Flexible and Efficient Toolbox for Information Retrieval

MIRACLE group

José Miguel Goñi-Menoyo (UPM)
José Carlos González-Cristóbal (UPM-Daedalus)
Julio Villena-Román (UC3M-Daedalus)

Our approach

- New Year's Resolution: work with all languages in CLEF
  - adhoc, image, web, geo, iclef, qa…
- Wish list:
  - Language-dependent stuff
  - Language-independent stuff
  - Versatile combination
  - Fast
  - Simple for non computer scientists
- Not to reinvent the wheel again every year!
- Approach: Toolbox for information retrieval

Agenda

- Toolbox
- 2005 Experiments
- 2005 Results
- 2006 Homework

Toolbox Basics

- Toolbox made of small one-function tools
- Processing as a pipeline (borrowed from Unix):
  - Each tool combination leads to a different run approach
- Shallow I/O interfaces:
  - tools in several programming languages (C/C++, Java, Perl, PHP, Prolog…),
  - with different design approaches, and
  - from different sources (own development, downloading, …)

MIRACLE Tools

- Tokenizer:
  - pattern matching
    - isolate punctuation
    - split sentences, paragraphs, passages
  - identifies some entities
    - compounds, numbers, initials, abbreviations, dates
  - extracts indexing terms
  - own-development (written in Perl) or “outsourced”
- Proper noun extraction
  - Naive algorithm: Uppercase words unless stop-word, stop-clef or verb/adverb
- Stemming: generally “outsourced”
- Transforming tools: lowercase, accents and diacritical characters are normalized, transliteration

More MIRACLE Tools

- Filtering tools:
  - stop-words and stop-clefs
  - phrase pattern filter (for topics)
- Automatic translation issues: “outsourced” to available on-line resources or desktop applications

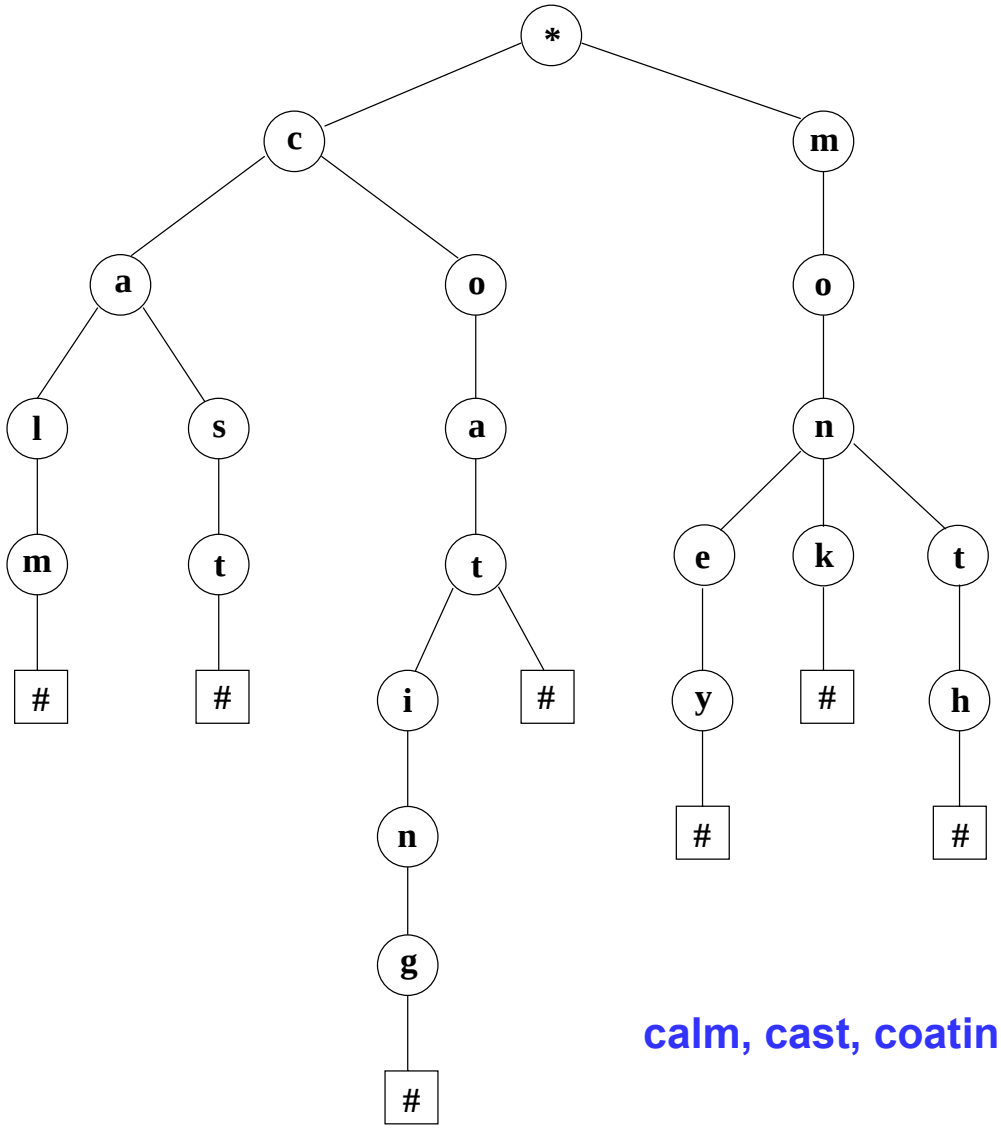
Bultra (En→Bu)	Webtrance (En→Bu)	AutTrans (Es→Fr, Es→Pt)
MoBiCAT (En→Hu)	Systran	BabelFish Altavista
Babylon	FreeTranslation	Google Language Tools
InterTrans	WordLingo	Reverso

- Semantic expansion
  - EuroWordNet
  - own resources for Spanish
- The philosopher's stone: indexing and retrieval system

Indexing and Retrieval System

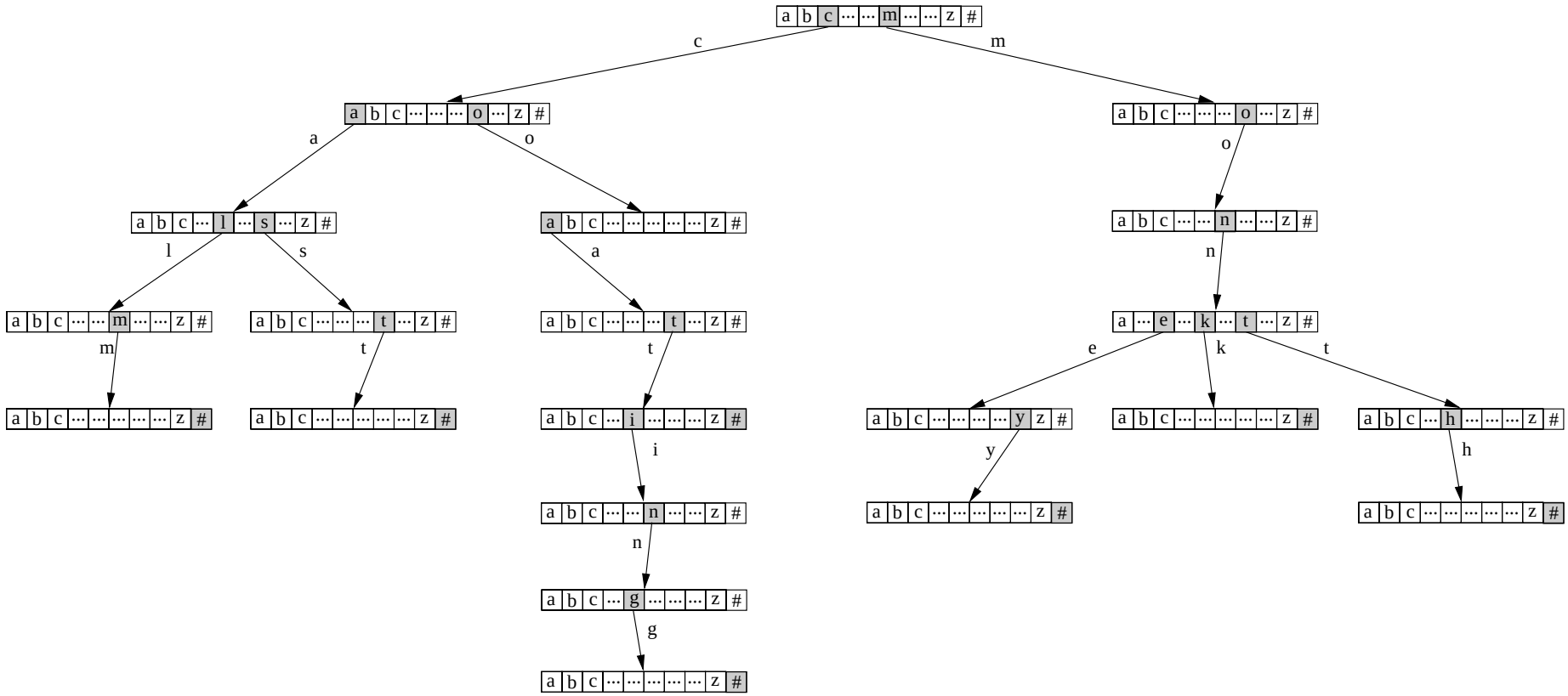
- Implements boolean, vectorial and probabilistic BM25 retrieval models
  - Only BM25 in used in CLEF 2005
  - Only OR operator was used for terms
- Native support for UTF-8 (and others) encodings
  - No transliteration scheme is needed
  - Good results for Bulgarian
- More efficiency achieved than with previous engines
  - Several orders of magnitude in indexing time

Trie-based index



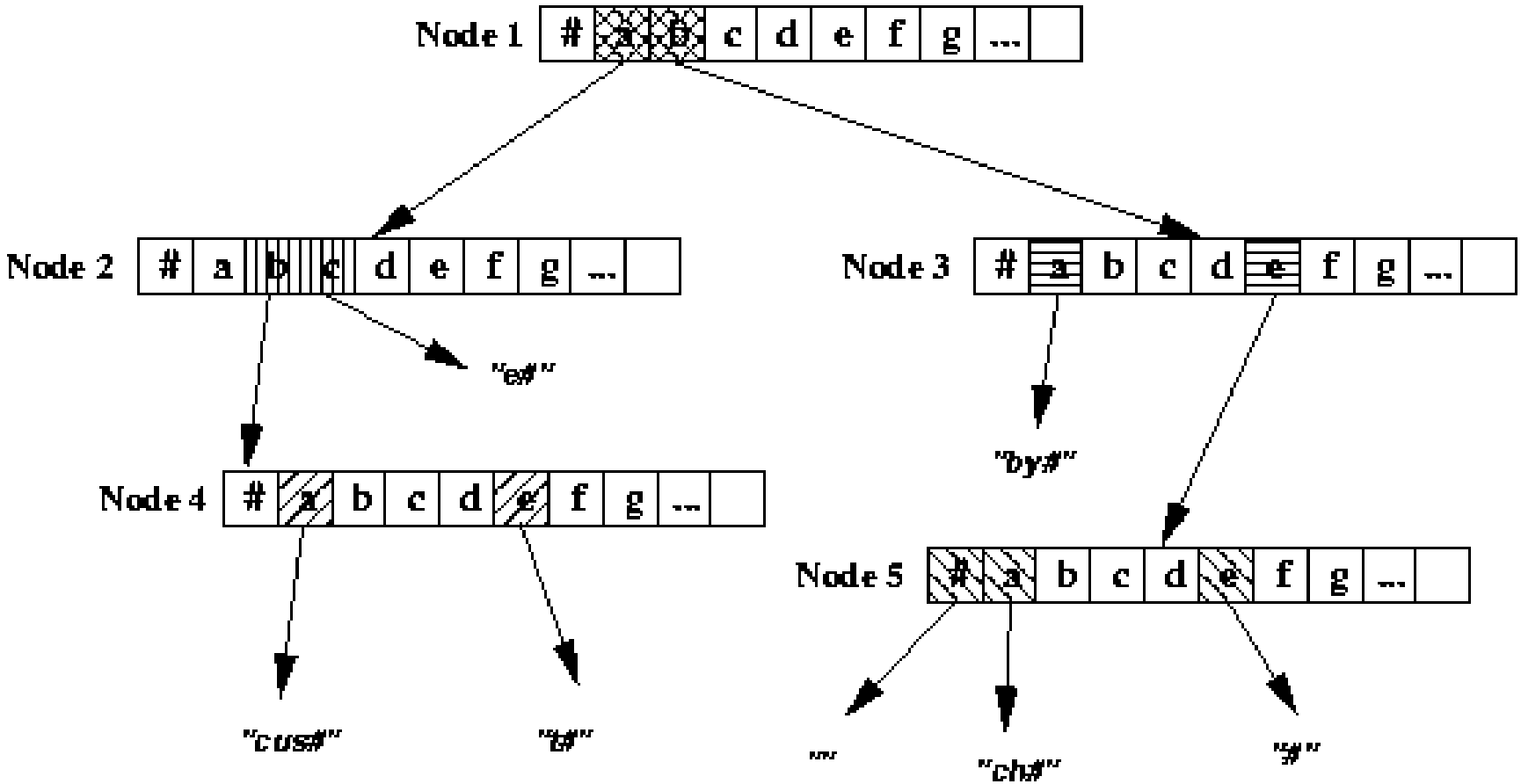
calm, cast, coating, coat, money, monk, month

1st course implementation: linked arrays



calm, cast, coating, coat, money, monk, month

Efficient tries: avoiding empty cells



abacus, abet, ace, baby be, beach, bee

Basic Experiments

- S: Standard sequence (tokenization, filtering, stemming, transformation)
- N: Non stemming

- R: Use of narrative field in topics
- T: Ignore narrative field
- r1: Pseudo-relevance feedback (with 1st retrieved document)
- P: Proper noun extraction (in topics)

→ SR, ST, r1SR, NR, NT, NP

Paragraph indexing

- H: Paragraph indexing
  - *docpars* (document paragraphs) are indexed instead of docs
    - term→ doc1#1, doc69#5 …
  - combination of *docpars* relevance:
    - $rel_N = rel_{mN} + \alpha / n * \sum_{j \neq m} rel_{jN}$

    n=paragraphs retrieved for doc N
    rel_{jN}=relevance of paragraph i of doc N
    m=paragraph with maximum relevance
    α=0.75 (experimental)

→ HR, HT

Combined experiments

- "Democratic system": documents with good score in many experiments are likely to be relevant

- a: Average:
  - Merging of several experiments, adding relevance
- x: WDX - asymmetric combination of two experiments:
  - First (more relevant) non-weighted D documents from run A
  - Rest of documents from run A, with W weight
  - All documents from run B, with X weight
  - Relevance re-sorting
    - Mostly used for combining base runs with proper nouns runs

→ aHRSR, aHTST, xNP01HR1, xNP01r1SR1

Multilingual merging

- Standard approaches for merging:
  - No normalization and relevance re-sorting
  - Standard normalization and relevance re-sorting
  - Min-max normalization and relevance re-sorting

- Miracle approach for merging:
  - The number of docs selected from a collection (language) is proportional to the average relevance of its first N docs (N=1, 10, 50, 125, 250, 1000). Then one of the standard approaches is used

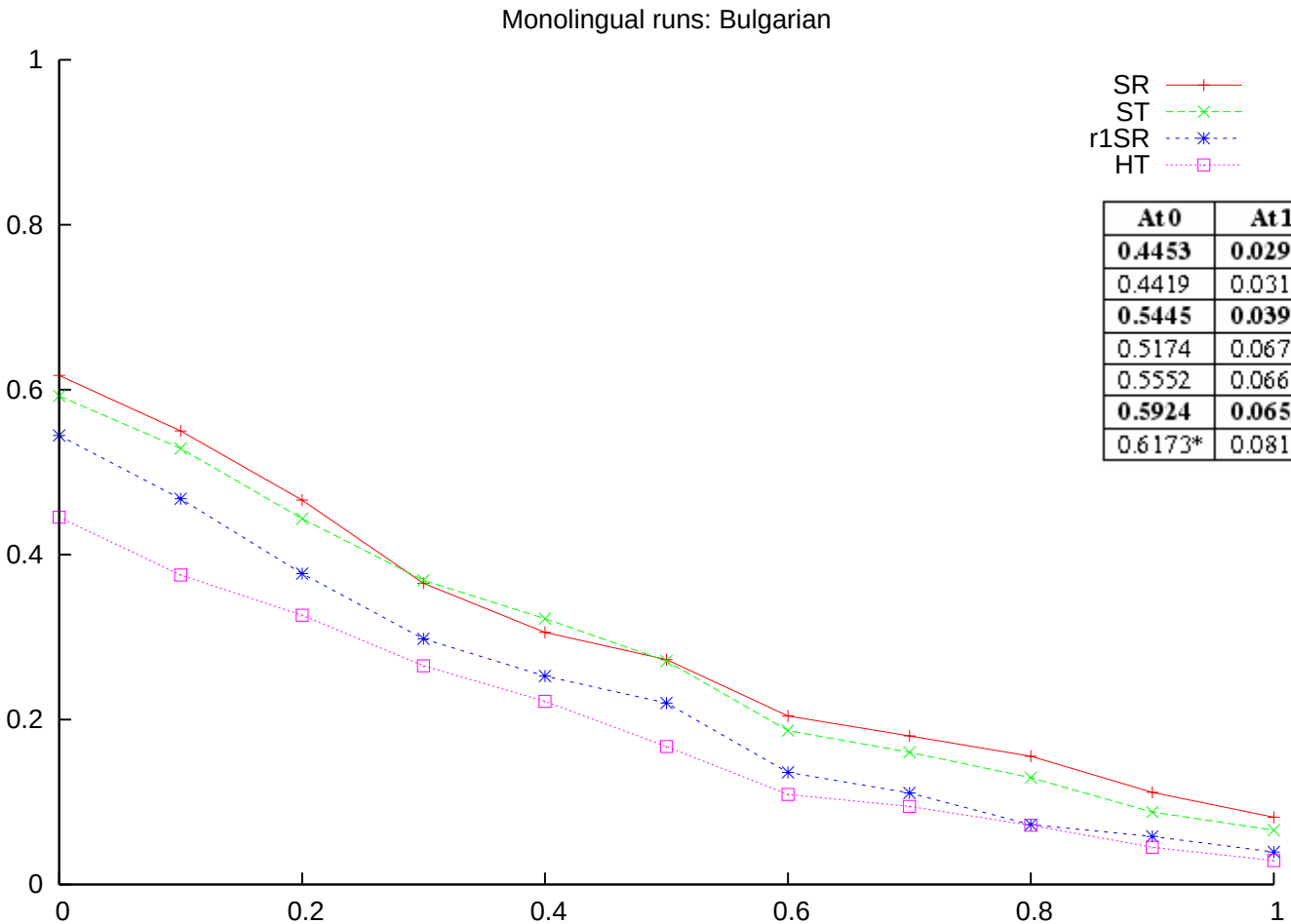
Results

We performed…

… countless experiments!

(just for the adhoc task)

Monolingual Bulgarian

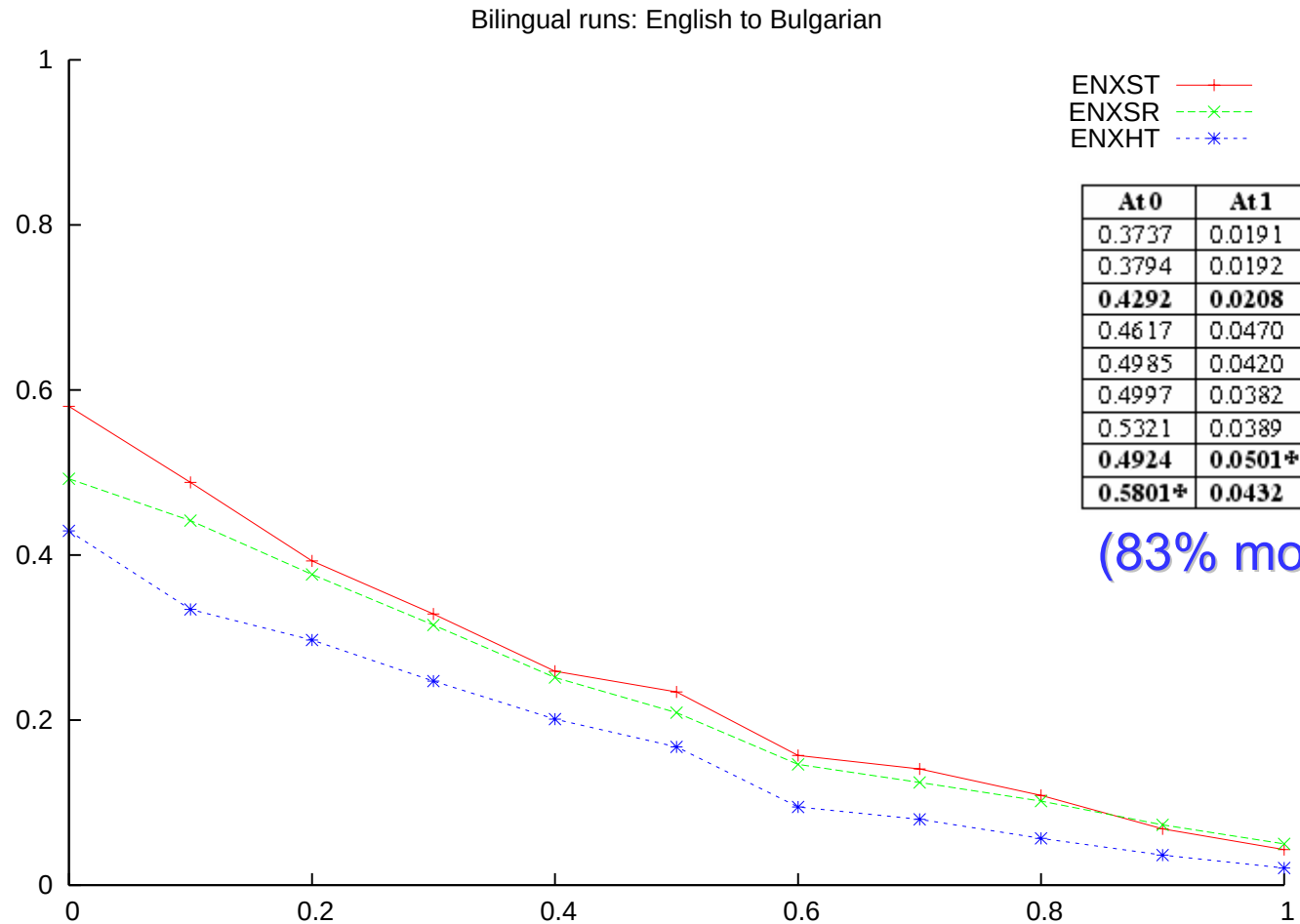


At 0	At 1	Avgp	%	Run
0.4453	**0.0291**	**0.1804**	**-36.01 %**	**HT**
0.4419	0.0315	0.1886	-33.10%	HR
0.5445	**0.0393**	**0.2191**	**-22.28 %**	**r1SR**
0.5174	0.0678	0.2200	-21.96%	NR
0.5552	0.0665	0.2328	-17.42%	NT
0.5924	**0.0659**	**0.2676**	**-5.07 %**	**ST**
0.6173*	0.0816*	0.2819*	-0.00%	SR

Rank: 4th

Stemmer (UTF-8): Neuchâtel

Bilingual English→Bulgarian



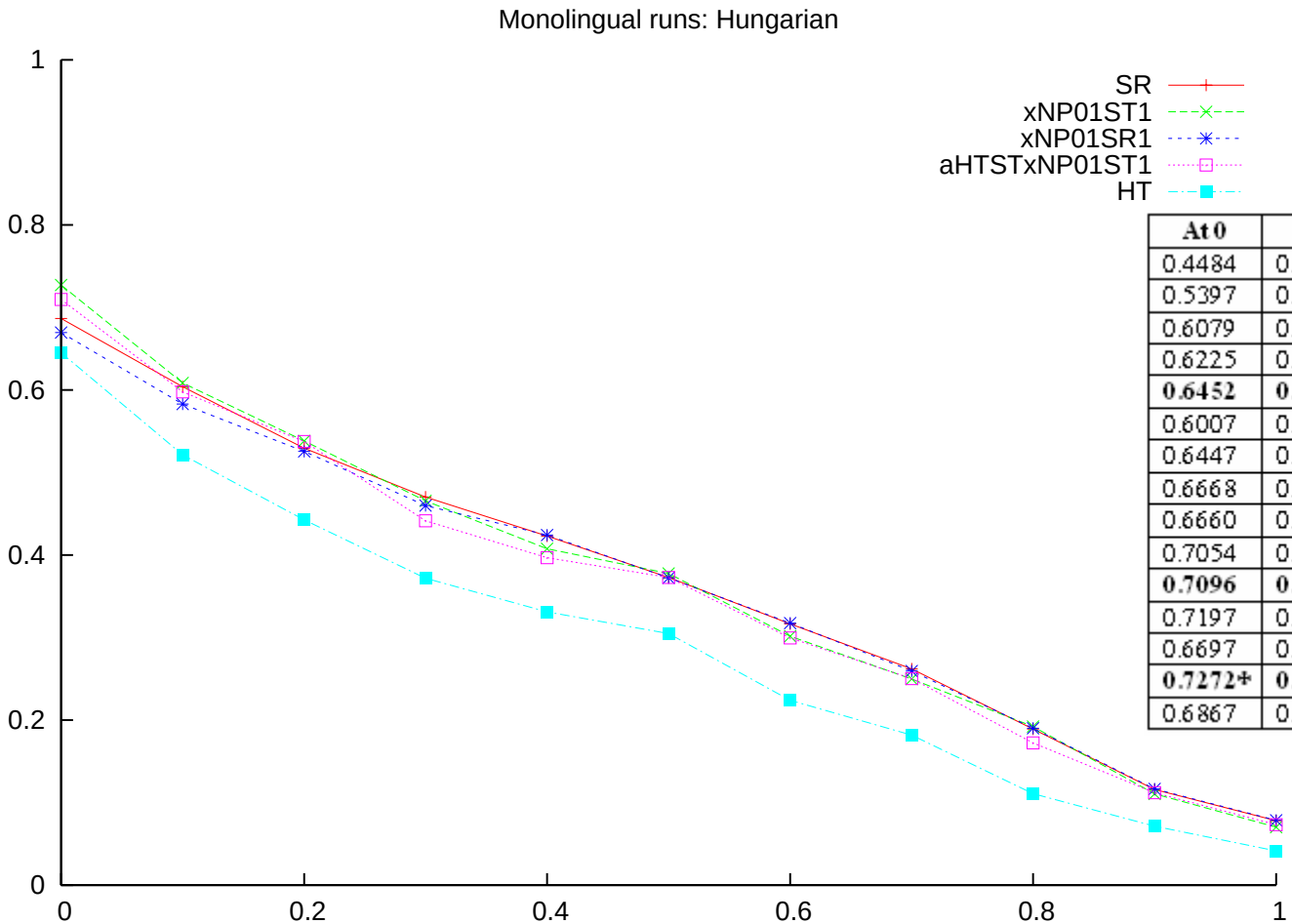
At 0	At 1	Avgp	%	Run
0.3737	0.0191	0.1405	-40.34%	ENBHT
0.3794	0.0192	0.1489	-36.77%	ENWHT
0.4292	**0.0208**	**0.1635**	**-30.57%**	**ENXHT**
0.4617	0.0470	0.1926	-18.22%	ENBSR
0.4985	0.0420	0.2014	-14.48%	ENBST
0.4997	0.0382	0.2112	-10.32%	ENWSR
0.5321	0.0389	0.2132	-9.47%	ENWST
0.4924	**0.0501***	**0.2194**	**-6.84%**	**ENXSR**
0.5801*	**0.0432**	**0.2355***	**-0.00%**	**ENXST**

(83% monolingual)

Rank: 1st

En→Bu: Bultra, Webtrance

Monolingual Hungarian

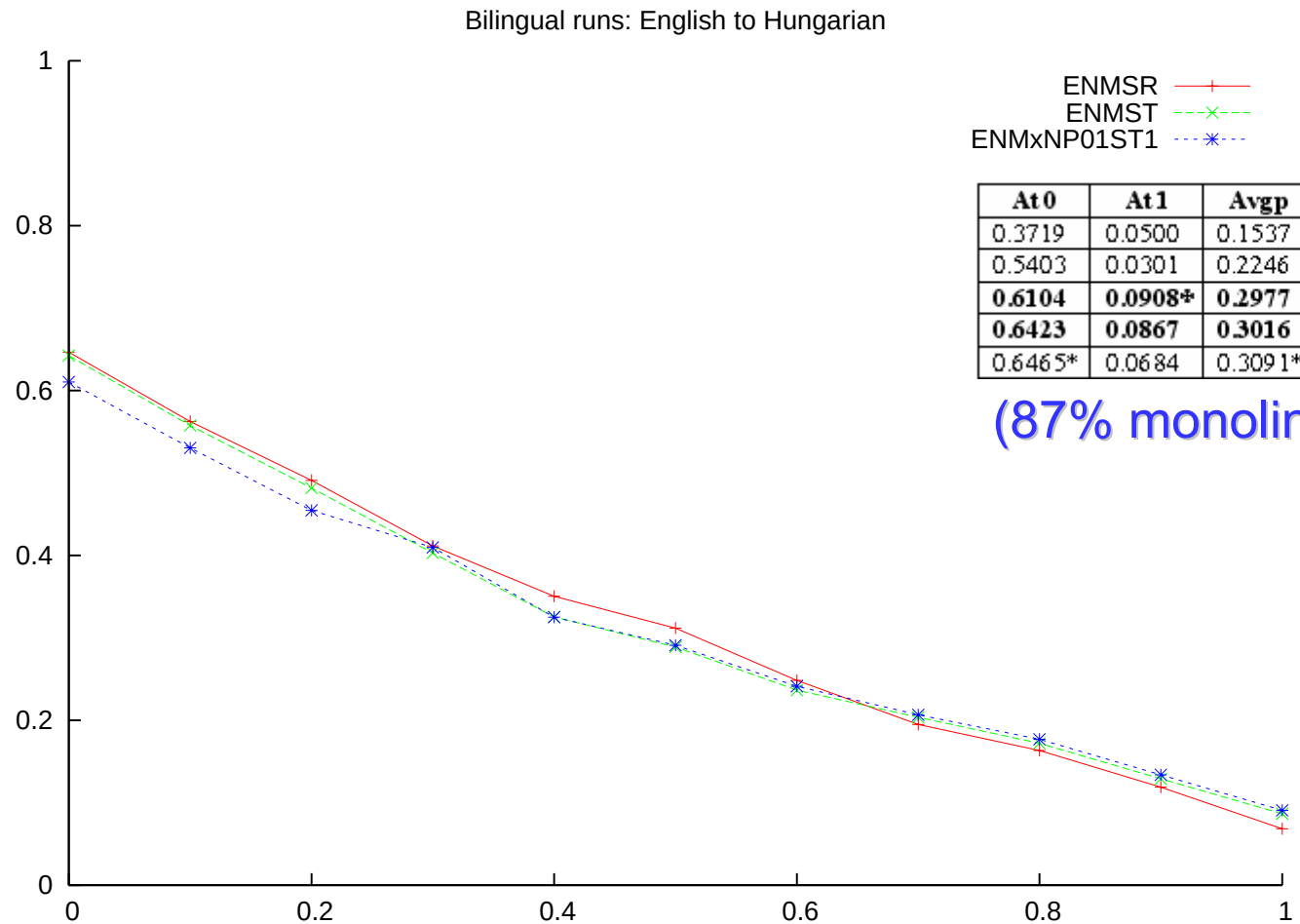


At 0	At 1	Avgp	%	Run
0.4484	0.0534	0.1776	-49.77%	NP
0.5397	0.0578	0.2263	-36.00%	NT
0.6079	0.0531	0.2641	-25.31%	NR
0.6225	0.0421	0.2721	-23.05%	xNP01HT1
0.6452	**0.0414**	**0.2770**	**-21.66%**	**HT**
0.6007	0.0426	0.2777	-21.46%	xNP01HR1
0.6447	0.0413	0.2843	-19.60%	HR
0.6668	0.0562	0.3085	-12.75%	aHTSTxNP01HT1
0.6660	0.0651	0.3266	-7.64%	aHRSRxNP01HR1
0.7054	0.0677	0.3373	-4.61%	aHRSR
0.7096	**0.0733**	**0.3435**	**-2.86%**	**aHTSTxNP01ST1**
0.7197	0.0703	0.3493	-1.22%	ST
0.6697	0.0785*	0.3501	-0.99%	xNP01SR1
0.7272*	**0.0703**	**0.3520**	**-0.45%**	**xNP01ST1**
0.6867	0.0781	0.3536*	-0.00%	SR

Rank: 3rd

Stemmer: Neuchâtel

Bilingual English→Hungarian



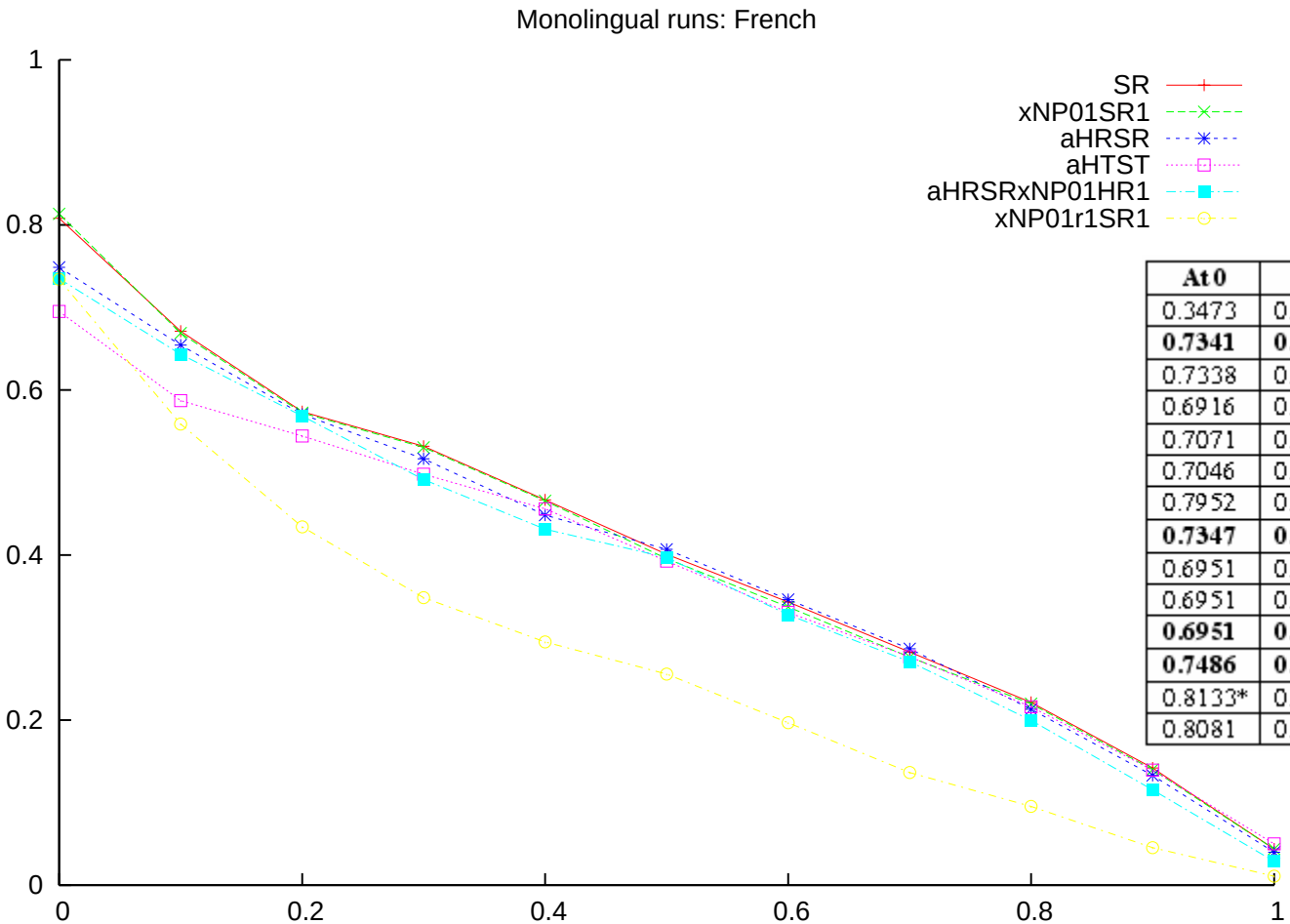
At 0	At 1	Avgp	%	Run
0.3719	0.0500	0.1537	-50.27%	ENMNP
0.5403	0.0301	0.2246	-27.34%	ENMHT
0.6104	**0.0908***	**0.2977**	**-3.69 %**	**ENMxNP01ST1**
0.6423	**0.0867**	**0.3016**	**-2.43 %**	**ENMST**
0.6465*	0.0684	0.3091*	-0.00%	ENMSR

(87% monolingual)

Rank: 1st

En→Hu: MoBiCAT

Monolingual French

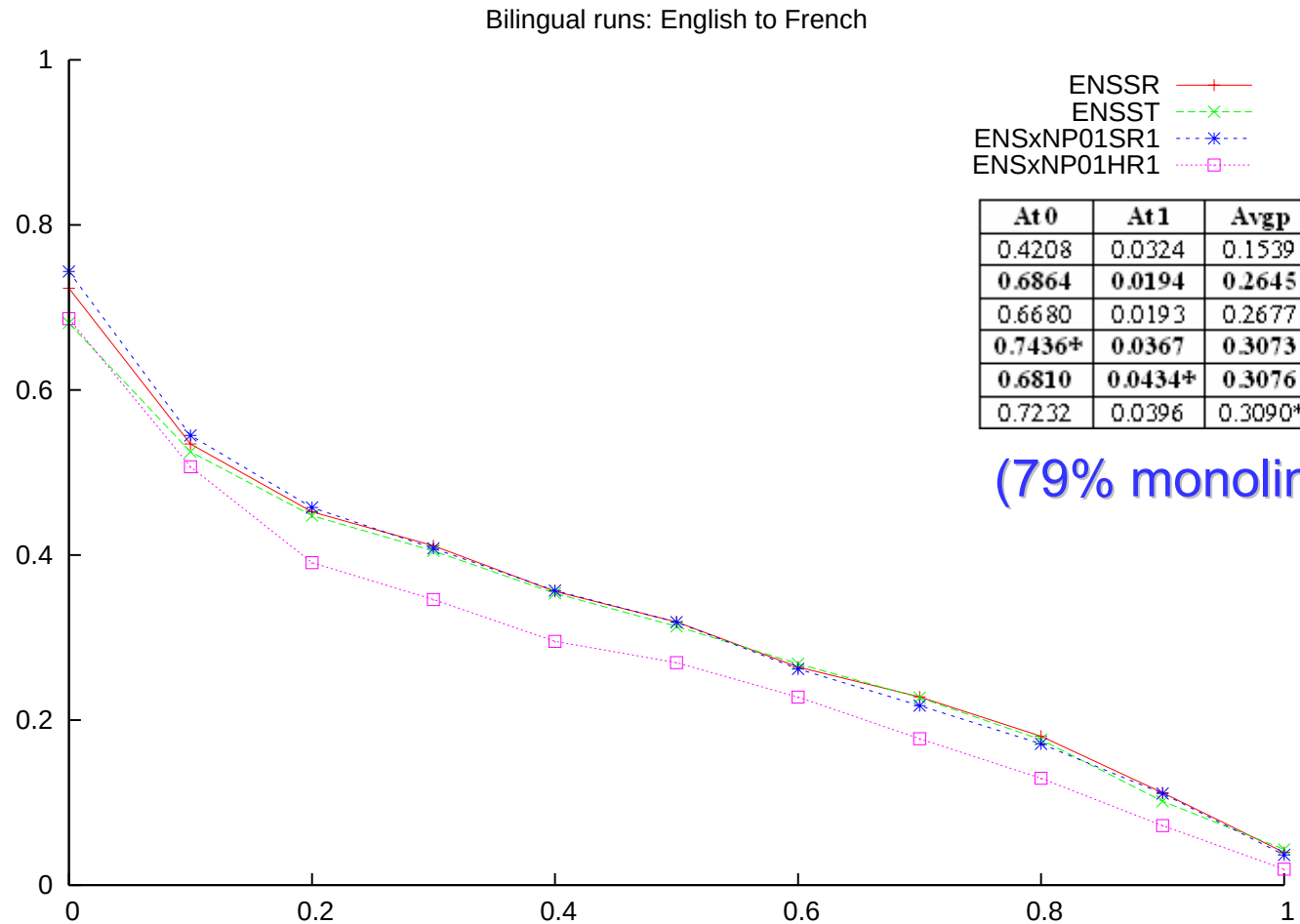


At 0	At 1	Avgp	%	Run
0.3473	0.0180	0.1254	-68.02%	NP
0.7341	**0.0109**	**0.2636**	**-32.77 %**	**xNP01r1SR1**
0.7338	0.0190	0.2713	-30.81%	r1SR
0.6916	0.0227	0.3157	-19.48%	NT
0.7071	0.0252	0.3298	-15.89%	xNP01HR1
0.7046	0.0251	0.3350	-14.56%	HR
0.7952	0.0279	0.3431	-12.50%	NR
0.7347	**0.0289**	**0.3675**	**-6.27 %**	**aHRSRxNP01HR1**
0.6951	0.0501*	0.3692	-5.84%	ST
0.6951	0.0501*	0.3692	-5.84%	HT
0.6951	**0.0501***	**0.3692**	**-5.84 %**	**aHTST**
0.7486	**0.0398**	**0.3833**	**-2.24 %**	**aHRSR**
0.8133*	0.0437	0.3883	-0.97%	xNP01SR1
0.8081	0.0437	0.3921*	-0.00%	SR

Rank: >5th

Stemmer: Snowball

Bilingual English→French



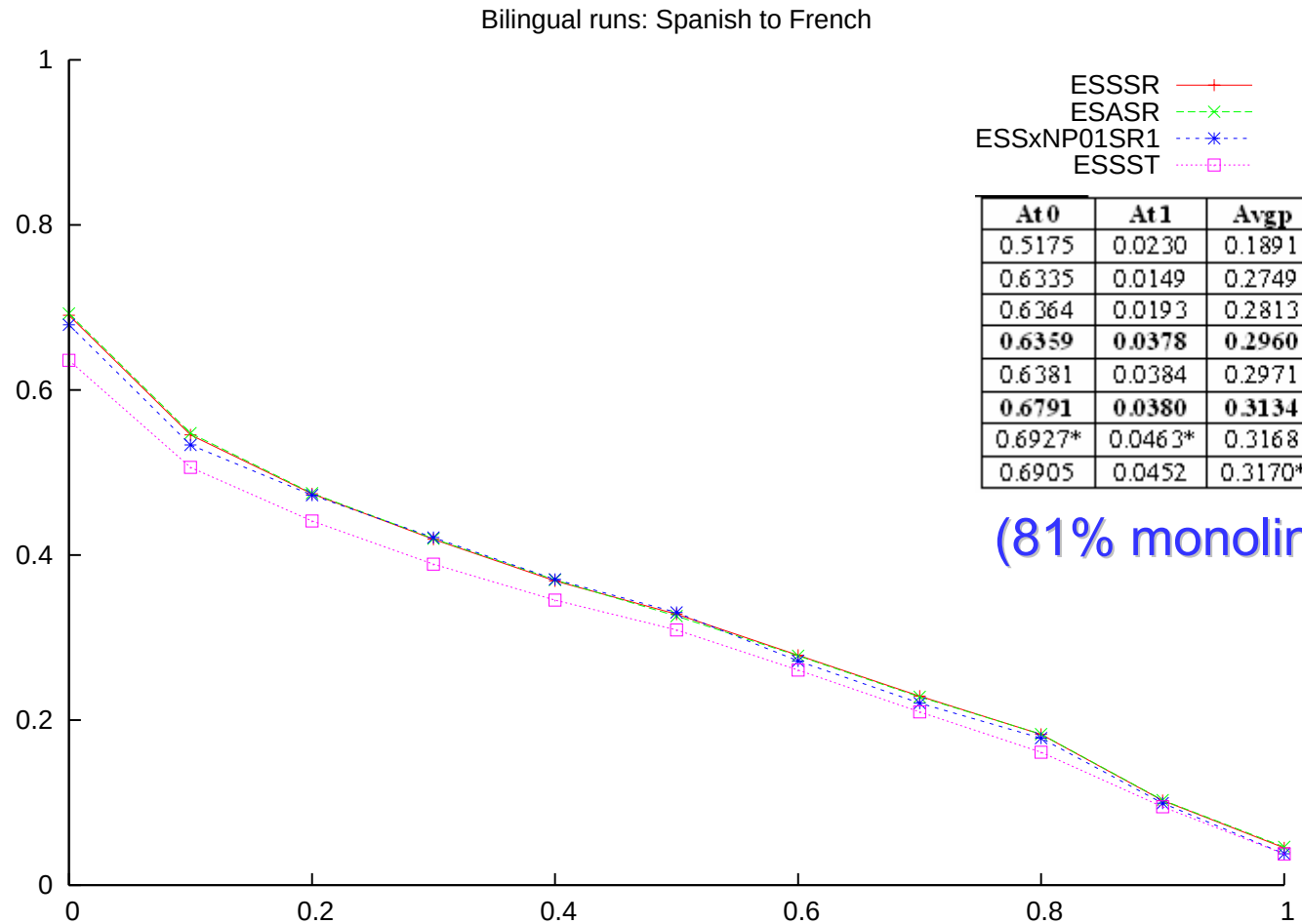
At 0	At 1	Avgp	%	Run
0.4208	0.0324	0.1539	-50.19%	ENSNP
0.6864	**0.0194**	**0.2645**	**-14.40%**	**ENSxNP01HR1**
0.6680	0.0193	0.2677	-13.37%	ENSHR
0.7436*	**0.0367**	**0.3073**	**-0.55%**	**ENSxNP01SR1**
0.6810	**0.0434***	**0.3076**	**-0.45%**	**ENSST**
0.7232	0.0396	0.3090*	-0.00%	ENSSR

(79% monolingual)

Rank: 5th

En→Fr: Systran

Bilingual Spanish→French



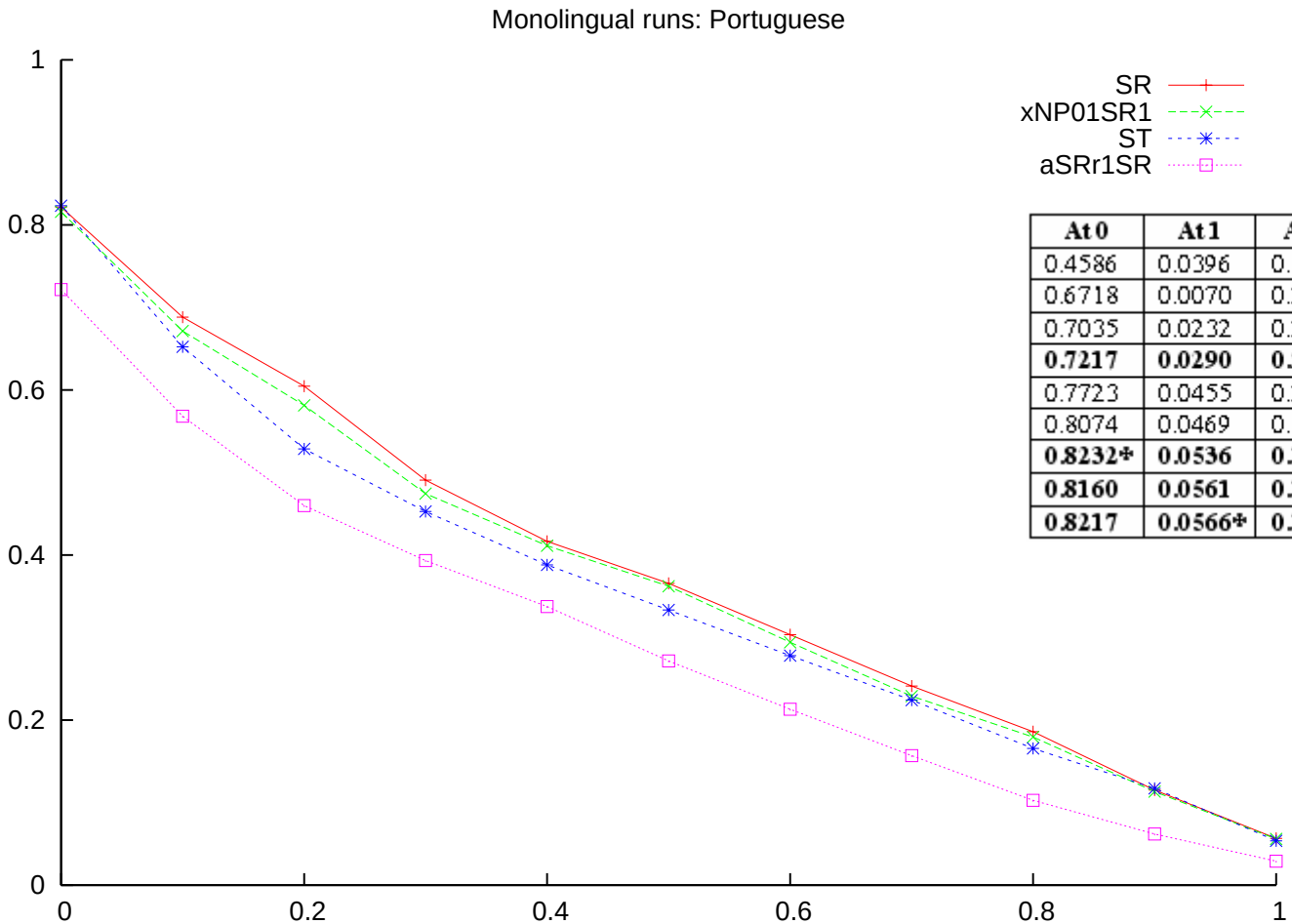
At 0	At 1	Avgp	%	Run
0.5175	0.0230	0.1891	-40.35%	ESSNP
0.6335	0.0149	0.2749	-13.28%	ESSxNP01HR1
0.6364	0.0193	0.2813	-11.26%	ESSHR
0.6359	**0.0378**	**0.2960**	**-6.62%**	**ESSST**
0.6381	0.0384	0.2971	-6.28%	ESAST
0.6791	**0.0380**	**0.3134**	**-1.14%**	**ESSxNP01SR1**
0.6927*	0.0463*	0.3168	-0.06%	ESASR
0.6905	0.0452	0.3170*	-0.00%	ESSSR

(81% monolingual)

(Rank: 5th)

Es→Fr: ATrans, Systran

Monolingual Portuguese

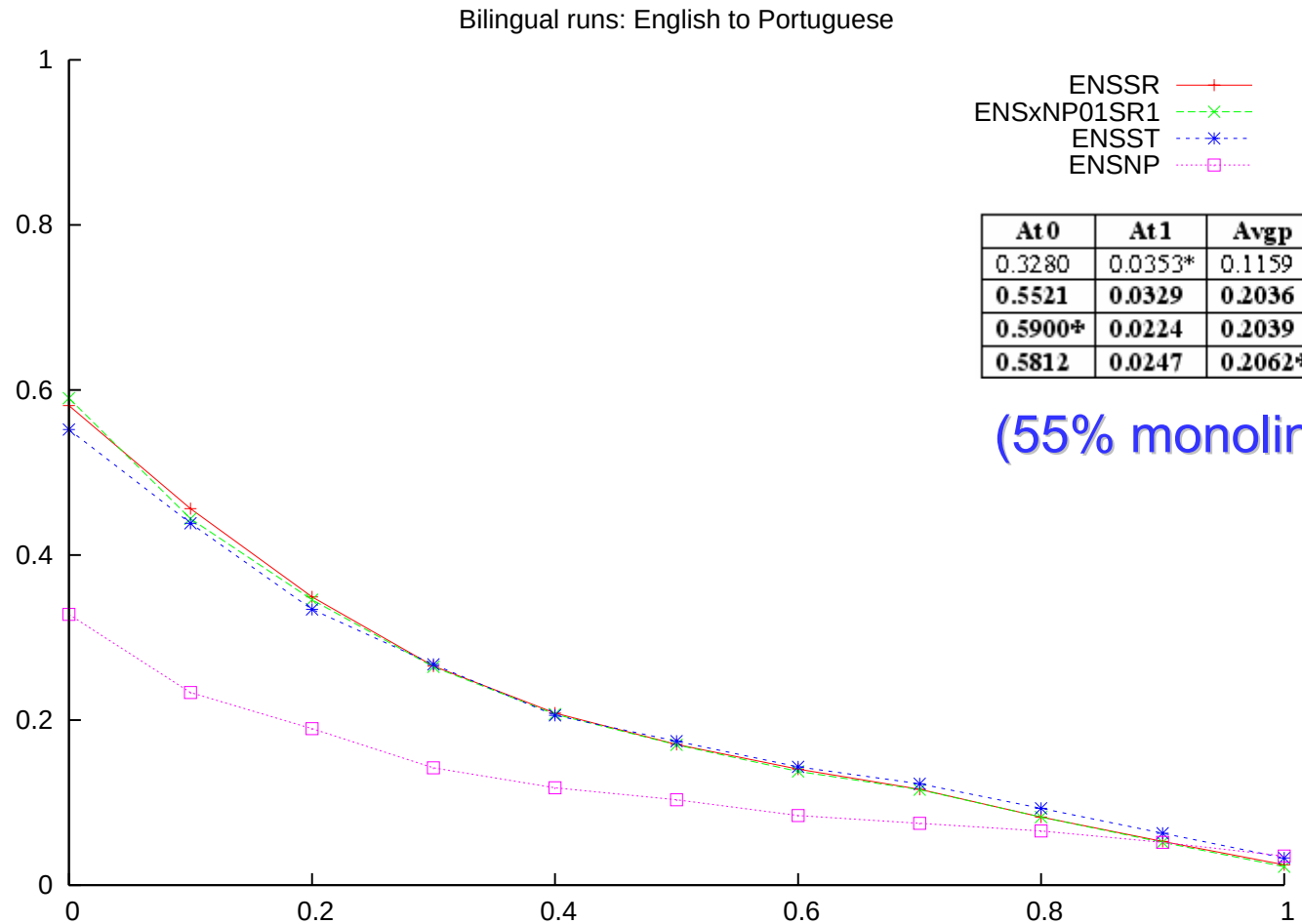


At 0	At 1	Avgp	%	Run
0.4586	0.0396	0.1669	-54.87%	NP
0.6718	0.0070	0.2214	-40.13%	xNP01r1SR1
0.7035	0.0232	0.2358	-36.24%	r1SR
0.7217	**0.0290**	**0.2832**	**-23.42%**	**aSRr1SR**
0.7723	0.0455	0.2957	-20.04%	NT
0.8074	0.0469	0.3198	-13.52%	NR
0.8232*	**0.0536**	**0.3456**	**-6.54%**	**ST**
0.8160	**0.0561**	**0.3628**	**-1.89%**	**xNP01SR1**
0.8217	**0.0566***	**0.3698***	**-0.00%**	**SR**

Rank: >5th (4th)

Stemmer: Snowball

Bilingual English→Portuguese



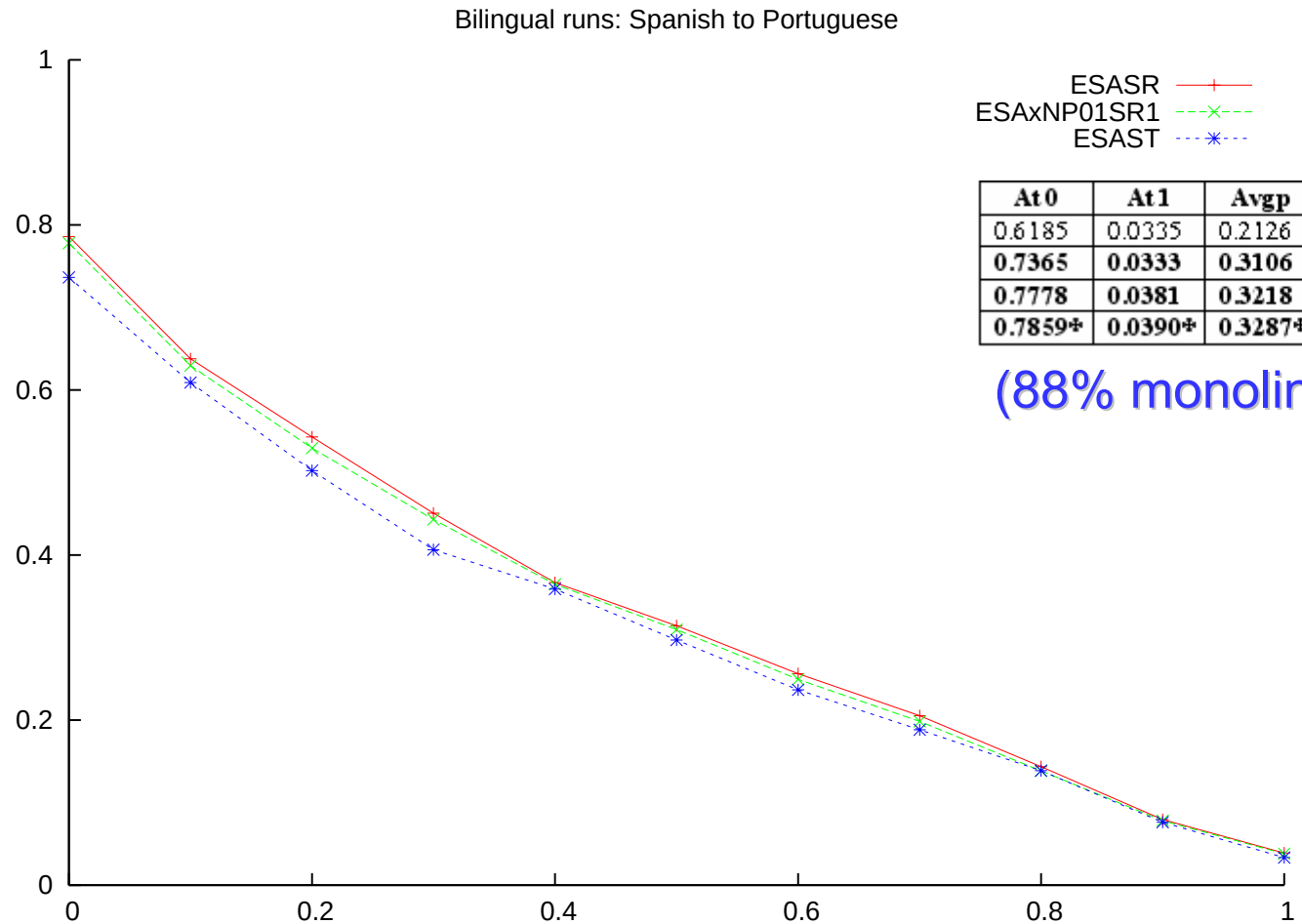
At 0	At 1	Avgp	%	Run
0.3280	0.0353*	0.1159	-43.79 %	ENSNP
0.5521	**0.0329**	**0.2036**	**-1.26 %**	**ENSST**
0.5900*	**0.0224**	**0.2039**	**-1.12 %**	**ENSxNP01SR1**
0.5812	**0.0247**	**0.2062***	**-0.00 %**	**ENSSR**

(55% monolingual)

Rank: 3rd

En→Pt: Systran

Bilingual Spanish→Portuguese



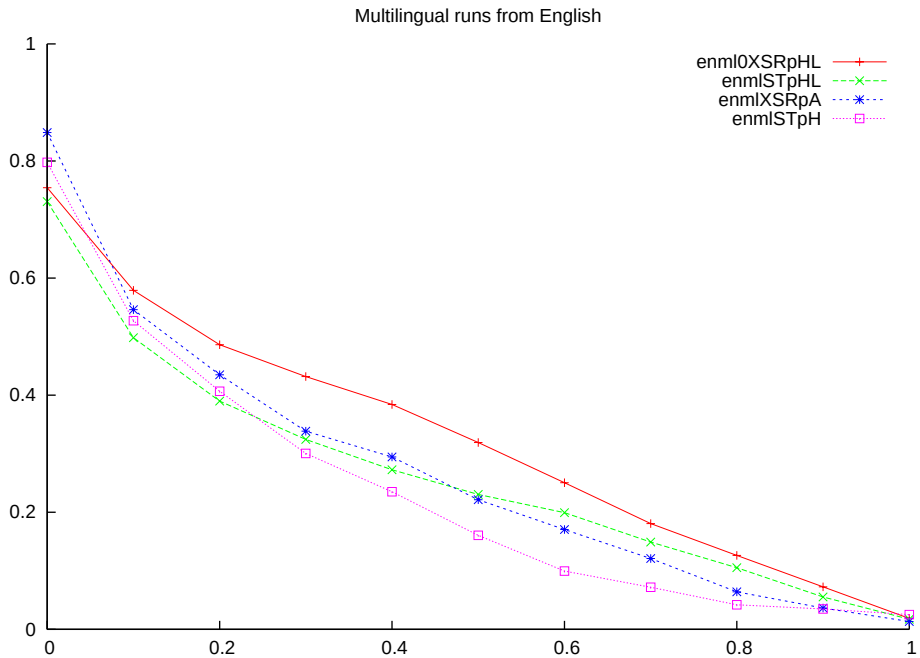
At 0	At 1	Avgp	%	Run
0.6185	0.0335	0.2126	-35.32%	ESANP
0.7365	**0.0333**	**0.3106**	**-5.51 %**	**ESAST**
0.7778	**0.0381**	**0.3218**	**-2.10 %**	**ESAxNP01SR1**
0.7859*	**0.0390***	**0.3287***	**-0.00 %**	**ESASR**

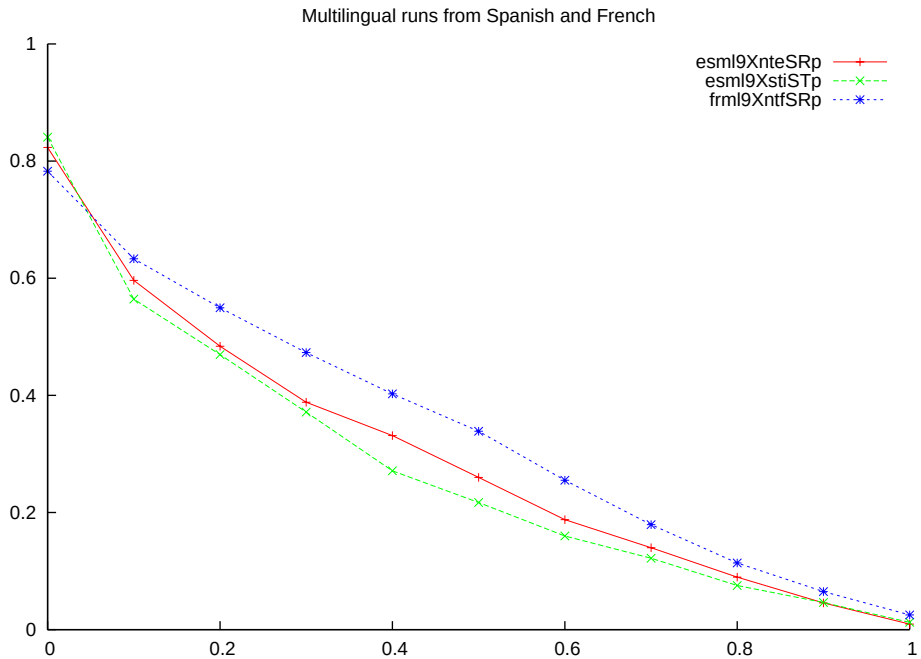
(88% monolingual)

(Rank: 2nd)

Es→Pt: ATrans

Multilingual-8 (En, Es, Fr)





At 0	At 1	Avgp	%	Run
0.8232	0.0094	0.2828*	-15.12%	Best Spanish
0.7543	0.0186	0.3078*	-7.6%	Best English
0.7828*	0.0250*	0.3332*	-0.00%	Best French

Rank: 2nd [Fr, En]
3rd [Es]

Conclusions and homework

- Toolbox = “imagination is the limit”
- Focus on interesting linguistic things instead of boring text manipulation
- Reusability (half of the work is done for next year!)

- Keys for good results:
  - Fast IR engine is essential
  - Native character encoding support
  - Topic narrative
  - Good translation engines make the difference

- Homework:
  - further development on system modules, fine tuning
  - Spanish, French, Portuguese…