Natural Language Processing of Recipes

Supervisor: Claire Gardent

Ripan Hermawan - Le Dieu Thu - Jenia Berzak

Menu

- STARTER
  - – Introduction
  - – Resources & Data
- MAIN COURSE
  - – Linguistic Processing
  - – Information Extraction
- DESSERT
  - – Vectors & Clustering
  - – Conclusions

Menu

- **STARTER**
  - **Introduction**
  - **Resources & Data**
- MAIN COURSE
  - Linguistic Processing
  - Information Extraction
- DESSERT
  - Vectors & Clustering
  - Conclusions

From Computer Cooking Contest
CCC 2009 @ ICCBR 2009

Who says that only human beings are able to cook delicious meals?

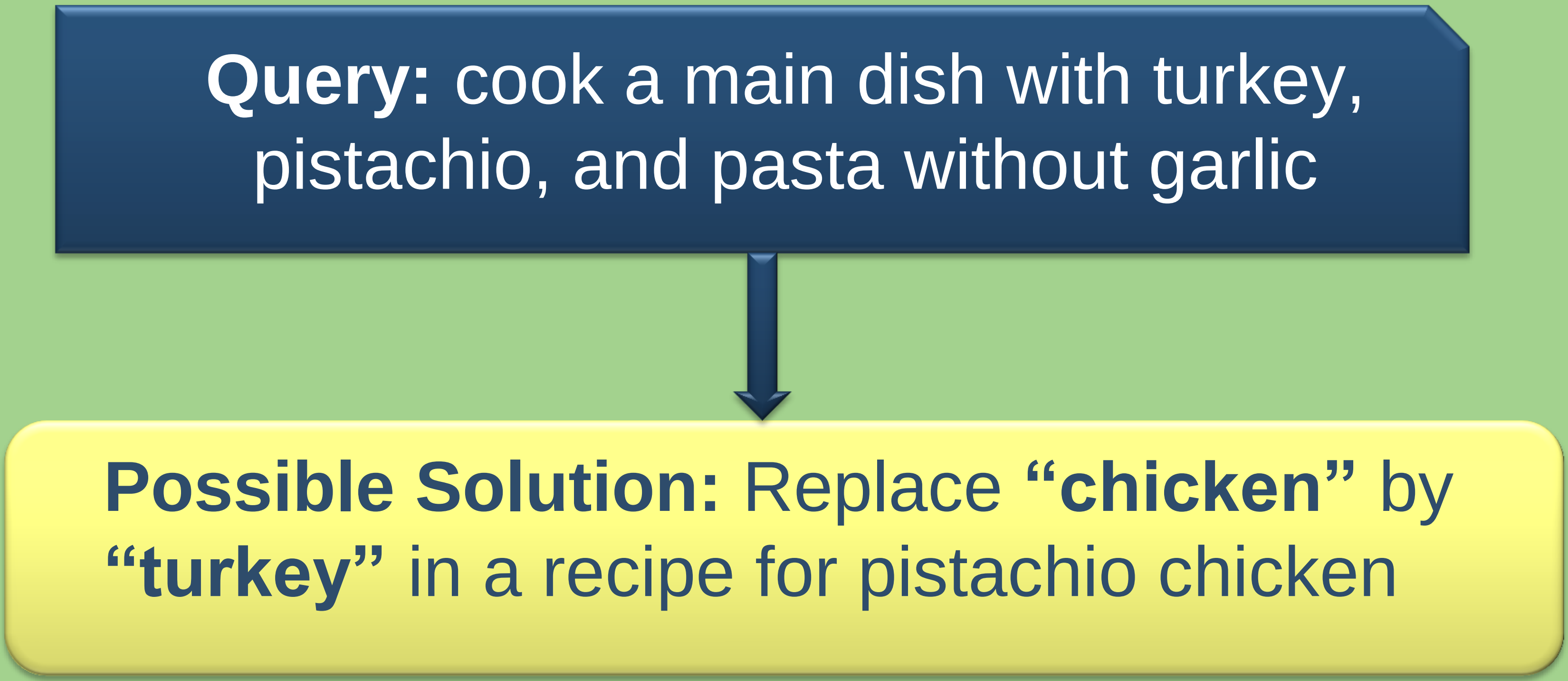
We aim to teach our computers the cuisine…

Problem definition

How to match the content of the fridge to a dish?

- Computer Cooking Contest @ICCBR 2009: the International Conference on Case-Based Reasoning
- Using a computer for the design of the menu
- When appropriate, **modify the recipe** to match the ingredients we have

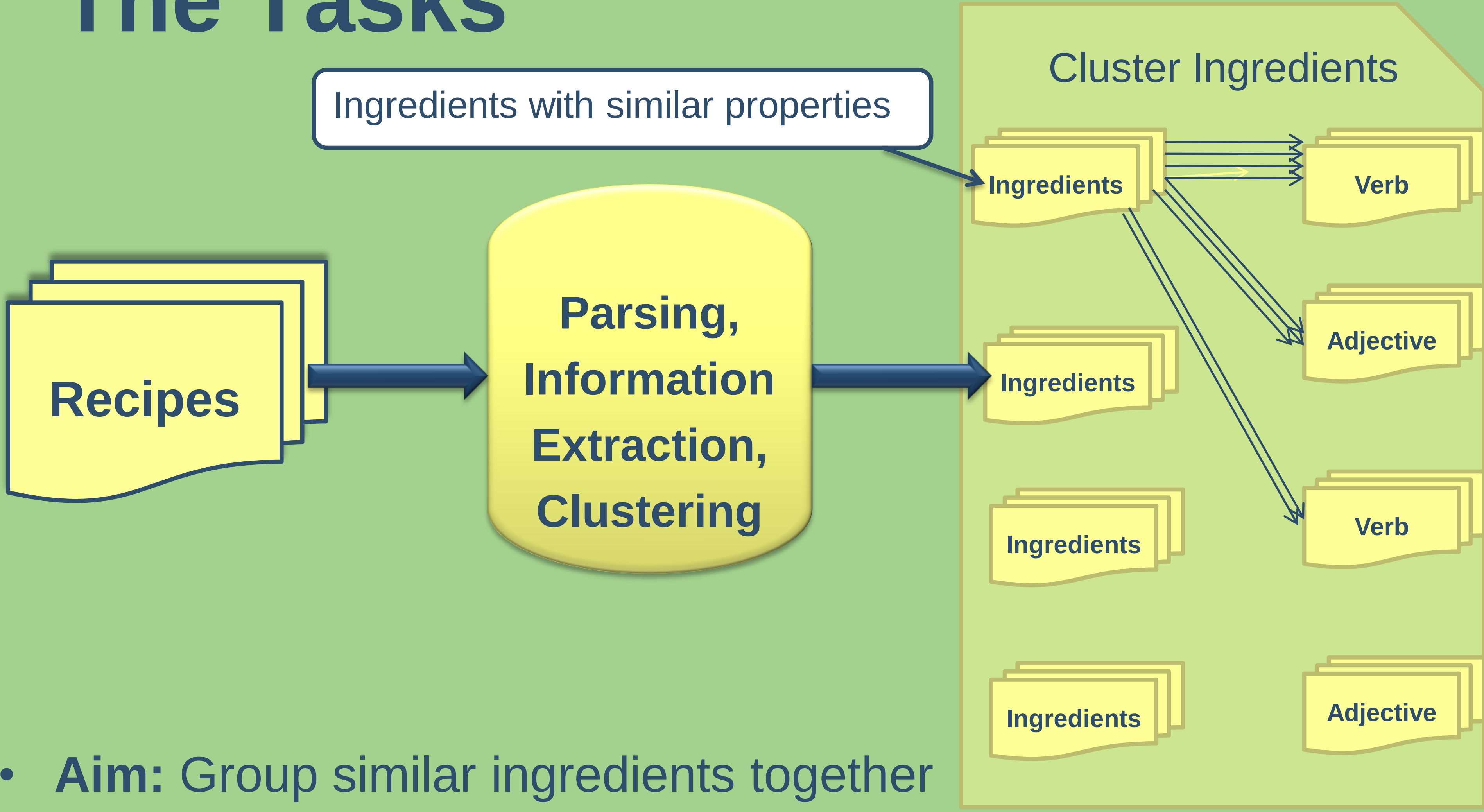
Recipe Adaptation



Query: cook a main dish with turkey, pistachio, and pasta without garlic

Possible Solution: Replace **“chicken”** by **“turkey”** in a recipe for pistachio chicken

The Tasks



- **Aim:** Group similar ingredients together
- **Strategy:** Ingredients are similar if they participate in the similar actions, have similar properties

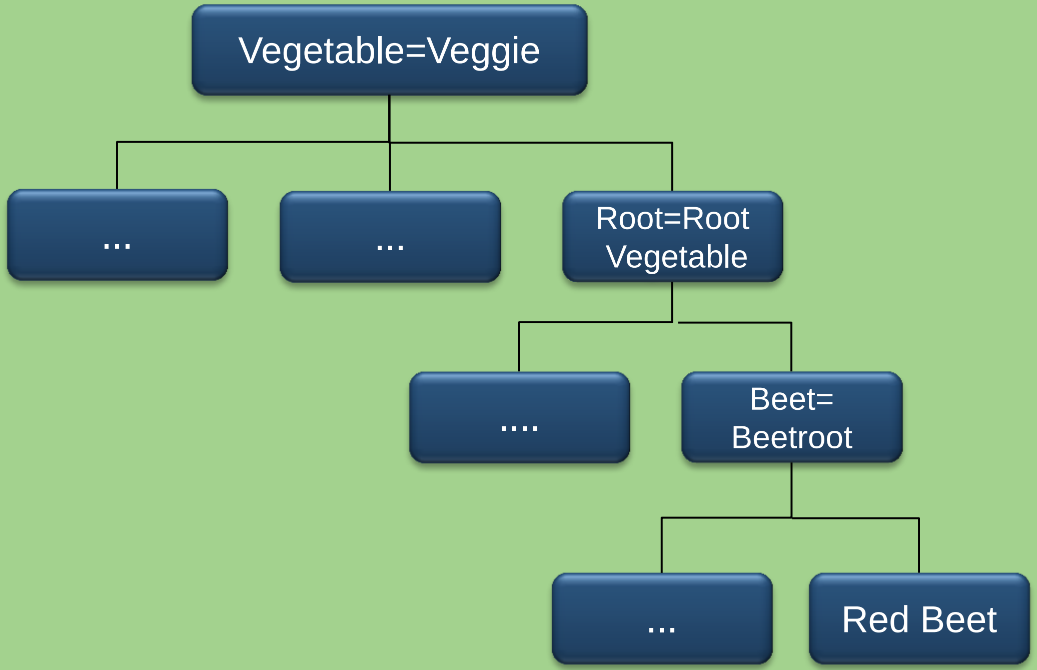
Resources & Data

- **A database of 1,459 recipes**:
  - Title
  - Ingredients description
  - Preparation instructions
- **Challenges for Linguistic Processing**
  - Imperative sentences
  - Unusual use of words
  - Missing verb complement
  - Referring expression and anaphora
  - Domain specific vocabulary

```
<RECIPE>
<TI>Almond Roca Cookies</TI>
<IN>1 c Butter</IN>
<IN>1/2 c Brown sugar</IN>
<IN>1/2 c Sugar</IN>
<IN>1  Egg yolk</IN>
<IN>1 ts Vanilla</IN>
<IN>2 c Flour</IN>
<IN>10 oz Chocolate chips</IN>
<IN>1 c Finely chopped nuts</IN>
<PR> Cream butter and sugar, add egg yolk and
     vanilla. Stir in flour. Spread mixture thinly on
     greased cookie sheet. Bake at 350 degrees for
     15 to 20 minutes. Melt chocolate over hot water
     and spread over warm baked cookies. Sprinkle
     on nuts and press them firmly. Cut in bars while
     still warm. Let stand until chocolate is dry.
</PR>
</RECIPE>
```

Resources & Data

- **Hierarchy of Ingredients**
  - Extracted from external database by **Orpailleur** group
  - Each node may contain the synonyms of the same ingredients
- **Recipes database with annotated ingredients**

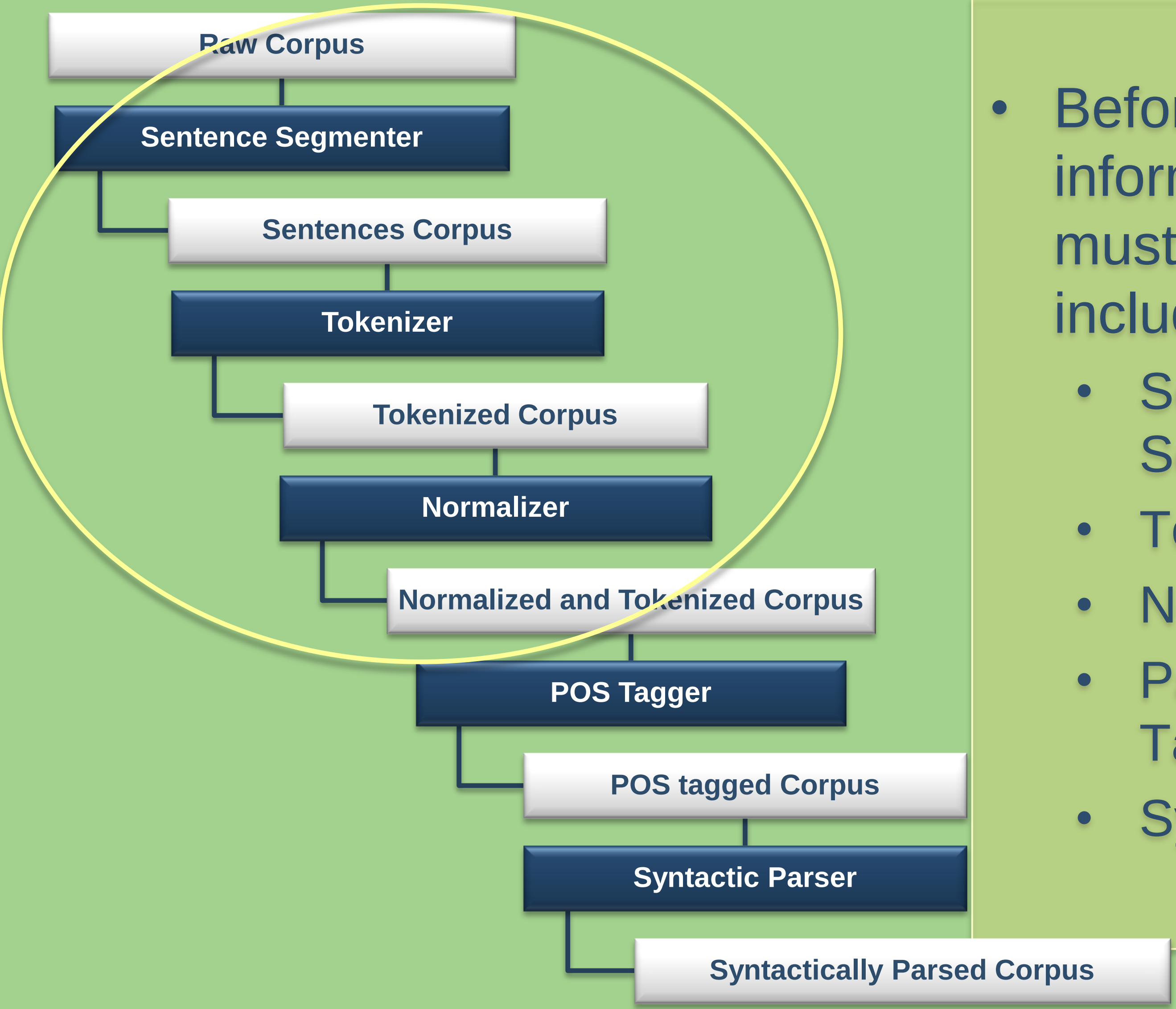


<TI>Almond Roca Cookies</TI>
<IN>
<IN_I>1 ts Vanilla</IN_I>
<ING>vanilla</ING>
<QT>1</QT>
<U>tsp</U>…
</IN>

Menu

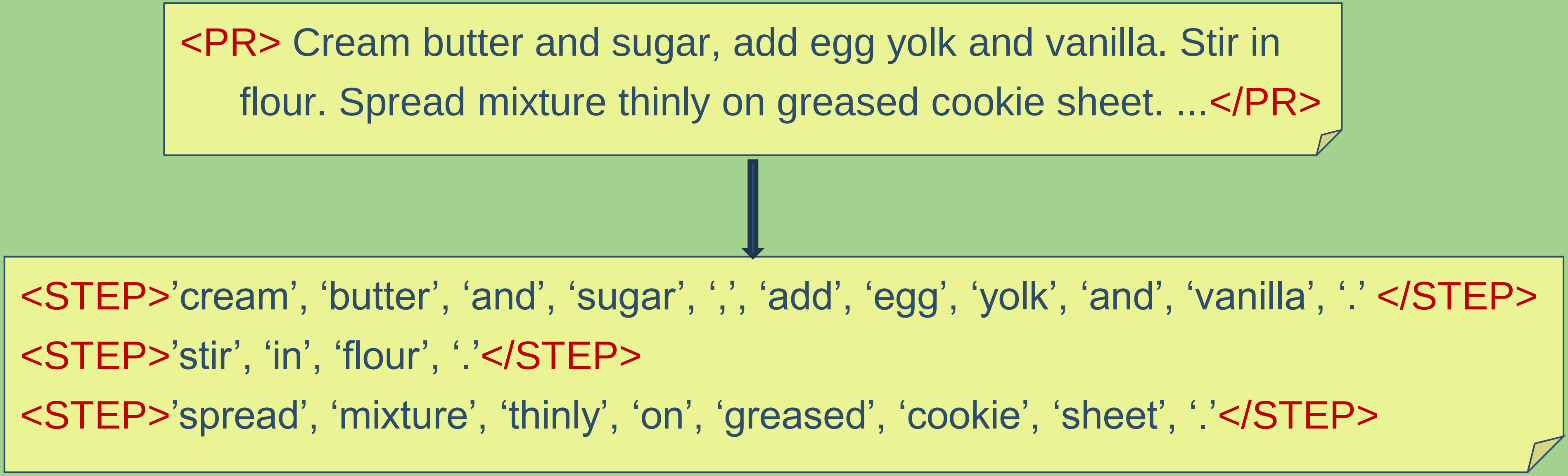
- STARTER
  - Introduction
  - Resources & Data
- **MAIN COURSE**
  - **Linguistic Processing**
  - **Information Extraction**
- DESSERT
  - Vectors & Clustering
  - Conclusions

Linguistic Processing



- Before extracting information, the corpus must be pre-processed including:
  - Sentence Segmentation
  - Tokenization
  - Normalization
  - Part of Speech Tagging
  - Syntactic Parsing

Segmentation, Tokenization & Normalization



<PR> Cream butter and sugar, add egg yolk and vanilla. Stir in flour. Spread mixture thinly on greased cookie sheet. ...</PR>

<STEP>’cream’, ‘butter’, ‘and’, ‘sugar’, ‘,’, ‘add’, ‘egg’, ‘yolk’, ‘and’, ‘vanilla’, ‘.’ </STEP>

<STEP>’stir’, ‘in’, ‘flour’, ‘.’</STEP>

<STEP>’spread’, ‘mixture’, ‘thinly’, ‘on’, ‘greased’, ‘cookie’, ‘sheet’, ‘.’</STEP>

- Segment the string of characters into **sentences**
- Sentences are divided into **tokens**, basic unit for linguistic processing
- **Normalization:** convert to lowercase, filter invalid tokens
- **Difficulties:** sentence (word) boundaries, excluding punctuations (stop word, ice-cream), etc.

Part of Speech (POS) Tagging

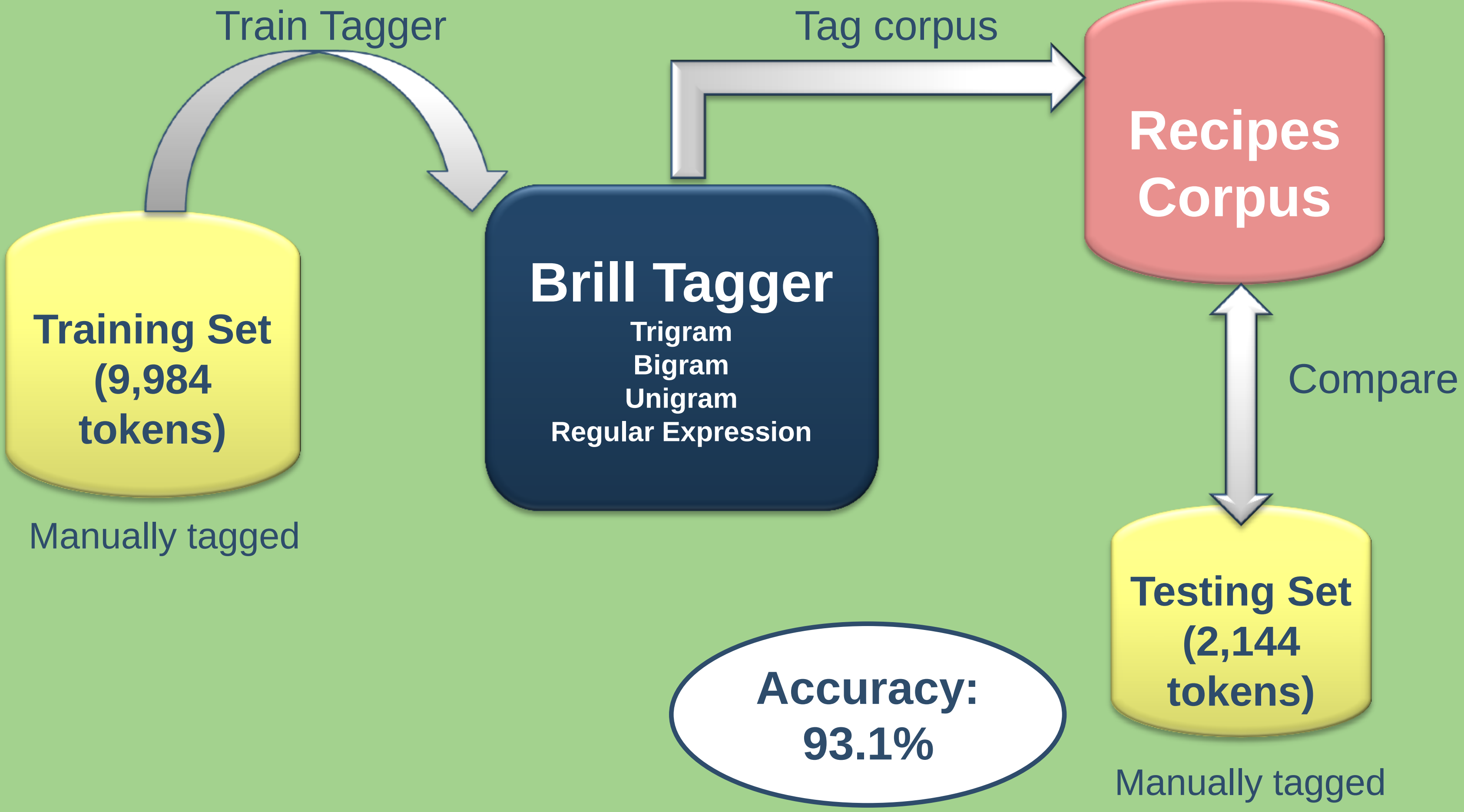
Syntactic Categorization of Words

melt/VB the/AT butter/NN
and/CC cook/VB the/AT
leeks/NNS and/CC garlic/NN
for/IN 3/CD minutes/NNS
over/IN medium/JJ heat/NN ./.

Tag	Meaning
NN	noun, singular, common
NNS	noun, plural, common
VB	verb, base: uninflected present, imperative or infinitive
CC	conjunction, coordinating
CD	numeral, cardinal
JJ	adjective

- Assign to each word a syntactic category using Brown tag set

Part of Speech Tagging (cont)



Train Tagger

Tag corpus

Training Set (9,984 tokens)

Manually tagged

Brill Tagger

Trigram
Bigram
Unigram
Regular Expression

Recipes Corpus

Compare

Testing Set (2,144 tokens)

Manually tagged

Accuracy: 93.1%

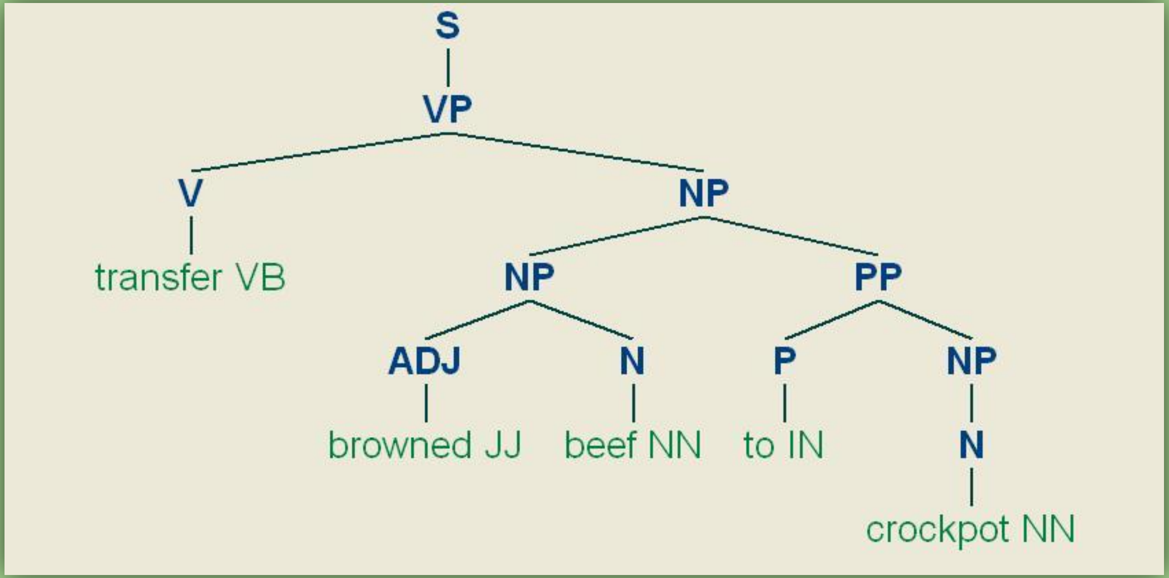
Parsing
Syntactic Analysis

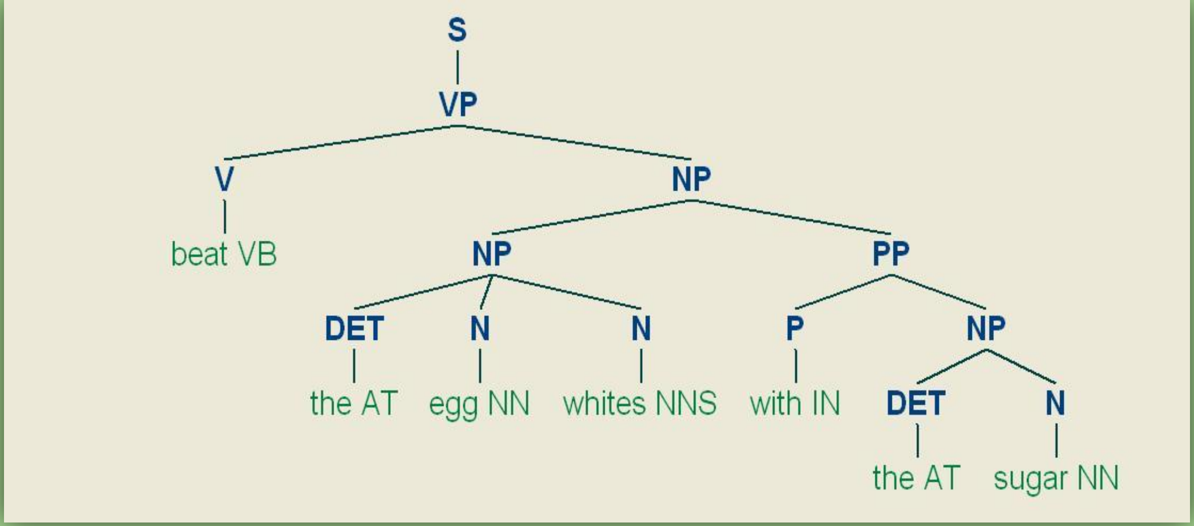
Parsing Grammar

```
DET: {<DT|AT|ABN|ABL|ABX|AP|AP$|PP$|WDT|CD>}    # Determiner
ADV: {<QL|QLP|RB|RB$|RBR|RBT>}                  # Adverb
ADJ: {<AP|JJ|JJ+JJ|JJR|JJS|JJT >}               # Adjective
N: {<NN|NNS|NPS|NPS$|PPS|PPSS>}                 # Noun
COR: {<CC>}                                     # Coordinating Conjunction
NP: {(<DET>*<ADV>*<ADJ>*<N>+)}                  # Noun Phrase
NP: {<NP>(<COR><NP>)+}                          # Noun Phrase
NP: {<NP><PP>}                                  # Noun Phrase
P: {<IN>}                                       # Preposition
PP: {<P><NP>}                                   #Prepositional Phrase
V: {((<BE.*>*<V.*>+)|(<H.*>*<V.*>+)|(<M.*>*<V.*>+)|(<DO.*><V.*>))<RP>*}    # Verb
VP :{(<ADV>?<V><ADV>?(<COR>?<ADV>?<V><ADV>?)*)<PP>*<ADV>?<NP><PP>*} # Verb Phrase
S: {<NP>*<VP>}                                  # Sentence
```

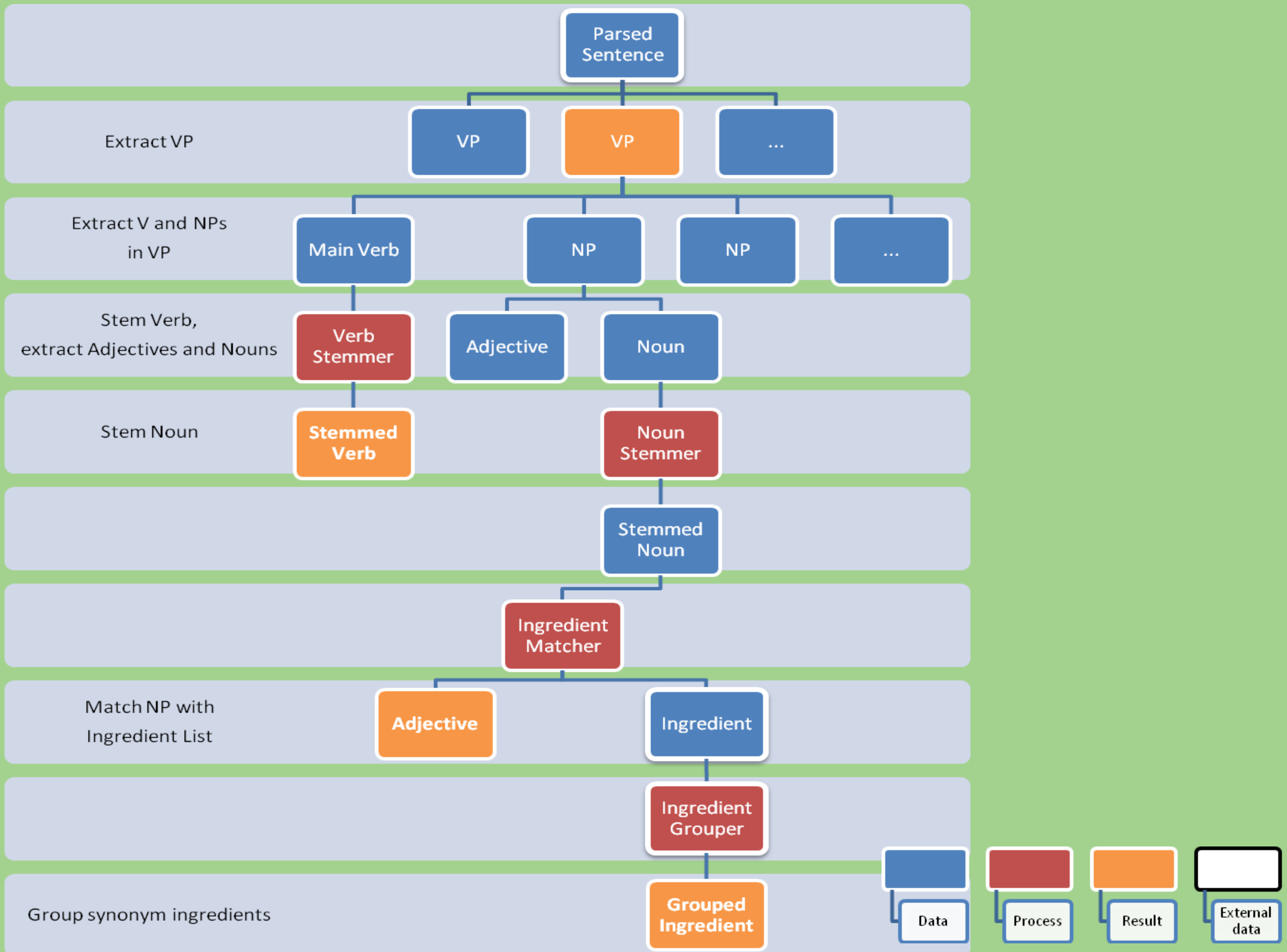
- Develop a parser for the corpus based on POS tags
- Define the grammar tailored for the corpus

Parsing: Results Examples





Information Extraction from Parsed Corpus



Parsed Sentence

Extract VP — VP, VP, ...

Extract V and NPs in VP — Main Verb, NP, NP, ...

Stem Verb, extract Adjectives and Nouns — Verb Stemmer, Adjective, Noun

Stem Noun — Stemmed Verb, Noun Stemmer

Stemmed Noun

Ingredient Matcher

Match NP with Ingredient List — Adjective, Ingredient

Ingredient Grouper

Group synonym ingredients — Grouped Ingredient

🟦 Data

🟥 Process

🟧 Result

⬜ External data

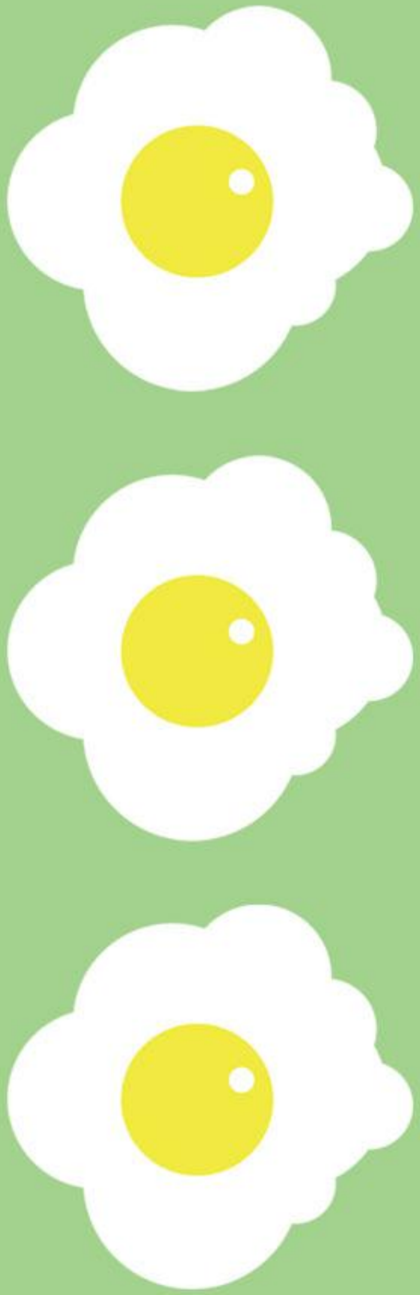
Morphological Normalizing, Ingredient Matching, Grouping

- **Stemming:** reduce a word-form to a base-form/root
  (carrot**s** -> carrot, mix**ing** -> mix)
- **Ingredient Matching:** match Noun Phrase with Ingredient
- **Ingredient Grouping:** group same ingredients that have different names
  (milk powder, dry milk)

<VP>
<ORG>mix/VB **margarine/NN** and/CC **milk/NN powder/NN</ORG>**
<V>mix</V>
<NP>
    <ING> **margarine oleo, margarine** </ING>
</NP>
<NP>
    <ING>**milk powder, dry milk**</ING>
</NP>
</VP>

Menu

- STARTER
  - Introduction
  - Resources & Data
- MAIN COURSE
  - Linguistic Processing
  - Information Extraction
- **DESSERT**
  - **Vectors & Clustering**
  - **Conclusions**



Syntactic Vectors

Goals: 1. Associate each ingredient with all its actions and properties.

2. Group ingredients according to this information

Extracted Information:

- Co-occurrences of verbs and ingredients:
  if both appear in the same verb phrase
- Co-occurrences of adjectives and ingredients:
  if both appear in the same immediate noun phrase

- Syntactic vectors summarize the co-occurrences information extracted from the parser output.

Syntactic Vectors

- A vector is simply a frequency distribution of verbs and adjectives co-occurrences for a given ingredient.
- Each **ingredient** is represented as a **vector of features**.
- Each **feature** is either a **verb or an adjective**.
- Each feature **value** represents the **number of co-occurrences** of the ingredient with the feature in the recipes corpus.

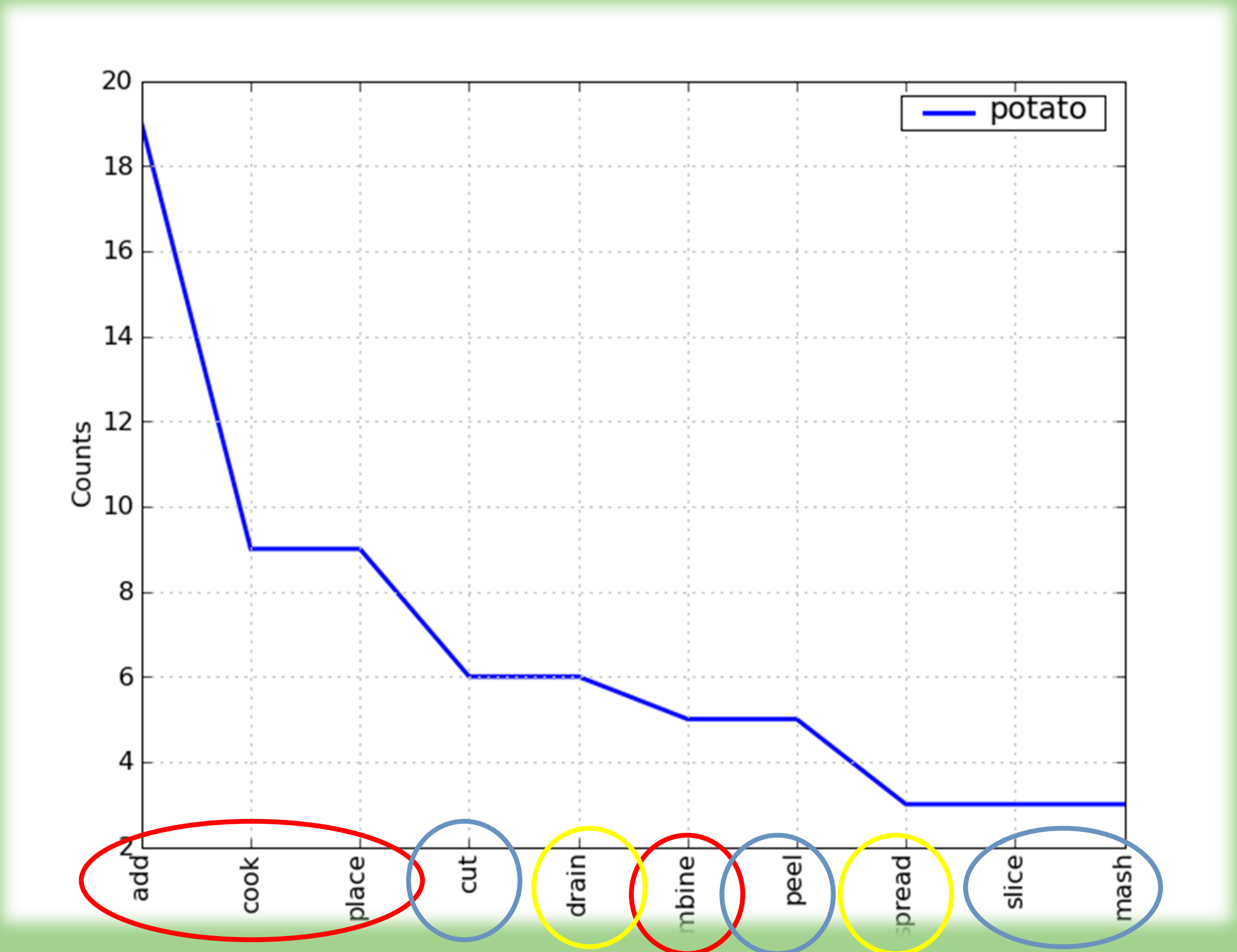
Syntactic Vectors

- 537 vectors (ingredients)
- 514 features (dimensions):
  - 300 verb features
  - 214 adjective features
- Examples of vectors

Ingredient	Verbs				Adjectives		
	add	beat	brown	…	hard	chopped	…
Onion	78	0	15	…	0	8	…
Egg, whole egg	58	98	0	…	2	0	…

Vectors

- Frequent verb features of “potato”



Clustering

- Clustering ingredients using K-Means clustering algorithm.
- Experimenting with different configurations:
  - Different numbers of clusters
  - Clustering verbs using ingredients as features

Clustering Results

- Most clusters are hard to interpret.
- “good” clusters:
  - garlic, onion
  - fettucine/fettuccine, green bean, macaroni, ravioli
- “bad” clusters:
  - coconut, duck, popcorn
  - milk, salt, tomato, vanilla

Future Improvements

- Bigger dataset
- Normalization of the vectors
- Adding features (other syntactic relations, semantics)
- Using other clustering methods (hierarchical clustering, concept analysis with a lattice)

Future Improvements

- **Improving features:**
  - Unifying verbs and adjectives that convey the same meaning (e.g. boil and boiled)
  - Separating different grammatical functions / thematic roles (e.g. pour the sauce over the potatoes)
  - Filtering/weighting frequent verbs that take almost any ingredient
  - Filtering/weighting very low frequency features

Summary

- Grouping ingredients using a distributional analysis of syntactically parsed text for the task of ingredient substitution for recipe adaptation.
- Current results:
  - good linguistic processing and information extraction
  - clusters are difficult to interpret
- Experiments with some of the suggested changes yield promising results.

Bon Appétit!