Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields

Mohamed Khemakhem, Luca Foppiano, Laurent Romary

Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields

Mohamed Khemakhem[1,2,4], **Luca Foppiano**[1], **Laurent Romary**[1,2,3]

[1]Inria - ALMAnaCH, 2 Rue Simone IFF 75012, Paris
[2] Centre Marc Bloch, Friedrichstrasse 191 10117, Berlin
[3] Berlin-Brandenburgische Akademie der Wissenschaften, Jaegerstrasse 22-23 10117, Berlin
[4]University Paris Diderot, 5 Rue Thomas Mann 75013, Paris
E-mail: mohamed.khemakhem@inria.fr, luca.foppiano@inria.fr, laurent.romary@inria.fr

Abstract

This paper presents an open source machine learning system for structuring dictionaries in digital format into TEI (Text Encoding Initiative) encoded resources. The approach is based on the extraction of overgeneralised TEI structures in a cascading fashion, by means of CRF (Conditional Random Fields) sequence labelling models. Through the experiments carried out on two different dictionary samples, we aim to highlight the strengths as well as the limitations of our approach.

Keywords: automatic structuring, digitized dictionaries, TEI, machine learning, CRF

1. Introduction

An important number of digitized lexical resources remain unexploited due to their unstructured content. Manually structuring such resources is a costly task given their multifold complexity. Our goal is to find an approach to automatically structure digitized dictionaries, independently of the language or the lexicographic school or style. In this paper we present a first version of GROBID-Dictionaries[1], an open source machine learning system for lexical information extraction.

2. Approach

By observing how the lexical information is organised in different paper dictionaries, it is clear that the majority of these lexical resources share the same visual layout to represent the same categories of text information. That served as our starting point to develop our approach for dismantling the content of digitized dictionaries. We tried to build cascading models for automatically extracting TEI (Text Encoding Initiative) (Budin et al., 2012) constructs and make sure that the final output is aligned with current efforts to unify the TEI representations of lexical resources. To be easily adaptable to new dictionary samples, we chose machine learning over rule-based techniques.

2.1 Cascading extraction models

We followed a divide-and-conquer strategy to dismantle text constructs in a digitized dictionary, based initially on observations of their layout. Main pages (see Figure1) in almost any dictionary share three blocks: a header (green), a footer (blue) and a body (orange). The body is, in its turn, made of several entries (red). Each lexical entry can be further broken down (see Figure 2) into: form (green), etymology (blue), sense (red) or/and related entry.

[1] `https://github.com/MedKhem/grobid-dictionaries`

Layout features become less relevant when the segmentation process reaches a deeper information level and we consequently give them up for the corresponding models. The same logic could be applied further for each extracted block, as long as the finest TEI elements are not yet reached. But in the scope of this paper, we focus just on the first six models, details which are given below.

Fig. 1: First and second segmentation levels of a dictionary page

Such a cascading approach ensures a better understanding of the learning process' output and consequently simplifies the feature selection process. Limited exclusive text blocks per level help significantly to diagnose the cause of prediction errors. Moreover, it would be possible to detect and replace early on any irrelevant selected features that can bias a trained model. In such a segmentation, it becomes more straightforward to notice that, for instance, the token position in the page is very relevant to detect headers and footers but has almost no relevance for capturing a sense in a lexical entry, which is very often split over two pages.

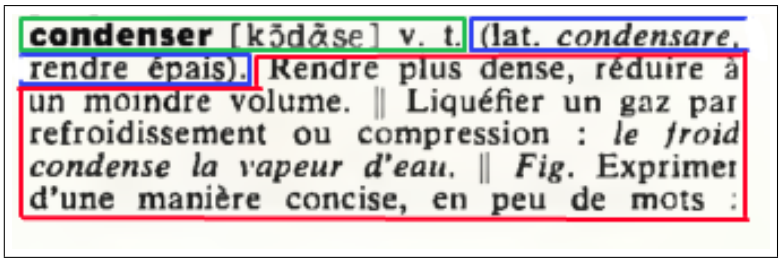


Fig. 2: Example of the segmentation performed by the Lexical Entry model

2.2 Towards a more unified TEI modelling

Our choice for TEI, as the encoding format for the detected structures, is based on its widespread use in lexicographic projects, as well as on some technical factors which will be detailed in the following section. The domination of the lexicographic landscape by TEI is due to the fact that this initiative has provided the lexicographic community with diverse alternatives for encoding different kinds of lexical resources, as well as for modelling the same lexical information. However, the flexibility that this standard ensures has led to an explosion of TEI schemes and, consequently, limited the possibilities for exchange and exploitation.

Our cascading models are conceived in a way to support the encoding of the detected structures in multiple TEI schemes. But to avoid falling into the diversity trap, we are adopting a format that generalises over existing encoding practices. The final scheme has not yet been finalised, but we are continuously refining our guidelines as we move deeper with our models and apply them to new dictionary samples. We are aiming to ensure a maximal synchronisation with existing research efforts in this direction, by collaborating with COST ENeL and ISO committee TC 37/SC 4.

Presenting the details of our encoding choices is beyond the scope of this paper, since we are still shaping them, especially for fine grained information. But we aim to highlight about some constraining decisions we made for the upper levels, to give an idea about our modelling direction. A lexical entry, for instance, is always encoded using <entry> exclusively, which means we do not make use of any possible alternatives, such as <superEntry> and <entryFree>. The semantic loss is not important in this case, since the nature of the entry could be inferred from the elements it contains. As for lexical entries, they can be completely encoded using 5 main elements: <form> for morphological and grammatical information of the whole entry, <etym> for etymological information, <sense> for semantic and syntactic information, <re> for related entries and <dictScrap> for any text that does not belong to the previous elements. Note here that we are trying to use the more generic elements to encode the lexical information in each level, which will be more refined in the following levels.

3. GROBID-Dictionaries

To implement our approach, we took up the available infrastructure from GROBID (Lopez and Romary, 2015) and we adapted it to the specificity of the use case of digitized dictionaries.

3.1 GROBID

GROBID (GeneRation Of Bibliographic Data) is a machine learning system for parsing and extracting bibliographic metadata from scholar articles, mainly text documents in PDF format. It relies on CRF (Lavergne et al., 2010) to perform a multi-level sequence labelling of text blocks in a cascade fashion which are then extracted and encoded in TEI elements. Such an approach has been very accurate for that use case and the system's Java API has been one of the most used by bibliography research platforms and research bodies worldwide.

We have been struck by the analogy between the structures that can be extracted by GROBID, in the case of full scientific articles, and the actual constructs we wanted to extract from a digitized dictionary. At its first extraction level, GROBID detects the main blocks of a paper such as the header, the body, the references, annexes, etc. These main parts will be further structured at the following level, like the references which will be extracted in separate items and then parsed one by one to detect the titles, the authors and the other publication details. By recalling the segmentation steps presented in the previous section, there is a clear analogy between the case of a reference in a scientific document and a lexical entry in a dictionary.

This correspondence is reinforced by the fact that GROBID relies on layout, as well as text features, to perform the supervised classification of the parsed text and generates a TEI compliant encoding where the various segmentation levels are associated with an appropriate XML tessellation.

3.2 GROBID-Dictionaries

Due to the above-mentioned similarities, we undertook the adaptation of GROBID for the case of digitized dictionaries in order to build a system, which uses the core utilities of GROBID and applies them for lexical information processing. In building GROBID-Dictionaries, we faced several challenges, the three major ones being detailed in the following.

3.2.1 TEI cascade modelling

After having fully encoded a lexical entry, the task became more specific and more challenging when it comes to defining the TEI structures to be extracted by each model. It is a question of finding the appropriate mapping between the TEI elements and the labels to be set for the models that share the task of structuring the text in cascade. In addition, the process is at the same time constrained by the need to avoid having structures from different hierarchy levels being extracted at once. In fact, the CRF models, as they could be used from GROBID core, do not allow the labelling of nested text sequences. We clarify this technical point by explaining how the sequence labelling process works in the case of segmenting a lexical entry.

The following matrix represents the set of feature vectors corresponding to the lexical entry *condenser*, which will be labelled by a first version of the "Lexical Entry" model. The latter has the task of detecting the 5 main blocks in a lexical entry, if they exist. For

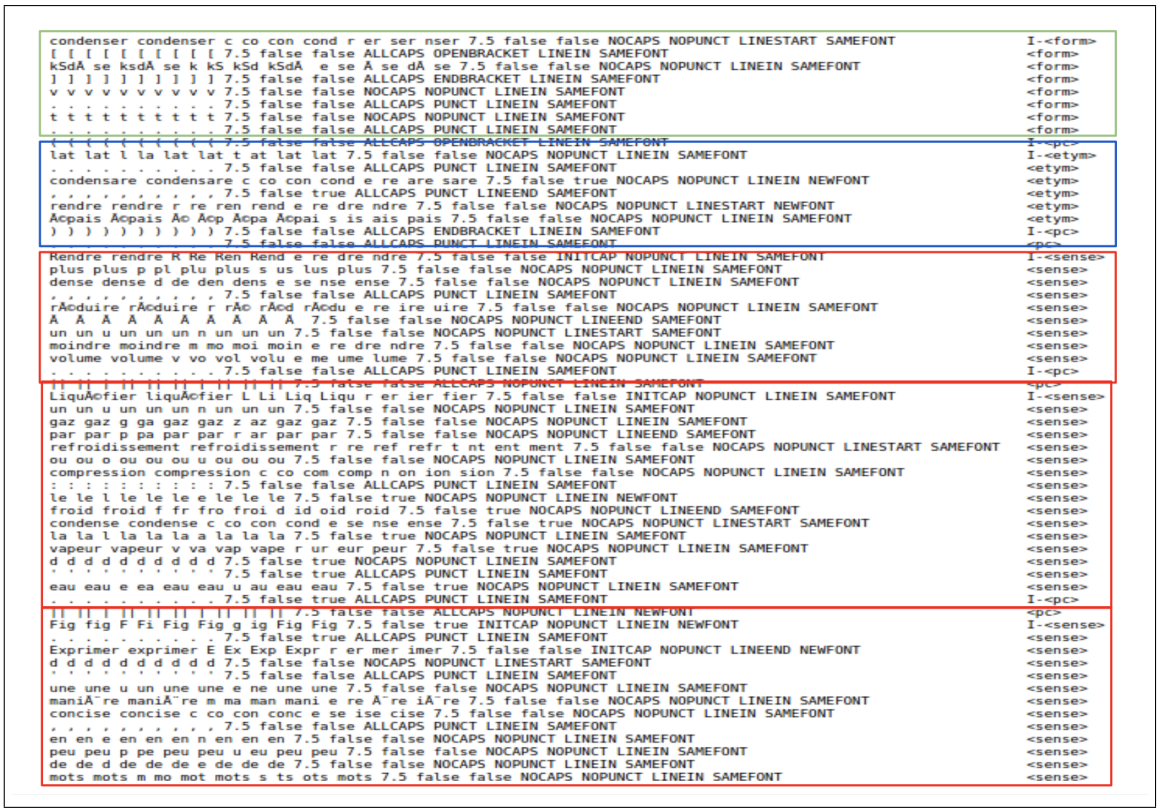


Fig. 3: Sequence labelling using a first version of the "Lexical Entry" segmentation model

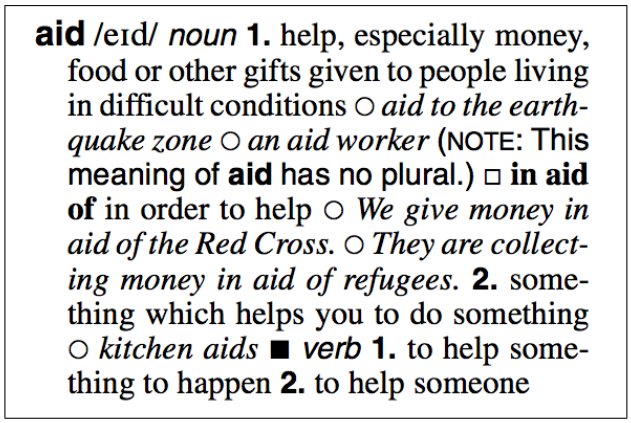
the sense information, the model has been trained to extract each parsed text sequence representing a sense.

Each vertical column is a specific feature for all the tokens of the lexical entry and each horizontal line corresponds to all the features of each token. For this model, a set of features is going to be assigned to each token based on criteria we chose in the feature selection process. In the second phase, comes the role of the trained model to give a prediction of a suitable label for each token, based on all its feature values. A structure corresponds then to the sequence of tokens having the same label, where the *I-Label* marks the beginning of a new sequence. Following this technique, it is obviously not possible in this model to structure the example *"le froid condense la vapeur d'eau"*(see Figures 2 and 3) in the sense, since just one label is allowed per token. Therefore, the segmentation of the examples should be delegated to another model that follows the current one.

3.2.2 Sample annotation

This is the phase where the previous rules will be applied on different instances, to annotate data for training the models. An adjustment of the directives is necessary to make the models more general, as soon as new instances appear to show the modelling limits of our current guidelines. To illustrate such a case, we could take the example of the previously defined "Lexical Entry" model and apply it to the lexical entry *aid*.

The TEI encoding for this entry with the "Lexical Entry" model is the following (see Figure 5):



aid /eɪd/ *noun* **1.** help, especially money, food or other gifts given to people living in difficult conditions ○ *aid to the earthquake zone* ○ *an aid worker* (NOTE: This meaning of **aid** has no plural.) □ **in aid of** in order to help ○ *We give money in aid of the Red Cross.* ○ *They are collecting money in aid of refugees.* **2.** something which helps you to do something ○ *kitchen aids* ■ *verb* **1.** to help something to happen **2.** to help someone

Fig. 4: Lexical entry having more than one POS

```
<entry>
    <form>aid /e□d/ noun</form>
    <sense>1. help, especially money, <lb/>food or other gifts given to people living <lb/>
        in difficult conditions aid to the earth-<lb/>quake zone an aid worker (NOTE: This <lb/>
        meaning of aid has no plural.) in aid <lb/>of in order to help We give money in <lb/>
        aid of the Red Cross. They are collect-<lb/>ing money in aid of refugees</sense><pc>. </pc>
    <sense>2. some-<lb/>thing which helps you to do something <lb/>kitchen aids</sense> í
    <sense>verb 1. to help some-<lb/>thing to happen</sense>
    <sense>2. to help someone</sense>
</entry>
```

Fig. 5: Structured output of the "Lexical Entry" model's primary version

We could notice that the model presented in Figure 3 is no longer valid to perform the segmentation of senses aggregated by part of speech (POS), with respect to avoiding nested constructs. This issue could be fixed by having a first model that does not find the boundaries of the senses of a part of speech in this level.

```
<entry>
    <form>aid /e□d/ noun</form>
    <sense>1. help, especially money, <lb/>food or other gifts given to people living <lb/>
in difficult conditions aid to the earth-<lb/>quake zone an aid worker (NOTE: This <lb/>
meaning of aid has no plural.) in aid <lb/>of in order to help We give money in <lb/>
aid of the Red Cross. They are collect-<lb/>ing money in aid of refugees.2. some-<lb/>
thing which helps you to do something <lb/>kitchen aids</sense> í
    <sense>verb 1. to help some-<lb/>thing to happen 2. to help someone</sense>
</entry>
```

Fig. 6: Structured output of the "Lexical Entry" model's adjusted version

This segmentation of main POS-aggregated senses should be performed by a second model, called "Sense" for example, to find the limits of each sense as well as any grammatical information, if any exists.

The labelling and extraction of the TEI structures should be performed further for the other blocks, by following the same approach. For the case of the *aid* entry, a dedicated model should be used to segment the <form> block by extracting the morphological and grammatical information and decide about of the parent of the latter. In the current case, the <gramGrp> will be the direct child node of the entry, since it carries information

```
<entry>
    <form>aid /eɪd/ noun</form>
    <sense>
        <sense>1. help, especially money, <lb/>food or other gifts given to people living <lb/>
            in difficult conditions aid to the earth-<lb/>quake zone an aid worker (NOTE: This
            <lb/>meaning of aid has no plural.) in aid <lb/>of in order to help We give money in <lb/>
            aid of the Red Cross. They are collect-<lb/>ing money in aid of refugees</sense><pc>. </pc>
        <sense>2. some-<lb/>thing which helps you to do something <lb/>kitchen aids</sense> í
    </sense>
    <sense>
        <gramGrp>verb</gramGrp>
        <sense>1. to help some-<lb/>thing to happen</sense>
        <sense>2. to help someone</sense>
    </sense>
</entry>
```

Fig. 7: Structured output of the "Sense" model

about the sense of the entry given a POS, and not about the lemma. The <gramGrp> block will, in its turn, have another specific model to structure its content. Figure 8 shows the final output generated by our cascading model tree.

```
<entry>
   <form type="lemma">
      <orth>aid</orth>
      <pron>/eɪd/</pron>
   </form>
   <gramGrp>
      <pos>noun</pos>
   </gramGrp>
   <sense>
      <sense>1. help, especially money, <lb/>food or other gifts given to people living <lb/>
         in difficult conditions aid to the earth-<lb/>quake zone an aid worker (NOTE: This
         <lb/>meaning of aid has no plural.) in aid <lb/>of in order to help We give money in <lb/>
         aid of the Red Cross. They are collect-<lb/>ing money in aid of refugees</sense><pc>. </pc>
      <sense>2. some-<lb/>thing which helps you to do something <lb/>kitchen aids</sense> í
   </sense>
   <sense>
      <gramGrp>
         <pos>verb</pos>
      </gramGrp>
      <sense>1. to help some-<lb/>thing to happen</sense>
      <sense>2. to help someone</sense>
   </sense>
</entry>
```

Fig. 8: Final output of all the models

Annotation guidelines seem to be mandatory here to guide the process since an annotator, especially with a linguistic or lexicographic background, could be easily biased by the TEI practices and tags which are used differently in our cascading approach but will converge in the final output. We noticed this issue after having lexicographers annotate a few samples and we therefore, defined a first version of the guidelines[2], which we are actively maintaining.

[2] https://github.com/MedKhem/grobid-dictionaries/wiki/How-to-Annotate%3F

3.2.3 Feature selection

In this phase, the cumulated data will be used for generating features that will be used by the models to discriminate between their labels. For the first model, we kept the line based features used in GROBID's first model[3]. Our choice was based simply on the assumption of the general nature of such features. Moreover, the experiments on several samples showed a high and fast performance.

As explained in our approach, we tried to rely on a restricted list of features for the rest of the models, where we drop the ones that are most likely to produce bias. We chose to use features on the token level to structure the lexical information. For the first version of our system, we are experimenting the use of one list with 16 features[4]: 8 based on the text and the rest carrying the layout aspects of each token, such as the change of font or line breaks.

4. Experiments

4.1 Models

The resulting models and their corresponding labels are the following:

- **Dictionary Segmentation:** This is the first model and has as its goal the segmentation of each dictionary page into three main blocks, where each block corresponds to a TEI label: <fw type="header"> for information in the header, <ab type="page"> for all the text in the body of a page and <fw type="footer"> for footer information. For the sake of simplicity, for training the models (see Section 4.3) we use: <headnote> to refer to <fw type="header">, <body> referring to <ab type="page"> and <footer> to refer to <fw type="footer">. But we respect the original labels for the final TEI output.

- **Dictionary Body Segmentation:** The second model gets the page body, recognized by the first model, and processes it to recognize the boundaries of each lexical entry by labelling each sequence with <entry> label.

- **Lexical Entry:** The third model parses each lexical entry, recognized by the second model, to segment it into four main blocks: <form> for morphological and grammatical information, <etym> for etymology, <sense> for all sense information, <re> for related entries.

- **Form:** This model analyses the <form> block, generated by the Lexical Entry model, and segments its contained information . We have for the moment three labels for this model: <orth> for the lemma, <pron> for pronounciation and <gramGrp> for grammatical information, such as part of speech, gender, number,.etc.

- **Sense:** The Sense model has two goals. First, to extract the grammatical information <gramGrp>, that could exist. Second, to segment the first level senses, by structuring them in <sense> sequences.

[3] https://github.com/kermitt2/grobid/blob/master/grobid-core/src/main/java/org/grobid/core/features/FeaturesVectorSegmentation.java

[4] https://github.com/MedKhem/grobid-dictionaries/blob/master/src/main/java/org/grobid/core/features/FeatureVectorLexicalEntry.java

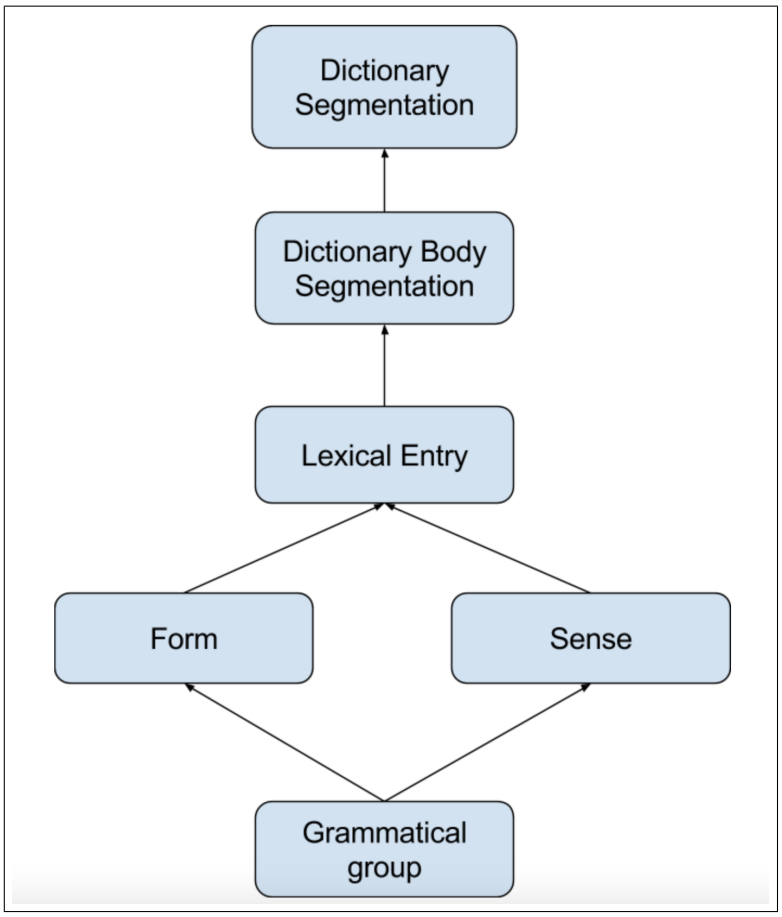


Fig. 9: Selected models

- **Grammatical group:** The last model in our temporary hierarchy has the of segmenting the grammatical information <gramGrp>, extracted by previous models

For each model, we reserved two extra labels: <pc> for punctuation such as separators between text information or any markup text. A second label, <dictScrap>, is used to contain any information that couldn't be classified in one of the main labels of the model.

4.2 Lexical Samples

We carried out our experiments by applying our models to several dictionaries and given the inconstancy that some presented, mainly due to digitization issues, we selected two resources that represent several differences on many levels.

4.2.1 Digital dictionary

"Easier English Basic Dictionary" (EEBD, 2009) is a monolingual dictionary for English which contains over 5,000 entries, published in 2009. For our experiments, we used the 370 pages containing the body of the dictionary. The version which we used, is a digitally born one. In other words, no OCR processing has been performed to generate the resource in its electronic format. As Figure 10 illustrates, the dictionary has a very modern and basic layout and its markup system is spread over the entries to mark the transition of the lexical information presented. We chose this digital sample to be our baseline, since it contains very clean text and clear lexical information modelling.

4.2.2 Digitized dictionary

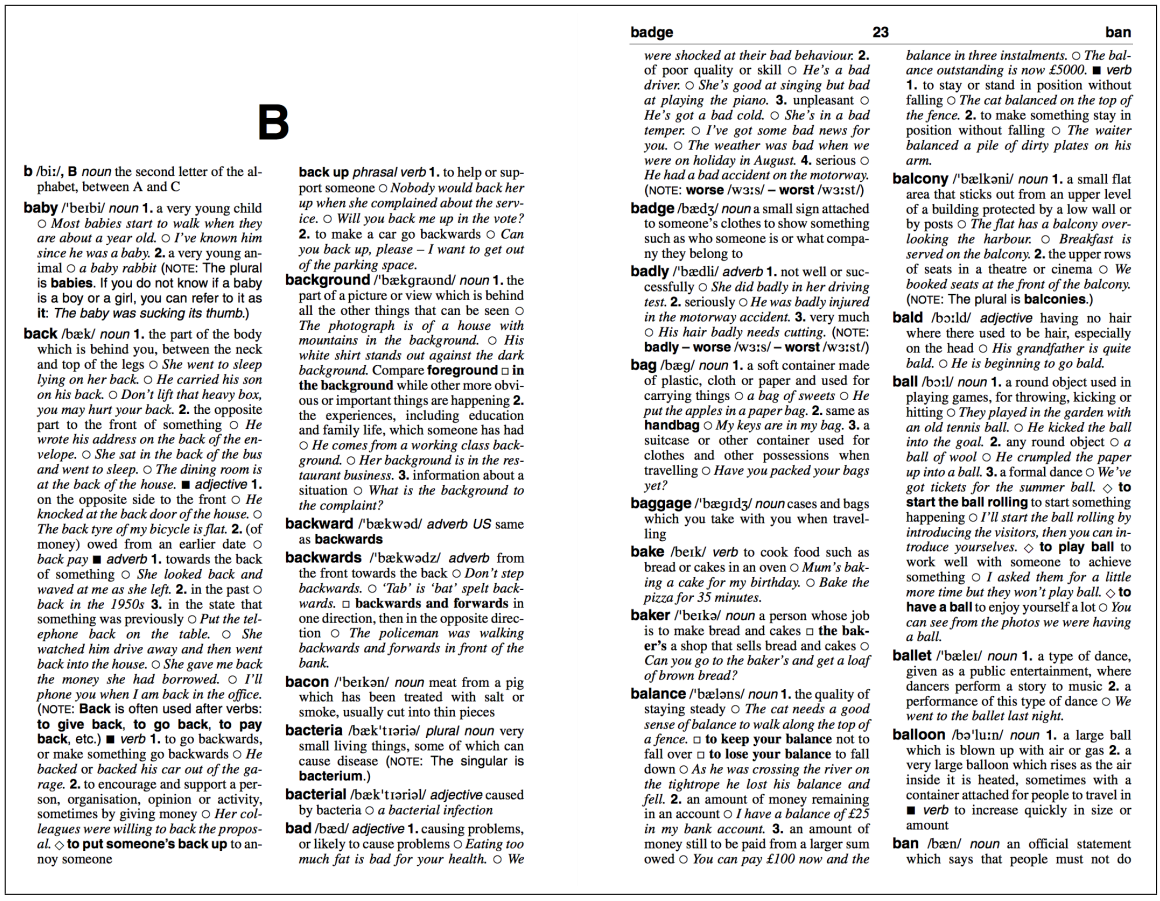
To take the experiments to the next level, we chose a dictionary that has been OCRized and that encloses totally different lexical information. The dictionary was published in 1964 but later digitized. The version we have is of relatively good quality but still presents some anomalies, where some text blocks are unextractable from the PDF.

The Fang-French & French-Fang dictionary (Galley, 1964) is a bilingual dictionary having over 500 pages of lexical entries split into two parts. As Figure 11 shows, the markup system is totally different from the EEBD, where field transition is mostly marked with a change of font rather than with specific markers. For our experiments, we worked on the first part, Fang-French Dictionary (FFD), containing over 390 pages

4.3 Results

For the sake of conciseness, in this paper we present an evaluation of just 4 selected models out of 6 implemented, for each dictionary. We used the benchmark module provided by GROBID to measure the precision, recall and F1 scores.

In the following tables, token level gathers the measures for each different token, field level is for each continuous sequence of the same label (so a field, a sequence of several tokens which all belongs to the same labelled chunk, e.g. a lexical entry).



B

b /biː/, **B** *noun* the second letter of the alphabet, between A and C

baby /'beɪbi/ *noun* **1.** a very young child ○ *Most babies start to walk when they are about a year old.* ○ *I've known him since he was a baby.* **2.** a very young animal ○ *a baby rabbit* (NOTE: The plural is **babies**. If you do not know if a baby is a boy or a girl, you can refer to it as **it**: *The baby was sucking its thumb.*)

back /bæk/ *noun* **1.** the part of the body which is behind you, between the neck and top of the legs ○ *She went to sleep lying on her back.* ○ *He carried his son on his back.* ○ *Don't lift that heavy box, you may hurt your back.* **2.** the opposite part to the front of something ○ *He wrote his address on the back of the envelope.* ○ *She sat in the back of the bus and went to sleep.* ○ *The dining room is at the back of the house.* ■ *adjective* **1.** on the opposite side to the front ○ *He knocked at the back door of the house.* ○ *The back tyre of my bicycle is flat.* **2.** (of money) owed from an earlier date ○ *back pay* ■ *adverb* **1.** towards the back of something ○ *She looked back and waved at me as she left.* **2.** in the past ○ *back in the 1950s* **3.** in the state that something was previously ○ *Put the telephone back on the table.* ○ *She watched him drive away and then went back into the house.* ○ *She gave me back the money she had borrowed.* ○ *I'll phone you when I am back in the office.* (NOTE: **Back** is often used after verbs: **to give back**, **to go back**, **to pay back**, etc.) ■ *verb* **1.** to go backwards, or make something go backwards ○ *He backed* or *backed his car out of the garage.* **2.** to encourage and support a person, organisation, opinion or activity, sometimes by giving money ○ *Her colleagues were willing to back the proposal.* ◇ **to put someone's back up** to annoy someone

back up *phrasal verb* **1.** to help or support someone ○ *Nobody would back her up when she complained about the service.* ○ *Will you back me up in the vote?* **2.** to make a car go backwards ○ *Can you back up, please – I want to get out of the parking space.*

background /'bækgraʊnd/ *noun* **1.** the part of a picture or view which is behind all the other things that can be seen ○ *The photograph is of a house with mountains in the background.* ○ *His white shirt stands out against the dark background.* Compare **foreground** □ **in the background** while other more obvious or important things are happening **2.** the experiences, including education and family life, which someone has had ○ *He comes from a working class background.* ○ *Her background is in the restaurant business.* **3.** information about a situation ○ *What is the background to the complaint?*

backward /'bækwəd/ *adverb US* same as **backwards**

backwards /'bækwədz/ *adverb* from the front towards the back ○ *Don't step backwards.* ○ *'Tab' is 'bat' spelt backwards.* □ **backwards and forwards** in one direction, then in the opposite direction ○ *The policeman was walking backwards and forwards in front of the bank.*

bacon /'beɪkən/ *noun* meat from a pig which has been treated with salt or smoke, usually cut into thin pieces

bacteria /bæk'tɪəriə/ *plural noun* very small living things, some of which can cause disease (NOTE: The singular is **bacterium**.)

bacterial /bæk'tɪəriəl/ *adjective* caused by bacteria ○ *a bacterial infection*

bad /bæd/ *adjective* **1.** causing problems, or likely to cause problems ○ *Eating too much fat is bad for your health.* ○ *We were shocked at their bad behaviour.* **2.** of poor quality or skill ○ *He's a bad driver.* ○ *She's good at singing but bad at playing the piano.* **3.** unpleasant ○ *He's got a bad cold.* ○ *She's in a bad temper.* ○ *I've got some bad news for you.* ○ *The weather was bad when we were on holiday in August.* **4.** serious ○ *He had a bad accident on the motorway.* (NOTE: **worse** /wɜːs/ **– worst** /wɜːst/)

badge /bædʒ/ *noun* a small sign attached to someone's clothes to show something such as who someone is or what company they belong to

badly /'bædli/ *adverb* **1.** not well or successfully ○ *She did badly in her driving test.* **2.** seriously ○ *He was badly injured in the motorway accident.* **3.** very much ○ *His hair badly needs cutting.* (NOTE: **badly – worse** /wɜːs/ **– worst** /wɜːst/)

bag /bæg/ *noun* **1.** a soft container made of plastic, cloth or paper and used for carrying things ○ *a bag of sweets* ○ *He put the apples in a paper bag.* **2.** same as **handbag** ○ *My keys are in my bag.* **3.** a suitcase or other container used for clothes and other possessions when travelling ○ *Have you packed your bags yet?*

baggage /'bægɪdʒ/ *noun* cases and bags which you take with you when travelling

bake /beɪk/ *verb* to cook food such as bread or cakes in an oven ○ *Mum's baking a cake for my birthday.* ○ *Bake the pizza for 35 minutes.*

baker /'beɪkə/ *noun* a person whose job is to make bread and cakes □ **the baker's** a shop that sells bread and cakes ○ *Can you go to the baker's and get a loaf of brown bread?*

balance /'bæləns/ *noun* **1.** the quality of staying steady ○ *The cat needs a good sense of balance to walk along the top of a fence.* □ **to keep your balance** not to fall over □ **to lose your balance** to fall down ○ *As he was crossing the river on the tightrope he lost his balance and fell.* **2.** an amount of money remaining in an account ○ *I have a balance of £25 in my bank account.* **3.** an amount of money still to be paid from a larger sum owed ○ *You can pay £100 now and the balance in three instalments.* ○ *The balance outstanding is now £5000.* ■ *verb* **1.** to stay or stand in position without falling ○ *The cat balanced on the top of the fence.* **2.** to make something stay in position without falling ○ *The waiter balanced a pile of dirty plates on his arm.*

balcony /'bælkəni/ *noun* **1.** a small flat area that sticks out from an upper level of a building protected by a low wall or by posts ○ *The flat has a balcony overlooking the harbour.* ○ *Breakfast is served on the balcony.* **2.** the upper rows of seats in a theatre or cinema ○ *We booked seats at the front of the balcony.* (NOTE: The plural is **balconies**.)

bald /bɔːld/ *adjective* having no hair where there used to be hair, especially on the head ○ *His grandfather is quite bald.* ○ *He is beginning to go bald.*

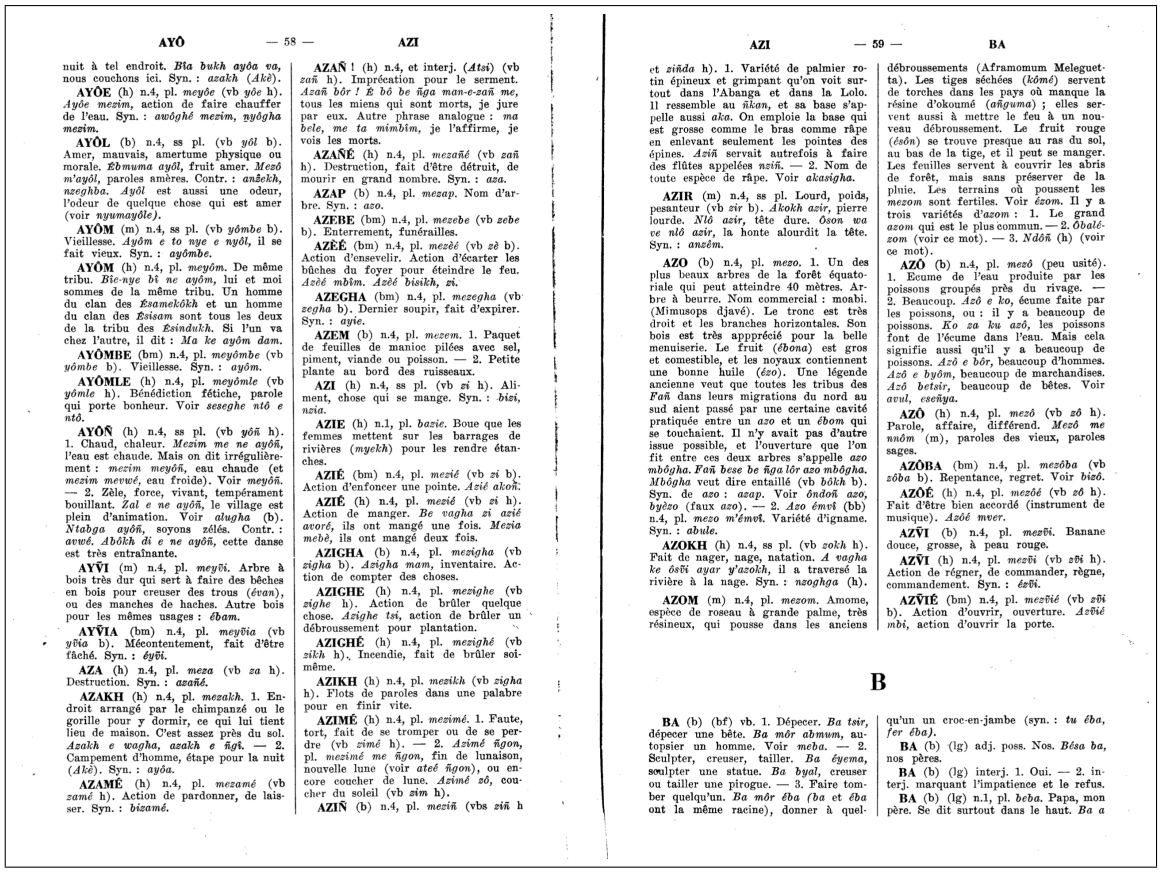
ball /bɔːl/ *noun* **1.** a round object used in playing games, for throwing, kicking or hitting ○ *They played in the garden with an old tennis ball.* ○ *He kicked the ball into the goal.* **2.** any round object ○ *a ball of wool* ○ *He crumpled the paper up into a ball.* **3.** a formal dance ○ *We've got tickets for the summer ball.* ◇ **to start the ball rolling** to start something happening ○ *I'll start the ball rolling by introducing the visitors, then you can introduce yourselves.* ◇ **to play ball** to work well with someone to achieve something ○ *I asked them for a little more time but they won't play ball.* ◇ **to have a ball** to enjoy yourself a lot ○ *You can see from the photos we were having a ball.*

ballet /'bæleɪ/ *noun* **1.** a type of dance, given as a public entertainment, where dancers perform a story to music **2.** a performance of this type of dance ○ *We went to the ballet last night.*

balloon /bə'luːn/ *noun* **1.** a large ball which is blown up with air or gas **2.** a very large balloon which rises as the air inside it is heated, sometimes with a container attached for people to travel in ■ *verb* to increase quickly in size or amount

ban /bæn/ *noun* an official statement which says that people must not do

Fig. 10: Two pages from EEBD side by side



AYÔ — 58 — AZI

nuit à tel endroit. *Bia bukh ayôa va*, nous couchons ici. Syn. : *azakh* (*Akè*).

AYÔE (h) n.4, pl. *meyôe* (vb *yôe* h). *Ayôe mezim*, action de faire chauffer de l'eau. Syn. : *awôghé mezim*, *nyôgha mezim*.

AYÔL (b) n.4, ss pl. (vb *yôl* b). Amer, mauvais, amertume physique ou morale. *Ébmuma ayôl*, fruit amer. *Mezó m'ayôl*, paroles amères. Contr. : *anôckh*, *nzoghba*. *Ayôl* est aussi une odeur, l'odeur de quelque chose qui est amer (voir *nyumayôle)*.

AYÔM (m) n.4, ss pl. (vb *yômbe* b). Vieillesse. *Ayôm e to nye e nyôl*, il se fait vieux. Syn. : *ayômbe*.

AYÔM (h) n.4, pl. *meyôm*. De même tribu. *Bie-nye bî ne ayôm*, lui et moi sommes de la même tribu. Un homme du clan des *Ezamekôkh* et un homme du clan des *Esisam* sont tous les deux de la tribu des *Esindukh*. Si l'un va chez l'autre, il dit : *Ma ke ayôm dam*.

AYÔMBE (bm) n.4, pl. *meyômbe* (vb *yômbe* b). Vieillesse. Syn. : *ayôm*.

AYÔMLE (h) n.4, pl. *meyômle* (vb *yômle* h). Bénédiction fétiche, parole qui porte bonheur. Voir *seseghe ntó e ntô*.

AYÔÑ (h) n.4, ss pl. (vb *yôñ* h). 1. Chaud, chaleur. *Mezim me ne ayôñ*, l'eau est chaude. Mais on dit irrégulièrement : *mezim meyôñ*, eau chaude (et *mezim mevwé*, eau froide). Voir *meyôñ*. — 2. Zèle, force, vivant, tempérament bouillant. *Żal e ne ayôñ*, le village est plein d'animation. Voir *alugha* (b). *Ntabga ayôñ*, soyons zélés. Contr. : *avwé*. *Abôkh dî e ne ayôñ*, cette danse est très entraînante.

AYŸI (m) n.4, pl. *meyŸi*. Arbre à bois très dur qui sert à faire des bêches en bois pour creuser des trous (*évan*), ou des manches de haches. Autre bois pour les mêmes usages : *ébam*.

AYŸIA (bm) n.4, pl. *meyŸia* (vb *yŸia* b). Mécontentement, fait d'être fâché. Syn. : *éyŸi*.

AZA (h) n.4, pl. *meza* (vb *za* h). Destruction. Syn. : *azaké*.

AZAKH (h) n.4, pl. *mezakh*. 1. Endroit arrangé par le chimpanzé ou le gorille pour y dormir, ce qui lui tient lieu de maison. C'est assez près du sol. *Azakh e wagha, azakh e ñgô*. — 2. Campement d'homme, étape pour la nuit (*Akè*). Syn. : *ayôa*.

AZAMÉ (h) n.4, pl. *mezamé* (vb *zamé* h). Action de pardonner, de laisser. Syn. : *bizamé*.

AZAÑ ! (h) n.4, et interj. (*Atsî*) (vb *zañ* h). Imprécation pour le serment. *Azañ bôr ! É bô be ñga man-e-zañ me*, tous les miens qui sont morts, je jure par eux. Autre phrase analogue : *ma bele, me ta mimbîm*, je l'affirme, je vois les morts.

AZAÑÉ (h) n.4, pl. *mezañé* (vb *zañ* h). Destruction, fait d'être détruit, de mourir en grand nombre. Syn. : *aza*.

AZAP (b) n.4, pl. *mezap*. Nom d'arbre. Syn. : *azo*.

AZEBE (bm) n.4, pl. *mezebe* (vb *zebe* b). Enterrement, funérailles.

AZÊÉ (bm) n.4, pl. *mezêé* (vb *zê* b). Action d'ensevelir. Action d'écarter les bûches du foyer pour éteindre le feu. *Azêé mbîm. Azêé bisikh, zi*.

AZEGHA (bm) n.4, pl. *mezegha* (vb *zegha* b). Dernier soupir, fait d'expirer. Syn. : *ayie*.

AZEM (b) n.4, pl. *mezem*. 1. Paquet de feuilles de manioc pilées avec sel, piment, viande ou poisson. — 2. Petite plante au bord des ruisseaux.

AZI (h) n.4, ss pl. (vb *zi* h). Aliment, chose qui se mange. Syn. : *bisi*, *nzia*.

AZIE (h) n.1, pl. *bazie*. Boue que les femmes mettent sur les barrages de rivières (*myelch*) pour les rendre étanches.

AZIÉ (bm) n.4, pl. *mezié* (vb *zi* b). Action d'enfoncer une pointe. *Azié akoñ*.

AZIÉ (h) n.4, pl. *mezié* (vb *zi* h). Action de manger. *Be vagha zi azié avoré*, ils ont mangé une fois. *Mezia mebè*, ils ont mangé deux fois.

AZIGHA (b) n.4, pl. *mezigha* (vb *zigha* b). *Azigha mam*, inventaire. Action de compter des choses.

AZIGHE (h) n.4, pl. *mezighe* (vb *zighe* h). Action de brûler quelque chose. *Azighe tsi*, action de brûler un débroussement pour plantation.

AZIGHÉ (h) n.4, pl. *mezighé* (vb *zikh* h). Incendie, fait de brûler soi-même.

AZIKH (h) n.4, pl. *mezikh* (vb *zigha* h). Flots de paroles dans une palabre pour en finir vite.

AZIMÉ (h) n.4, pl. *mezimé*. 1. Faute, tort, fait de se tromper ou de se perdre (vb *zimé* h). — 2. *Azimé ñgon*, pl. *mezimé me ñgon*, fin de lunaison, nouvelle lune (voir *até ñgon*), ou encore coucher de lune. *Azimé sô*, coucher du soleil (vb *zim* h).

AZIÑ (b) n.4, pl. *meziñ* (vbs *ziñ* h

AZI — 59 — BA

et *ziñda* h). 1. Variété de palmier rotin épineux et grimpant qu'on voit surtout dans l'Abanga et dans la Lolo. Il ressemble au *ñkan*, et sa base s'appelle aussi *aka*. On emploie la base qui est grosse comme le bras comme râpe en enlevant seulement les pointes des épines. *Aziñ* servait autrefois à faire des flûtes appelées *nziñ*. — 2. Nom de toute espèce de râpe. Voir *akasigha*.

AZIR (m) n.4, ss pl. Lourd, poids, pesanteur (vb *zir* b). *Akokh azir*, pierre lourde. *Nlô azir*, tête dure. *Ósen wa ve nlô azir*, la honte alourdit la tête. Syn. : *anzêm*.

AZO (b) n.4, pl. *mezo*. 1. Un des plus beaux arbres de la forêt équatoriale qui peut atteindre 40 mètres. Arbre à beurre. Nom commercial : moabi. (Mimusops djavé). Le tronc est très droit et les branches horizontales. Son bois est très apppréció pour la belle menuiserie. Le fruit (*ébóna*) est gros et comestible, et les noyaux contiennent une bonne huile (*ézo*). Une légende ancienne veut que toutes les tribus des *Fañ* dans leurs migrations du nord au sud aient passé par une certaine cavité pratiquée entre un *azo* et un *ébom* qui se touchaient. Il n'y avait pas d'autre issue possible, et l'ouverture que l'on fit entre ces deux arbres s'appelle *azo mbôgha*. *Fañ bese be ñga lôr azo mbôgha*. *Mbôgha* veut dire entaillé (vb *bôkh* b). Syn. de *azo* : *azap*. Voir *ôndoñ azo*, *byèzo* (faux *azo*). — 2. *Azo émvî* (bb) n.4, pl. *mezo m'émvî*. Variété d'igname. Syn. : *abule*.

AZOKH (h) n.4, ss pl. (vb *zokh* h). Fait de nager, nage, natation. *A vagha ke êêlî ayar y'azokh*, il a traversé la rivière à la nage. Syn. : *nzoghga* (h).

AZOM (m) n.4, pl. *mezom*. Amome, espèce de roseau à grande palme, très résineux, qui pousse dans les anciens débroussements (Aframomum Meleguetta). Les tiges séchées (*kômé*) servent de torches dans les pays où manque la résine d'okoumé (*añguma*) ; elles servent aussi à mettre le feu à un nouveau débroussement. Le fruit rouge (*ésón*) se trouve presque au ras du sol, au bas de la tige, et il peut se manger. Les feuilles servent à couvrir les abris de forêt, mais sans préserver de la pluie. Les terrains où poussent les *mezom* sont fertiles. Voir *ézom*. Il y a trois variétés d'*azom* : 1. Le grand *azom* qui est le plus commun. — 2. *Óbalézom* (voir ce mot). — 3. *Ndôñ* (h) (voir ce mot).

AZÓ (b) n.4, pl. *mezó* (peu usité). 1. Ecume de l'eau produite par les poissons groupés près du rivage. — 2. Beaucoup. *Azó e ko*, écume faite par les poissons, ou : il y a beaucoup de poissons. *Ko za ku azó*, les poissons font de l'écume dans l'eau. Mais cela signifie aussi qu'il y a beaucoup de poissons. *Azó e bôr*, beaucoup d'hommes. *Azó e byôm*, beaucoup de marchandises. *Azó betsir*, beaucoup de bêtes. Voir *avul*, *eseñya*.

AZÓ (h) n.4, pl. *mezó* (vb *zó* h). Parole, affaire, différend. *Mezó me nnôm* (m), paroles des vieux, paroles sages.

AZÓBA (bm) n.4, pl. *mezóba* (vb *zóba* b). Repentance, regret. Voir *bizó*.

AZÓÉ (h) n.4, pl. *mezóé* (vb *zó* h). Fait d'être bien accordé (instrument de musique). *Azóé mver*.

AZŸI (b) n.4, pl. *mezŸi*. Banane douce, grosse, à peau rouge.

AZŸI (h) n.4, pl. *mezŸi* (vb *zŸi* h). Action de régner, de commander, règne, commandement. Syn. : *ézŸi*.

AZŸIÉ (bm) n.4, pl. *mezŸié* (vb *zŸi* b). Action d'ouvrir, ouverture. *AzŸié mbi*, action d'ouvrir la porte.

B

BA (b) (bf) vb. 1. Dépecer. *Ba tsir*, dépecer une bête. *Ba môr abmum*, autopsier un homme. Voir *meba*. — 2. Sculpter, creuser, tailler. *Ba éyema*, sculpter une statue. *Ba byal*, creuser ou tailler une pirogue. — 3. Faire tomber quelqu'un. *Ba môr éba (ba* et *éba* ont la même racine), donner à quelqu'un un croc-en-jambe (syn. : *tu éba*, *fer éba)*.

BA (b) (lg) adj. poss. Nos. *Bésa ba*, nos pères.

BA (b) (lg) interj. 1. Oui. — 2. interj. marquant l'impatience et le refus.

BA (b) (lg) n.1, pl. *beba*. Papa, mon père. Se dit surtout dans le haut. *Ba a*

Fig. 11: Two pages from FFD side by side

4.3.1 Dictionary Segmentation

For both dictionaries, we annotated 7 pages, which we split into 4 for training and 3 for evaluation.

4.3.2 Dictionary Body Segmentation

For EEBD, we annotated 5 pages, which we split into 50 lexical entries for training and 27 for evaluation. For FFD, we annotated 7 pages with 91 lexical entries for training and 45 for evaluation.

4.3.3 Lexical Entry

For EEBD, we annotated 8 pages, which we split into 76 entries for training and 24 for evaluation. For FFD, we annotated 3 pages, which we split into 47 for training and 24 for evaluation.

4.3.4 Sense

For EEBD, we annotated 6 pages, which we split into 15 sense blocks for training and 15 for evaluation.

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<body>               100          100          100          100
<headnote>           100          100          100          100
<dictScrap>          100          100          100          100

===== Field-level results =====

label                accuracy     precision    recall       f1

<body>               100          100          100          100
<headnote>           100          100          100          100
<dictScrap>          100          100          100          100
```

Table 1: Evaluation of "Dictionary Segmentation" model on EEBD

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<body>               99.23        99.21        100          99.6
<headnote>           99.23        100          66.67        80

===== Field-level results =====

label                accuracy     precision    recall       f1

<body>               57.14        50           33.33        40
<headnote>           85.71        100          66.67        80
```

Table 2: Evaluation of the "Dictionary Segmentation" model on FFD

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<entry>              100          100          100          100
<pc>                 100          100          100          100

===== Field-level results =====

label                accuracy     precision    recall       f1

<entry>              100          100          100          100
<pc>                 100          100          100          100
```

Table 3: Evaluation of the "Dictionary Body Segmentation" model on EEBD

```
===== Token-level results =====

label               accuracy     precision    recall       f1

<entry>             99.6         100          99.6         99.8
<pc>                99.6         75           100          85.71

===== Field-level results =====

label               accuracy     precision    recall       f1

<entry>             75           61.02        80           69.23
<pc>                88.28        75           100          85.71
```

Table 4: Evaluation of the "Dictionary Body Segmentation" model on FFD

```
===== Token-level results =====

label               accuracy     precision    recall       f1

<form>              99.59        99.26        97.12        98.18
<pc>                99.59        100          83.33        90.91
<re>                89.62        0            0            0
<sense>             88.97        86.75        98.26        92.15

===== Field-level results =====

label               accuracy     precision    recall       f1

<form>              90.09        73.08        82.61        77.55
<pc>                95.5         100          82.76        90.57
<re>                90.99        0            0            0
<sense>             79.28        54.29        73.08        62.3
```

Table 5: Evaluation of the "Lexical Entry" model on EEBD

```
===== Token-level results =====

label               accuracy     precision    recall       f1

<form>              90.77        57.94        75.26        65.47
<pc>                97.6         28.57        11.76        16.67
<sense>             92.45        96.46        93.59        95

===== Field-level results =====

label               accuracy     precision    recall       f1

<form>              59.12        2.94         4.17         3.45
<pc>                85.4         28.57        11.76        16.67
<sense>             57.66        0            0            0
```

Table 6: Evaluation of the "Lexical Entry" model on FFD

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<gramGrp>            99.12        100          50           66.67
<sense>              99.12        99.1         100          99.55

===== Field-level results =====

label                accuracy     precision    recall       f1

<gramGrp>            88.89        100          50           66.67
<sense>              77.78        83.33        83.33        83.33
```

Table 7: Evaluation of the "Sense" model on EEBD

For FFD, we annotated 4 pages, which we split into 71 sense blocks for training and 19 for evaluation.

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<sense>              100          100          100          100

===== Field-level results =====

label                accuracy     precision    recall       f1

<sense>              28.57        44.44        44.44        44.44
```

Table 8: Evaluation of the "Sense" model on FFD

4.4 Discussion

The evaluation on both dictionaries shows a high performance by the first and second models to detect, respectively, the body part of a page and the boundaries of lexical entries. The header and punctuation predictions for the first two models are however low for the digitized sample. This could be explained by the quality of the text which sometimes led to the generation of feature values that bias the learning.

For the "Lexical Entry" model, the performance of the system remains high for the extraction of grammatical and morphological information on the English dictionary but with low precision on the Fang-French sample. The detection of related entries, which are contained only in the English dictionary, shows the limitation of our model to extract these constructs with the actual setup. We hypothesis that it is related, firstly, to a lack of annotated data and, secondly, to a lack of discriminative features. Nonetheless, the model performs relatively well for sense block detection on the English dictionary and slightly worse on the bilingual dictionary. The detection of the punctuation, representing the transition between the main fields of the model, is also limited in this model.

The results of the final model reflect the reliability of our features to structure the sense information, when it has to focus on the boundaries of senses. But for the case of the senses aggregated by POS, more discriminative features should be added.

5. Related Works

This work takes place within the context of studies on lexicography and digital humanities fields, targeting the exploitation of digitized dictionaries. Most previous research (Khemakhem et al., 2009; Fayed et al., 2014; Mykowiecka et al., 2012) remained limited to the costly manual elaboration of lexical patterns, based on observing the organisation of the lexical information in a specific sample.

There have been, however, strong pointers to the usefulness of machine learning techniques, CRF in particular, to address the issue of decoding the complexity of lexical resources. Crist presented experiments for processing and automatically tagging linear text of two bilingual dictionaries, using CRF models. The goal has been purely experimental, proving the appropriateness of CRF for tagging tokens in digitized dictionaries. His exhaustive study also stressed the other processing issues, which are very important to the effectiveness and the evaluation of any parsing technique. Another recent study (Bago and Ljubešić, 2015), has addressed the issue of using CRF models to perform automatic language and structure annotation in a multilingual dictionary. The technique again has a very high accuracy in much less time than would be required for manual annotation.

Both of the mentioned machine learning approaches apply one CRF model to label the all the tokens of a dictionary. In such a bottom-up technique, the learner is overwhelmed by the number of labels to choose from at once, which increases the number of prediction errors. A huge amount of training data is also required per model to cover middle and high complexity dictionaries.

The novelty in our approach is that we reduce the scope of each bottom-up model by splitting the task over different models that process the lexical information in a top down fashion. Moreover, our system does not stop at the level of tagging the tokens, but enables the construction of blocks of lexical information in a format that facilitates the processing as well as the exchange of the output.

6. Conclusion and Future Work

GROBID-Dictionaries in its first version has shown the promise of CRF cascading models to structure digitally born and digitized dictionaries, independently of the language and lexicographic style. Our experiments had the goal of, firstly, verifying our assumptions and, secondly, highlighting the strengths and the limitations of the implemented models. It is obvious that more focus should be given to the feature selection process, in order to reinforce the prediction of the models for certain labels and fields. Feature tuning should also be applied on larger annotated data with more varied instances. Therefore, we are planning to build a smart annotation tool with strong guidelines, to simplify the annotation process.

Our open source system could be used, after more tuning, to radically speed up the structuring of many digitized dictionaries in a unified scheme or to measure the structurability of OCRized lexical resources.

7. Acknowledgement

This work was supported by PARTHENOS. We would like to thank Patrice Lopez, the main designer of GROBID, for his continuous support and valuable advice.

8. References

Bago, P. and Ljubešić, N. (2015). Using machine learning for language and structure annotation in an 18th century dictionary. In *Electronic lexicography in the 21st century: linking lexical data in the digital age*.

Budin, G., Majewski, S., and Mörth, K. (2012). Creating lexical resources in tei p5. a schema for multi-purpose digital dictionaries. *Journal of the Text Encoding Initiative*, (3).

Crist, S. (2011). Processing the text of bilingual print dictionaries.

EEBD (2009). *Easier English Basic Dictionary: Pre-Intermediate Level. Over 11,000 terms clearly defined*. Easier English. Bloomsbury Publishing.

Fayed, D. M., Fahmy, A. A., Rashwan, M. A., and Kamel Fayed, W. (2014). Towards structuring an arabic-english machine-readable dictionary using parsing expression grammars.

Galley, S. (1964). *Dictionnaire fang-francais et francais-fang; suivi d'une grammaire fang par Samuel Galley. Avec une pref. de M.L. Durand-Reville*. H. Messeiller.

Khemakhem, A., Elleuch, I., Gargouri, B., and Hamadou, A. B. (2009). Towards an automatic conversion approach of editorial arabic dictionaries into lmf-iso 24613 standardized model.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513. Association for Computational Linguistics.

Lopez, P. and Romary, L. (2015). Grobid - information extraction from scientific publications. *ERCIM News*.

Mykowiecka, A., Rychlik, P., and Waszczuk, J. (2012). Building an electronic dictionary of old polish on the base of the paper resource. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage at LREC*, pages 16–21.