Using Comparable Collections of Historical Texts for Building a Diachronic Dictionary for Spelling Normalization

Marilisa Amoia and José Manuel Martínez

Institute for Applied Linguistics, Translation and Interpreting
Saarland University
Support research in the Humanities through NLP.

Advantages of digital historical resources:

- faster and wider data retrieval,
- provide new insights and/or a more consistent and reliable account of findings,
- further data annotation and inter-liking with related resources in a centralized way.
State of Art

- In the last years: flourishing of large digitization programs in most European countries
- historical corpora,
- metadata annotation: year, author, geographic location, language, etc.
To automatize the process of annotation (e.g. pos tagging, syntactic parsing, semantic parsing) is a problematic issue:

- the noise introduced by deviant linguistic data,
- spelling/orthography variation,
- lack of sentence boundaries, etc.
- historical false friends, e.g. \textit{statt} (instead of) vs. \textit{stadt} (city), \textit{Bett} (bed) vs. \textit{bete} (pray)
Our approach

Comparable corpora have proven very useful in MT if parallel corpora are not available. We exploit ideas and techniques from MT for automatic extraction of diachronic dictionaries and spelling normalizing.

- we build a diachronic comparable corpus of German cooking recipes,
- we apply clustering techniques for finding word variants.
The Saarbrücken Cookbook Corpus is a historical comparable corpus made of recipe repertoires published in German language during the Early Modern Age.

- SaCoCo is one of the first attempts to build a comparable historical corpus of German.

- The corpus is made up of two collections:
  - *Historical subcorpus*, a historical comparable dataset aligned at recipe level providing multiple versions of the same dish across the time span of the core corpus;
  - *Contemporary subcorpus*, a contemporary comparable dataset providing contemporary German versions for each recipe.
SaCoCo Historical Corpus

- recipe collection spans two hundred years: 1569-1729
- recipe books by 14 different German authors
- total of 430 recipes (about 45,000 tokens)
- 107 average length of recipe (in tokens)
Digitization Strategy

- Manual transcription was part of a PhD Thesis (Andrea Wurm) → diplomatic transcription
- Some standardization: punctuation, hyphenation
- No standardization: spellchecking, word separation
- The corpus is encoded in UTF-8.
SaCoCo Contemporary Corpus

- recipe collection from Internet
- total of 1500 recipes (about 500,000 tokens)
- 325 average length of recipe (in tokens)
SaCoCo Alignment Strategy

- Historical as well as contemporary recipes have been manually annotated with main ingredient and cooking method information
- This information is used to extract comparable recipes, e.i recipes describing the preparation of the same dish

| MainIngredient=Huhn and CookingMethod=Suppe (chicken & soup) |
|-------------------------------------------------------------|
| **Historical:**                                              |
| 1800: Eine ordinaire Hühnersuppe mit Perlgrauken            |
| 1800: Hühnersuppe mit Reis                                  |
| 1800: Hühnersuppe mit Reis auf eine andere Art              |
| 1698: Suppe Sante, von Hünern und Pastinaken garniret       |
| 1698: Suppe von Macronen mit jungen Hünern                  |
| 1686: Hünlein in einer schwartzen Suppen                    |
| **Contemporary:**                                           |
| 2000: Hühnersuppe                                           |
| 2000: Einfache Hühnersuppe                                  |
| 2000: Festliche Hühnersuppe                                 |
| 2000: Hühnerbrühe                                           |
Advantages of Comparable Resources

- (1) no norm is needed, no need of a gold standard, e.g. for languages with very few resources
- (2) apply well know MT techniques to digital humanities
Automatic Annotation

- Normalization
- Lemma, POS-tagging: TreeTagger (Schmidt, 1994), trained on the TBa-D/Z treebank (performance about 97.4%, 78% on unknown words)
Normalization

A two-step framework:

- String similarity: different spellings of the same word
- Distributional semantics: different spellings of the same word and/or semantic similar words e.g. synonyms
Normalization: String Similarity

- Clustering techniques based on string similarity measure: agglomerative hierarchical clustering, Levenshtein edit distance, (65% similarity)
- Historical word form variants:
  vnd_1569, vnnd_1569, vnd_1679, und_1698, und_2000
- Normalized form: is the most modern form among the historical variants → und
Normalization: Distributional Semantics

Distributional semantic techniques based on measure of mutual information (Lin 1998):

- start by generating a list of trigrams from the corpus,
- assign to each pair of tokens in the corpus a value for their mutual information,
- assign to each pair of tokens in the corpus a value for their similarity,
- take the most similar token as the normalized form.
Normalization: Distributional Semantics

Mutual Information: \( I(t_1, t_2) = \log \frac{\|t_1,*,t_2\|\|*,*,*\|}{\|t_1,*,*\|\|*,*,t_2\|} \)

Semantic similarity: \( \text{sim}(w_1, w_2) = \frac{\sum_{T_{w_1} \cap T_{w_2}} I(w_1,*) + I(w_2,*)}{\sum I(w_1,*) + \sum I(w_2,*)} \)
Normalization: Distributional Semantics

- Distributional semantic techniques based on measure of mutual information (Lin 1998):
- Historical word form variants:
  - Zwippeln::Suppengemüse#0.1408360910393916 (e.g. onion:: soup vegetable)
  - Ulmer=Gerstlein::Gerstlein#0.6729067734961974, von#0.035148440000209266 (e.g. barley, a sort of cereal)
  - köcheln::garen#0.14688687822072227, aufkochen#0.051675148156741894 (e.g. fermenting, boil up)
- Normalized form: is the most similar among the historical variants
  → Suppengemüse
### Preliminary Evaluation

- **Subcorpus**: recipes on how to roast a chicken (32 historical and 52 modern recipes)
- **7103 words** (about 8% of whole corpus)

| Strategy                | Lemma (%) | POS (%) |
|-------------------------|-----------|---------|
| no-Normalization:       | 73%       | 80%     |
| string-similarity:      | 81%       | 81.4%   |
| semantic-similarity:    | 82.5%     | 82%     |
Corpus Query

SaCoCo allows queries at different level of annotation:
- lemma, POS:
- historical word forms
- normalized form:
- shallow semantic: main ingredient, cooking method
An Example Annotation

<recipe id="24" ingredient="Huhn" cookingMethod="Pastete">
...
<seg type="newline">
  nimm   VVIMP  nehmen
  dann   ADV    dann
  ein    ART    ein
  hand   NN     Hand
  voll   ADJD   voll
  oder   KON    oder
  zwei   CARD   zwei
  gut    ADJD   gut
  grün   ADJD   grün
  Kraut  NN     Kraut
</seg>
<seg type="newline">
  Binetsch  NN    UNKNOWN
  oder     KON    oder
  Mangold  NN     Mangold
</seg>
...
<seg type="newline">
  vnd     KON    und
  thu     VVFIN  tun
  es      PPER    es
  an      APPR    an
  den     ART     d
  Boden   NN     Boden
</seg>
The diachronic corpus on the web:

http://fedora.clarin-d.uni-saarland.de/sacoco
SaCoCo Web Interface

- CQPweb is a web-based graphical user interface for the CQP query processor (part of CWB - The IMS Open Corpus WorkBench, originally developed at Stuttgart University)
- CQP-web allows easy corpus access
- CQP-web implements some useful corpus query functionalities such as frequency distribution, collocations
Corpus Querying
The Imperative Form

Menu

Corpus queries
- Standard query
- Restricted query
- Word lookup
- Frequency lists
- Keywords

User controls
- User settings
- Query history
- Saved queries
- Categorised queries
- Upload a query
- Create/edit subcorpora

Corpus info
- View corpus metadata
- Corpus documentation
- STTS

---

SaCoCo: diachronic corpus: powered by CQPweb

Standard Query

[pos="VVIMP,.*"]

Query mode: CQP syntax
Number of hits per page: 10
Restriction: None (search whole corpus)

Simple query language syntax

Start query
Reset query
| Page | Corpus Line | Translation |
|------|-------------|-------------|
| 22   | Corpus 1602.23 | Ist so feine Es in einem Schlüssel beseitige sie mit Zucker und | gib sie hin. Ein Füllwirt dem Fleisch zu machen. Wenn du |
| 23   | Corpus 1602.24 | ein obert stein Speck und hacke das wo klein durchschneide und | schlage dem noch drei fünf oder sechs laug dem noch das Fleisch groß ist das |
| 24   | Corpus 1602.25 | oder Essig laß es sieden und wenn die Schüssel weil riemit werden so | treib das so durch ein Toch oder Durchschlag der nicht weirtschén ist thue |
| 25   | Corpus 1602.26 | Rindfleisch die nicht alt zu sehr getreut wurde das gedornen im Mösel und | treib es durch ein Toch und leg die gekochte Brühe daraus wärte sie |
| 26   | Corpus 1602.27 | drauf las sie also drinnen probeln das sie nicht alt zu dick werden | nimm Andern von der Brühe die Rinder und zu dem zu pulver thue das darum wärte es |
| 27   | Corpus 1602.27 | das darum wärte es mit Ingwer Nelken Zimt und Pfeffer nicht viel | gib sie auf die Schüssel das Salz haben ingwercken. Schüppen Fleisch in |
| 28   | Corpus 1602.29 | alt zu in den Weihe thue daraus und geißel weis Brod und | treib das so durch ein Toch wärte es mit Pfeffer Nelken Saffran und |
| 29   | Corpus 1602.4  | man machen ein Häschens Wildpreß von Fischen man mag draus machen kleine | stick und mache drauf ein brann Sohlt und wärte es mit sätleren zugehörung |
| 30   | Corpus 1602.4  | machen ein gansere von Mandeln und Rosenkeln und richte es zu und | gib es hin. Ein Rehen von Fischen. Wibda machen ein Rehnen- |
| 31   | Corpus 1602.5  | so nim Karpen und schappe sie und wenn sie geschlampf sehn so | reiß sie von einander und leg sie in eine Pfanne oder reifen Kessel |
| 32   | Dunkworth 1729.18 | anstalt Speck darauf Musenwelenblum laß es damit hoch kochen dem so | nimm lange Roten gib etwas Weiß dazu ein wenig Sau gerhau Semmel laß |
| 33   | Dunkworth 1729.18 | ein wenig Sau gerhau Semmel laß sie damit kochern dann so | nimm das Fleisch geßt die Brühe davon und die andere wieder darauf mit |
| 34   | Dunkworth 1729.15 | esset kalt davon. 40. Schüssel = Pasteur. 41. Andere Art. | Nonum 2. Pfund mager Kaff = Fleisch koch es so lang bil kein Blut mehr |
| 35   | Dunkworth 1729.4  | auf Kohlen und las es gant kochen. Wenn du konst so | gib ihr unter und oben Feuer wo nicht so muß die Schüssel umgeschlagen |
| 36   | Dunkworth 1729.4  | umgeschlagen werden das das Unterlieb kommt. Wenn es gahr so | gib ausgelassen Butter darüber und bereite es mit Zucker und Zimt. 6. |
| 37   | Dunkworth 1729.4  | die Karpe ab und koche sie mit Salz und Wasser halb gant | gib das Wasser rein davon und thue daran Weiß ein wenig Weiß-Fellig Butter |
## Corpus Querying

### The Imperative Form

| No. | Search result | No. of occurrences | Percent |
|-----|---------------|---------------------|---------|
| 1   | was            | 195                 | 20.81%  |
| 2   | uncomp        | 150                 | 16.01%  |
| 3   | mac           | 65                  | 6.94%   |
| 4   | tell          | 51                  | 5.44%   |
| 5   | dub           | 43                  | 4.59%   |
| 6   | ged           | 39                  | 4.15%   |
| 7   | race          | 33                  | 3.52%   |
| 8   | were          | 27                  | 2.88%   |
| 9   | did           | 21                  | 2.24%   |
| 10  | ret           | 21                  | 2.24%   |
| 11  | leg           | 17                  | 1.81%   |
| 12  | butter        | 15                  | 1.6%    |
| 13  | help          | 14                  | 1.49%   |
| 14  | kiosk         | 13                  | 1.39%   |
| 15  | adder         | 12                  | 1.29%   |
| 16  | miltai        | 11                  | 1.17%   |
| 17  | schac         | 8                   | 0.85%   |
| 18  | watch         | 8                   | 0.85%   |
### Corpus Querying

**The Imperative Form**

#### Distribution breakdown for query "[pos="VVIMP.*"]": this query returned 937 matches in 446 different texts

| Categories: | General Information | Show as: | Bar chart: |
|-------------|---------------------|-----------|------------|
| Category for crosstab: | No crosstabs | Show distribution | No |

#### Based on classification: collection

| Category | contemporary | historical |
|----------|--------------|------------|
| Hits     | 242          | 695        |
| Cat size (MW) | 0.31 | 0.05 |
| Freq per M | 781.35 | 15439.3 |

#### Based on classification: decade

| Category | 1560_1570 | 1570_1580 | 1600_1610 | 1610_1620 | 1640_1650 | 1670_1680 | 1680_1690 | 1690_1700 | 1700_1710 | 1710_1720 | 1720_1730 | 1780_1790 | 1800_1810 | 2000_2010 |
|----------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| Hits     | 144        | 151        | 29         | 19         | 10         | 145        | 3          | 98         | 57         | 31         | 6          | 2          | 0          | 1          |
| Cat size (MW) | 0          | 0          | 0          | 0          | 0.01       | 0.01       | 0          | 0          | 0.01       | 0          | 0          | 0          | 0          | 0          |
| Freq per M | 42427.81   | 50841.75   | 7401.74    | 4853.13    | 6983.24    | 20845.31   | 597.37     | 26322.86   | 18957.06   | 4077.83    | 2054.79    | 2935.86    | 0          | 1223.24    |
Thank you!

SaCoCo: http://fedora.clarin-d.uni-saarland.de/sacoco
Marilisa Amoia m.amoia@mx.uni-saarland.de
José Manuel Martínez j.martinez@mx.uni-saarland.de