UNIVERSITY OF PADUA
Department of Information Engineering

University of Padua at CLEF 2003: Probabilistic Models for Automatic Stemmer Generation

G. M. Di Nunzio, N. Ferro, M. Melucci and N. Orio
{giorgio.dinunzio, nicola.ferro, massimo.melucci, nicola.orio}@unipd.it

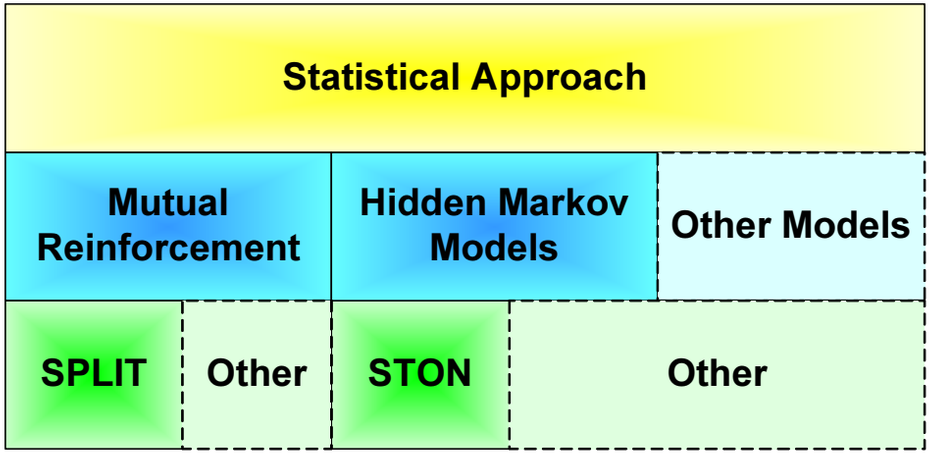
Information Management Systems Research Group

Outline

- Objectives of the Research;
- The SPLIT Stemming Algorithm;
- The STON Stemming Algorithm;
- The IRON Prototype Information Retrieval System;
- Experimental Results.

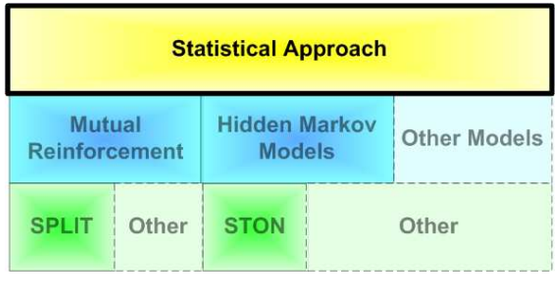
Objectives and Main Steps

The objective of the research is to develop algorithms able to automatically generate stemmers for Western languages.



- we propose a *statistical approach*;
- we employ two different *models*: one based on the notion of Mutual Reinforcement, the other based on Hidden Markov Models (HMM);
- we design two *algorithms*: **SPLIT** (Stemming Program for Language Independent Task) and **STON** (STemming ON).

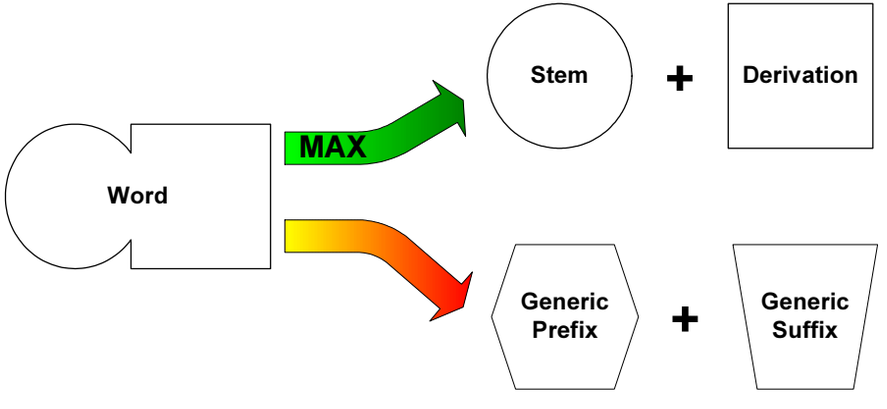
Statistical Approach



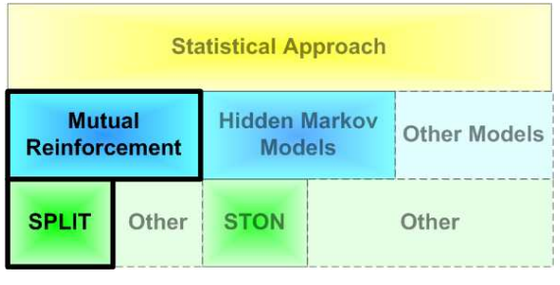
Statistical Approach

A word can be considered as the concatenation of a pair of sub–strings, named, respectively, prefix and suffix.

- given the morphology of the language, the pairs are not equiprobable;
  - the concatenation of a stem with a derivation is assumed to be an event more probable than the concatenation of two generic prefixes and suffixes;
  - a maximization criterion can be employed for identifying the most probable pair of sub–strings, i.e. stem and derivation;
- the probability distribution of the pairs can be estimated from a sample of words of a given language.

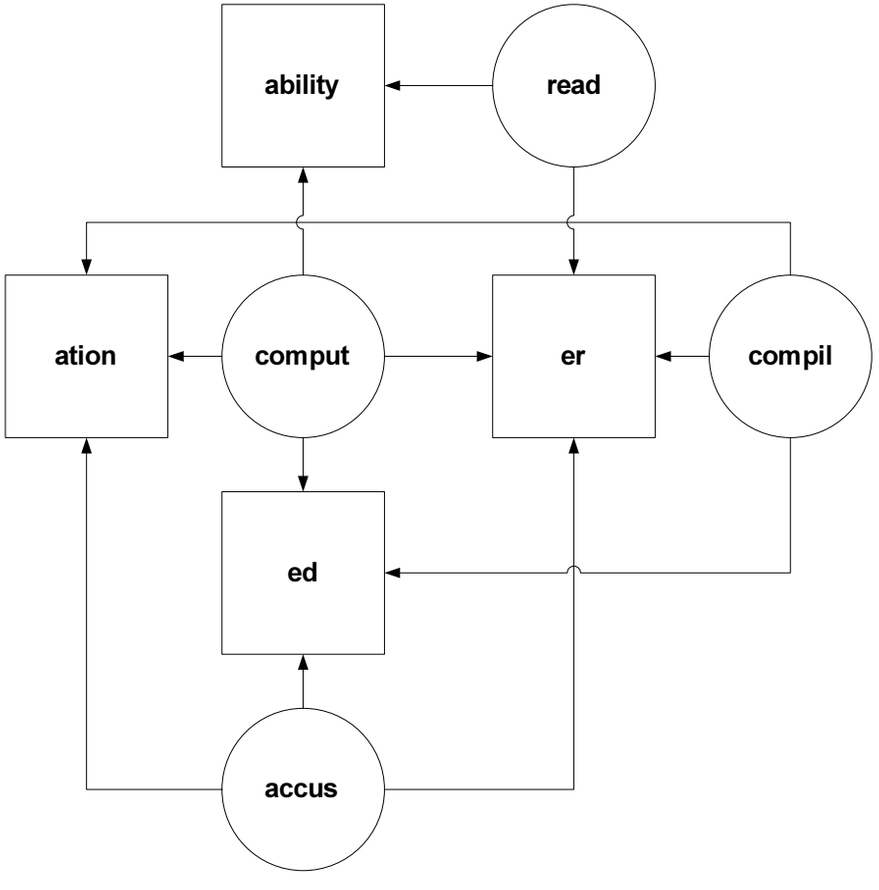


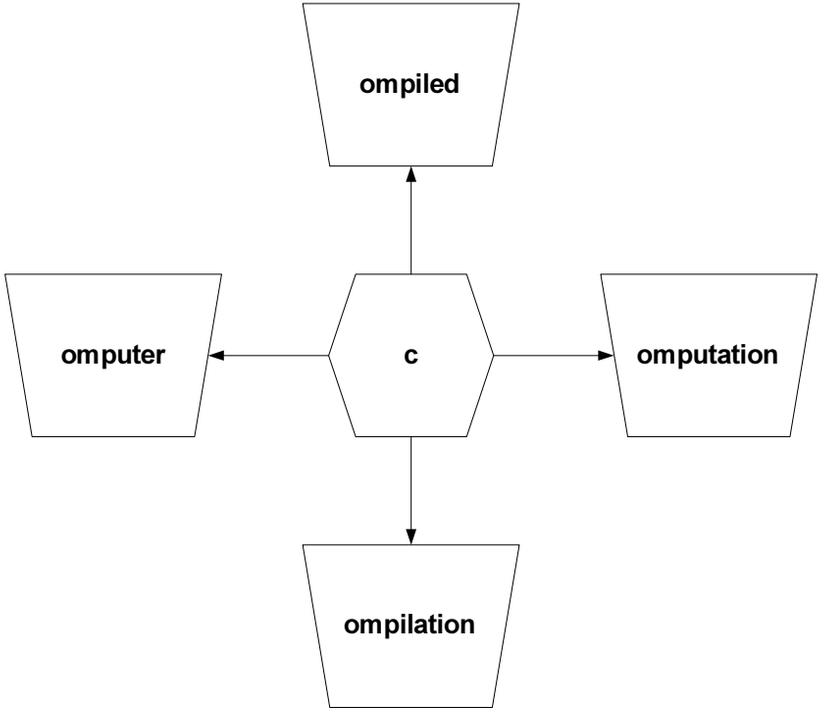
The SPLIT Algorithm



Mutual Reinforcement Model

Stems are prefixes that are more likely to be completed by derivations, and derivations are suffixes that are more likely to complete stems.





The SPLIT Algorithm

- *prefix/suffix estimation*: is a global step regarding the whole word collection, in which the algorithm iteratively computes for each prefix and suffix:

$$\Pr^{(t)}(x) = \sum_{y \in Y} \Pr(x \mid y) \Pr^{(t-1)}(y)$$

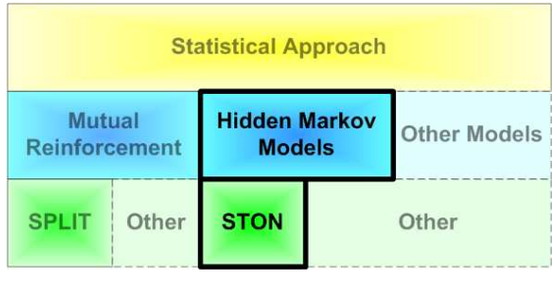
$$\Pr^{(t)}(y) = \sum_{x \in X} \Pr(y \mid x) \Pr^{(t)}(x)$$

- *stem/derivation estimation*: is a local step regarding a given word, in which the algorithm computes:

$$(x^*, y^*) = \arg\max_{(x,y)} Pr(x, y \mid w) = \arg\max_{(x,y)} \frac{Pr(w \mid x, y) Pr(x, y)}{Pr(w)} =$$

$$= \arg \max_{(x,y)|xy=w} Pr(x, y)$$

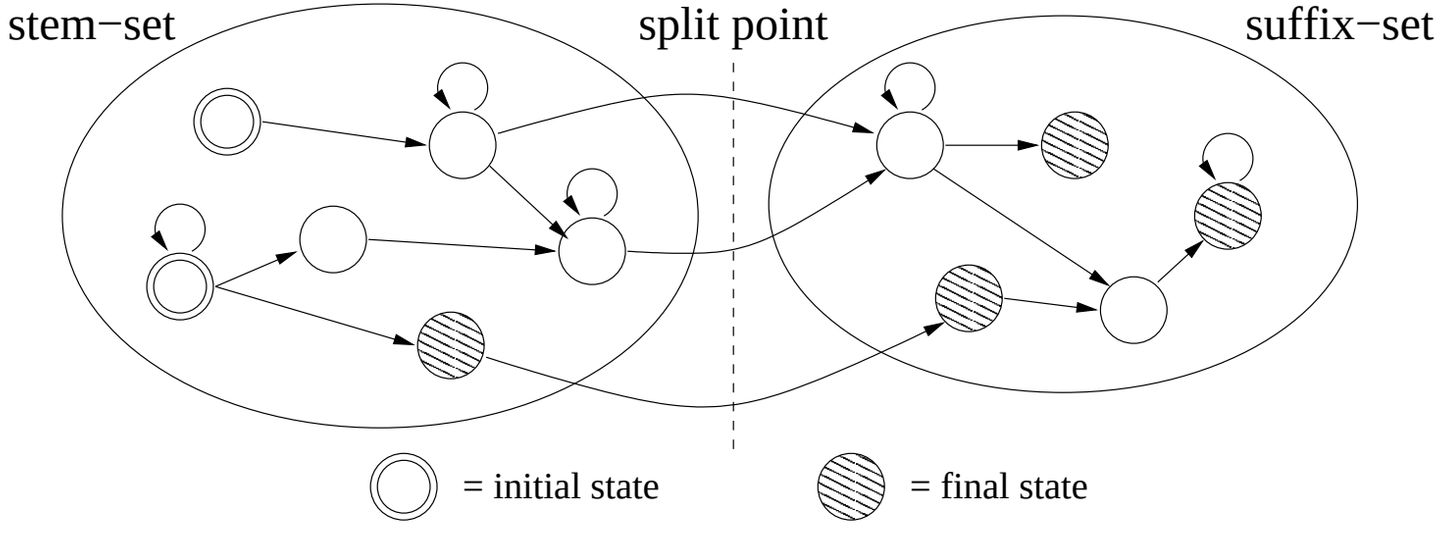
The STON Algorithm



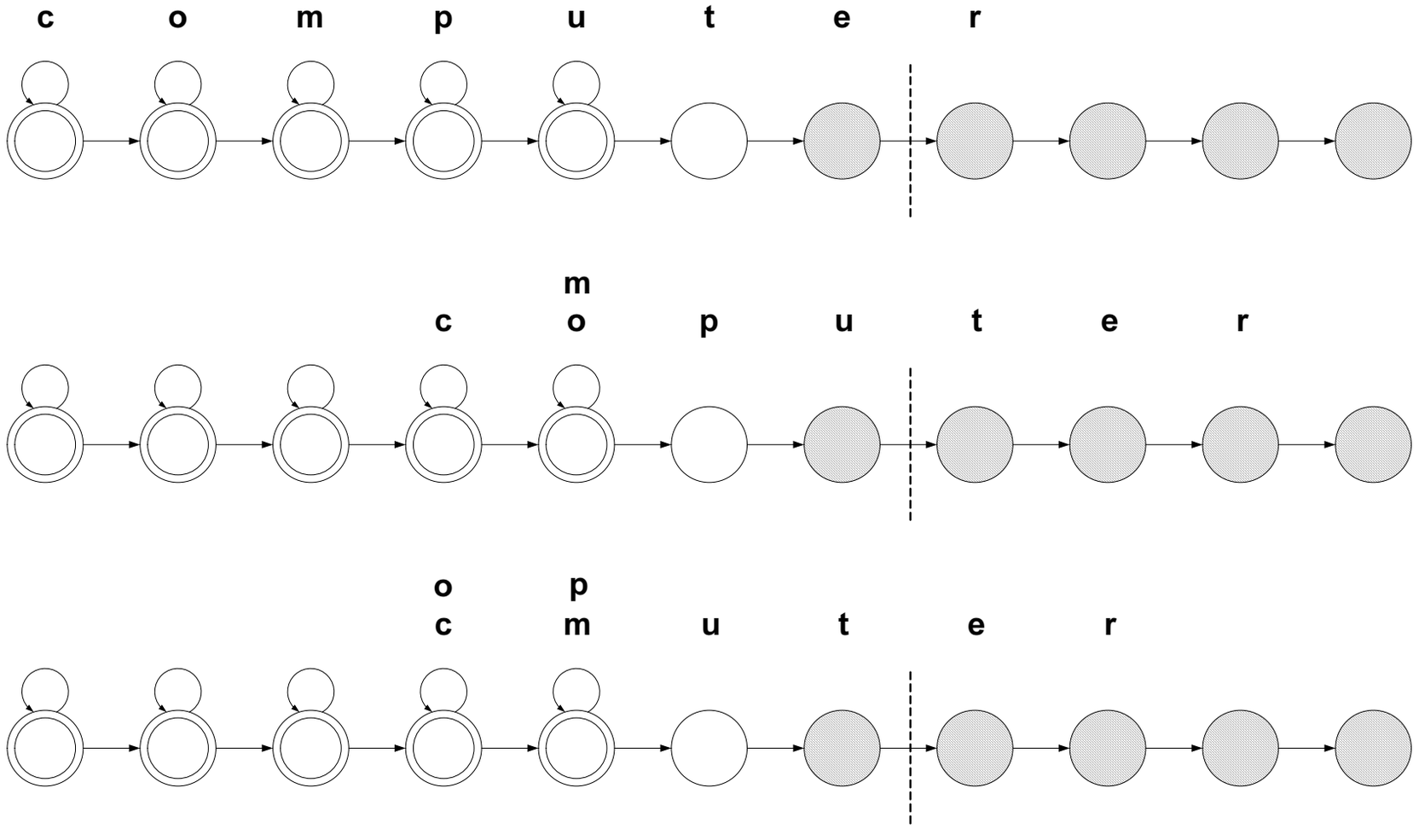
Stemming through Decoding

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} Pr(\mathbf{q} \mid \mathbf{O}, \lambda) = \arg\max_{\mathbf{q}} \frac{Pr(\mathbf{q}, \mathbf{O} \mid \lambda)}{Pr(\mathbf{O} \mid \lambda)} = \arg\max_{\mathbf{q}} Pr(\mathbf{q}, \mathbf{O} \mid \lambda)$$

- The process that generates a word can be divided in two sub-processes;
- The most probable path $\mathbf{q}$ among the states that corresponds to the observed sequence $\mathbf{O}$ is computed using the model λ;
- This path identifies the split point between the two processes, which is used to stem the word;

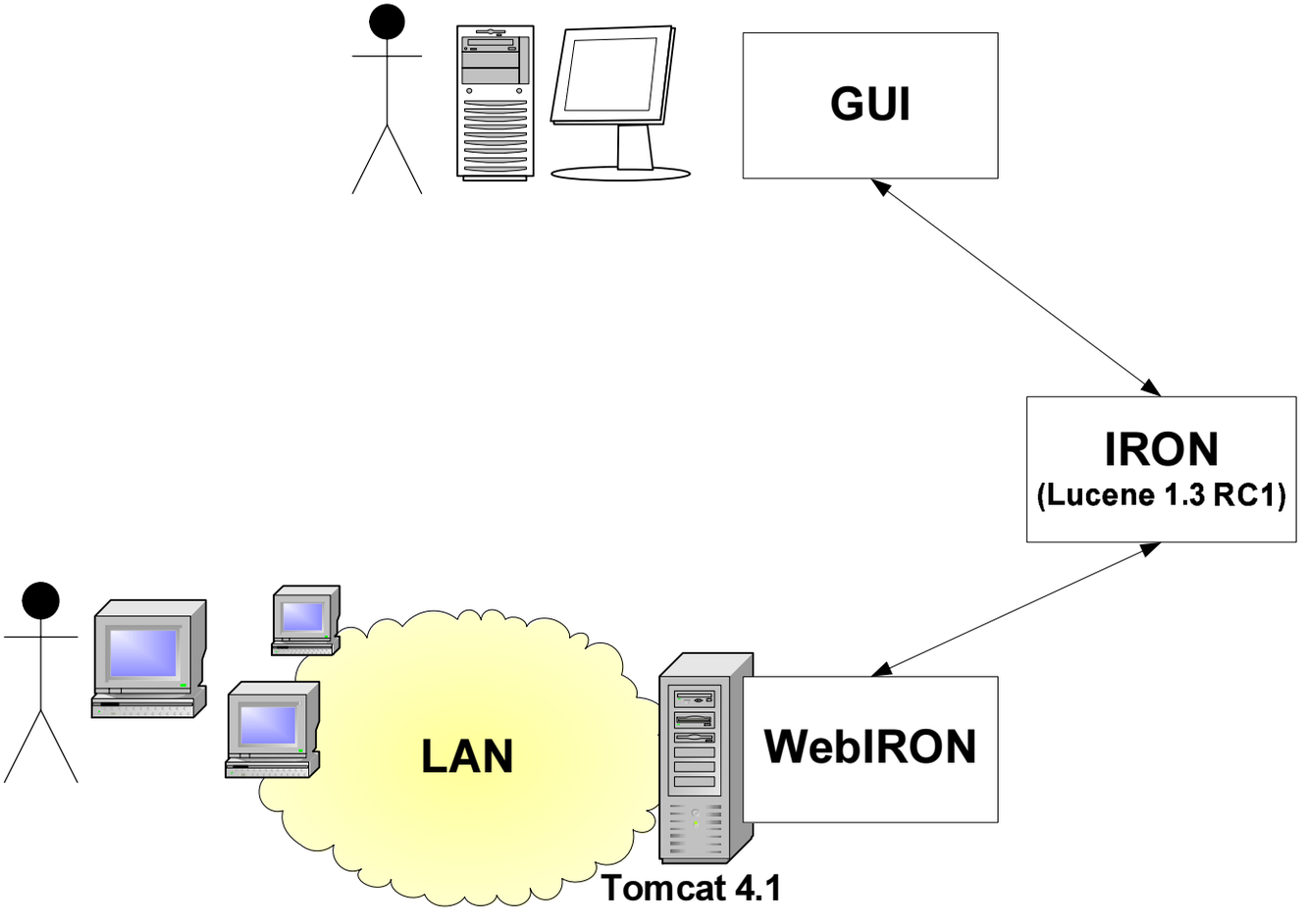


Example of HMM



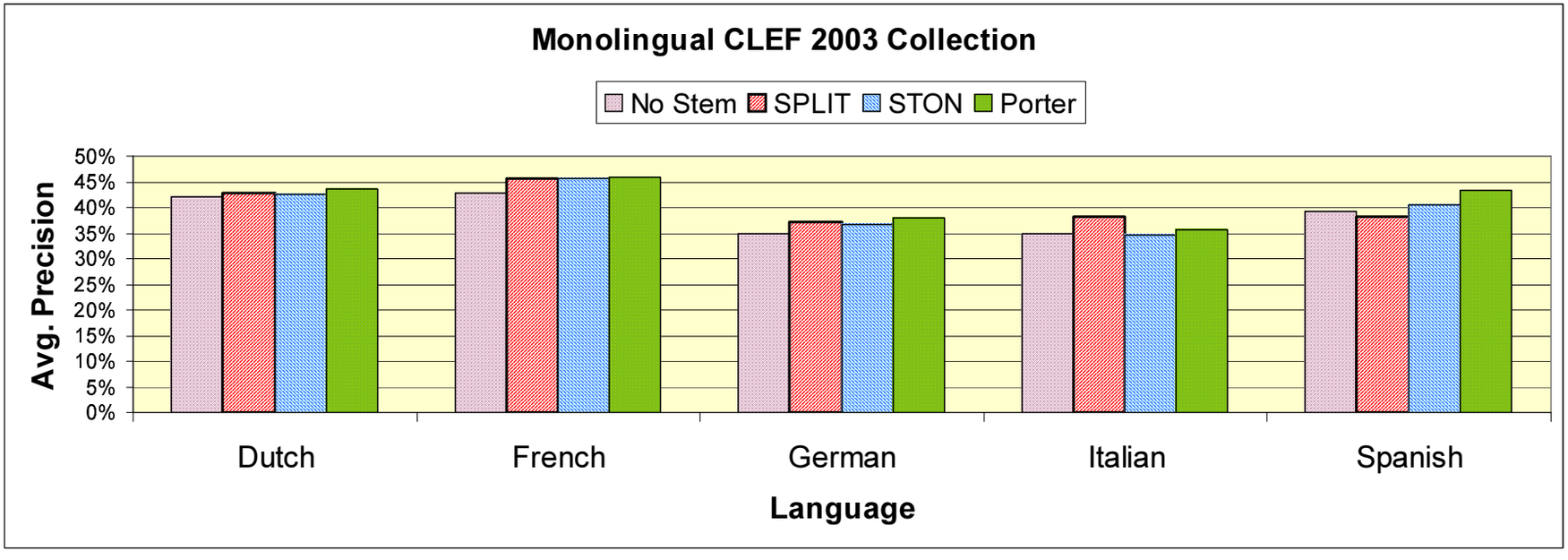
The IRON Prototype IRS

IRON or WebIRON?



Experimental Results

Average Precision



Algorithm	Average Precision (%)				
	Dutch	French	German	Italian	Spanish
No Stem	42.11	42.86	34.92	34.76	39.27
SPLIT	42.84	45.60	37.11	38.17	38.25
STON	42.57	45.67	36.68	34.66	40.56
Porter	43.49	45.87	37.88	35.53	43.42

Conclusions and Future Work

- The proposed stemming algorithms turned out to be as effective as the Porter's stemmer, for different European languages;
- The assumptions made in the statistical approach turned out to be reasonable;
- The statistical approach and the proposed algorithms will be studied in the context of other retrieval tasks or languages.