

Recenzní řízení



Zdroj: Hinari Acces To Research

Tipy

- Seznam všech CL konferencí (včetně termínů)
  - http://www.cs.rochester.edu/~tetreaul/conferences.html
- Kde hledat relevantní literaturu? On-line:
  - http://www.google.com
  - http://scholar.google.com
  - http://citeseer.ist.psu.edu/
  - http://www.aclweb.org/anthology-index/
  - http://www.springerlink.com
  - http://www.acm.org
- Učte se, jak psát, při čtení cizích prací

Triky

- Málo místa?
  - Neměňte font hlavního textu článku
  - Odstraňte zbytečné tabulky a grafy (případně je spojte)
  - Zmenšete font popisů tabulek a obrázků
  - Zmenšete mezery před a po matematických formulích
  - Zmenšete font bibliografie
  - Přesuňte text do poznámek pod čarou (také menší font)
- Málo času?
  - Požádejte o posunutí termínu :-)