Arif Ali
Max Kim
Mohammad Ali
Nick Chapman
LING/COSC 572 ENLP
Nov 04, 2016

A2

Exercise

Compare your tags to someone elses in your group. Measure accuracy, construct a confusion matrix, and compute Cohens . Look at a few of the differences and explain them as best you can in English.
The link to the lpp.ch4.fr.group3 is:
https://docs.google.com/a/georgetown.edu/spreadsheets/d/1dp2bDpxuDs7Adr0rmFu0hftjAeJXQQAaa9WPH3grufA/edit?u

Our group attempted parts of speech tagging on The Little Prince in the original French translation. In order to contruct the confusion matrices for individual tags and calculate cohen's κ, R code was used.

code to build confusion matrix and cohen's κ

```
cm = read.csv("confusion_matrix_POS.csv")
POS_cm = function(p1, p2){
  cm_pi = cm[, c(p1,p2)]
  levels(cm_pi[,p1]) = c(levels(cm_pi[,p1]),
                         levels(cm_pi[,p2])[-which(levels(cm_pi[,p2]) %in% levels(cm_pi[,p1]))])
  levels(cm_pi[,p2]) = c(levels(cm_pi[,p2]),
                         levels(cm_pi[,p1])[-which(levels(cm_pi[,p1]) %in% levels(cm_pi[,p2]))])

  confusion_matrix = table(cm_pi)
  confusion_matrix = confusion_matrix[order(rownames(confusion_matrix)), order(colnames(confusion_matri
  print("Confusion Matrix")
  print(confusion_matrix)
  p_0 = sum(as.character(cm_pi[,p1])==as.character(cm_pi[,p2]))/nrow(cm_pi)
  p_e = sum(rowSums(confusion_matrix)/nrow(cm_pi)*(colSums(confusion_matrix))/nrow(cm_pi))
  #P(A=X) =0 therefore P(A=X)*P(B=X) = 0

  print("Agreement rate")
  print(p_0)
  print("Cohen's Kappa")
  return((p_0-p_e)/(1-p_e))
}
```

Mohammad vs Arif

```
POS_cm("Mohammad","Arif")

## [1] "Confusion Matrix"
##          Arif
## Mohammad    ADJ ADP ADV AUX CONJ DET NOUN NUM PART PRON PROPN PUNCT SCONJ
##           0   0   0   0   0    0   0    0   0    0    0     0     0     0
##    ADJ    0   5   0   0   0    0   0    0   0    0    0     0     0     0
##    ADP    0   0  13   0   0    0   1    0   0    0    0     0     0     0
##    ADV    1   0   0   0   0    0   0    0   0    0    0     0     0     0
##    AUX    1   0   0   0   2    0   0    0   0    0    0     0     0     0
##    CONJ   0   0   0   0   0    1   0    0   0    0    0     0     0     0
##    DET    0   0   0   0   0    0  14    0   0    0    2     0     0     0
##    NOUN   1   0   0   0   0    1   0   14   1    0    0     5     0     0
##    NUM    0   0   0   0   0    0   2    0   1    0    0     0     0     0
##    PART   0   0   0   0   0    0   0    0   0    2    0     0     0     0
##    PRON   0   0   0   0   0    0   1    0   0    0    5     0     0     0
##    PROPN  0   0   0   0   0    0   0    0   0    0    0     0     0     0
##    PUNCT  0   0   0   0   0    0   0    0   0    0    0     0    12     0
##    SCONJ  0   0   0   0   0    1   0    0   0    0    2     0     0     1
##    VERB   1   0   0   0   0    0   1    0   0    0    0     0     0     0
##    X      0   0   0   0   0    0   0    0   0    0    0     0     0     0
##         Arif
## Mohammad VERB  X
##             0  5
##    ADJ      0  0
##    ADP      0  0
##    ADV      0  0
##    AUX      1  0
##    CONJ     0  0
##    DET      0  0
##    NOUN     0  0
##    NUM      0  0
##    PART     0  0
##    PRON     0  0
##    PROPN    0  0
##    PUNCT    0  0
##    SCONJ    0  0
##    VERB     8  0
##    X        0  0
## [1] "Agreement rate"
## [1] 0.7428571
## [1] "Cohen's Kappa"
## [1] 0.7128532
```

There were five blanks from my POS tags compared to Mohammed's. This is because Arif was unable to determine the proper tags for these words in French. As expected both of us agreed on the PUNCT tags. The use of PRON seems to be the biggest difference between us. Of the 9 that Arif tagged as PRON, only five agreed with Mohammed. Also, Mohammed never tagged PROPNs while Arif tagged 5 words as proper-nouns. This could be due to differences in what we both consider to be names in French. Some difficulty in the name could be due to the multiple parts of speech that make up a name. Another interesting case, Arif marked all blank values with X while Mohammed left them blank. Based on Landis and Koch interpretation of κ, the matching between Mohammed and Arif is substantial. The main reason for the differences could be because Mohammad has a significiant amount of experience in the French language from living in France for a while whereas Arif took 2 semester of French nearly three years ago with no real indepth use.

Mohammad vs Max

```
POS_cm("Mohammad","Max")

## [1] "Confusion Matrix"
##         Max
## Mohammad    ADJ ADP ADV AUX CONJ DET NOUN NUM PART PRON PROPN PUNCT SCONJ
##           5   0   0   0   0    0   0    0   0    0    0     0     0     0
##    ADJ    0   4   0   0   0    0   0    0   0    0    0     1     0     0
##    ADP    0   0  14   0   0    0   0    0   0    0    0     0     0     0
##    ADV    0   0   0   0   0    0   0    0   0    1    0     0     0     0
##    AUX    0   0   0   0   4    0   0    0   0    0    0     0     0     0
##    CONJ   0   0   0   0   0    1   0    0   0    0    0     0     0     0
##    DET    0   0   0   0   0    0  11    0   0    0    5     0     0     0
##    NOUN   0   0   0   0   0    0   0   17   0    0    0     5     0     0
##    NUM    0   0   0   0   0    0   0    0   3    0    0     0     0     0
##    PART   0   0   0   0   0    0   0    0   0    2    0     0     0     0
##    PRON   0   0   0   0   0    0   0    0   0    0    6     0     0     0
##    PROPN  0   0   0   0   0    0   0    0   0    0    0     0     0     0
##    PUNCT  0   0   0   0   0    0   0    0   0    0    0     0    12     0
##    SCONJ  0   0   0   0   0    0   0    0   0    0    0     0     0     4
##    VERB   0   0   0   0   0    0   0    0   0    0    0     0     0     0
##         Max
## Mohammad VERB
##             0
##    ADJ      0
##    ADP      0
##    ADV      0
##    AUX      0
##    CONJ     0
##    DET      0
##    NOUN     0
##    NUM      0
##    PART     0
##    PRON     0
##    PROPN    0
##    PUNCT    0
##    SCONJ    0
##    VERB    10
## [1] "Agreement rate"
## [1] 0.8857143
## [1] "Cohen's Kappa"
## [1] 0.872418
```

Overall, there was a very high raw agreement rate between Max and Mohammad. The two main categories of disagreement came from tagging PROPN and tagging PRON. In the case of PROPN, Mohammad never tagged any of the words as PROPN, which likely stems from a disagreement in what is considered a name, especially if the name contains common words/nouns (such as Congres International d'Astronomie). This was the largest source of disagreement. In the other major source of disagreement, Max tagged several words as PRON, where Mohammad tagged those same words as DET. This disagreement is seems to stem in how groupings of le/la/l' + verb were tagged. Max treated the le/la/l' as pronouns of direct objects of their adjacent verbs, while Mohammad treated these as determiners. The last point of disagreement stemmed from the word alors where Max tagged it as PART, but noted that he was unsure. Looking at the data with Max and Nick, it is likely that Max was mistaken on this tagging, given his uncertainty. The Landis and Koch interpretations of Kappa suggest that the matching between Mohammad and Max is substantial.

Mohammad vs Nick

```
POS_cm("Mohammad","Nick")

## [1] "Confusion Matrix"
##         Nick
## Mohammad    ADJ ADP ADV AUX CONJ DET NOUN NUM PART PRON PROPN PUNCT SCONJ
##           5   0   0   0   0    0   0    0   0    0    0     0     0     0
##    ADJ    0   4   0   0   0    0   0    0   0    0    0     1     0     0
##    ADP    0   0  12   0   0    0   1    0   0    1    0     0     0     0
##    ADV    0   0   0   1   0    0   0    0   0    0    0     0     0     0
##    AUX    0   0   0   0   4    0   0    0   0    0    0     0     0     0
##    CONJ   0   0   0   0   0    1   0    0   0    0    0     0     0     0
##    DET    0   2   0   0   0    0  13    0   0    0    1     0     0     0
##    NOUN   0   0   0   0   0    0   0   17   1    0    0     4     0     0
##    NUM    0   0   0   0   0    0   1    0   2    0    0     0     0     0
##    PART   0   0   0   0   0    0   0    0   0    2    0     0     0     0
##    PRON   0   0   0   0   0    0   0    0   0    0    6     0     0     0
##    PROPN  0   0   0   0   0    0   0    0   0    0    0     0     0     0
##    PUNCT  0   0   0   0   0    0   0    0   0    0    0     0    12     0
##    SCONJ  0   0   0   0   0    1   0    0   0    0    0     0     0     3
##    VERB   0   0   0   0   0    0   0    0   0    0    0     0     0     0
##         Nick
## Mohammad VERB
##             0
##    ADJ      0
##    ADP      0
##    ADV      0
##    AUX      0
##    CONJ     0
##    DET      0
##    NOUN     0
##    NUM      0
##    PART     0
##    PRON     0
##    PROPN    0
##    PUNCT    0
##    SCONJ    0
##    VERB    10
## [1] "Agreement rate"
## [1] 0.8761905
## [1] "Cohen's Kappa"
## [1] 0.8615057
```

Mohammad and Nick have an excellent agreement rate and an excellent Cohens Kappa as well. Their disagreements come on the easily controversial tags. For example, on PROPN they have no agreement; however this appears to be due to how Nick tagged the organization names in the text, opting to have every word in the organization name be a proper noun. Mohammad took the alternative approach which is to tag every word as its proper part of speech independent of whether it is part of a name. Their other disagreements were on similar types of issues where the confusion is easily explained. They had minor disagreement on ADP and DET, though in French this is a common confusion because of terms like de which can in some contexts mean from and in others be used to express that one has some of something. Additionally there was one disagreement where Nick said that something was a CONJ but Mohammad put SCONJ; both recognized the conjunction yet disagreed on whether it was subordinate. Overall, their agreement is excellent by any of the proposed metrics for assessing a Cohen Kappa.

Nick vs Arif

```
POS_cm("Nick", "Arif")
```

```
## [1] "Confusion Matrix"
##        Arif
## Nick       ADJ ADP ADV AUX CONJ DET NOUN NUM PART PRON PROPN PUNCT SCONJ
##          0   0   0   0   0    0   0    0   0    0    0     0     0     0
##   ADJ    0   4   0   0   0    0   0    0   0    0    2     0     0     0
##   ADP    0   0  11   0   0    0   1    0   0    0    0     0     0     0
##   ADV    1   0   0   0   0    0   0    0   0    0    0     0     0     0
##   AUX    1   0   0   0   2    0   0    0   0    0    0     0     0     0
##   CONJ   0   0   0   0   0    1   0    0   0    0    1     0     0     0
##   DET    0   0   1   0   0    0  14    0   0    0    0     0     0     0
##   NOUN   1   0   0   0   0    1   0   13   1    0    0     1     0     0
##   NUM    0   0   0   0   0    0   1    0   1    0    0     1     0     0
##   PART   0   0   1   0   0    0   0    0   0    2    0     0     0     0
##   PRON   0   0   0   0   0    0   2    0   0    0    5     0     0     0
##   PROPN  0   1   0   0   0    0   0    1   0    0    0     3     0     0
##   PUNCT  0   0   0   0   0    0   0    0   0    0    0     0    12     0
##   SCONJ  0   0   0   0   0    1   0    0   0    0    1     0     0     1
##   VERB   1   0   0   0   0    0   1    0   0    0    0     0     0     0
##   X      0   0   0   0   0    0   0    0   0    0    0     0     0     0
##        Arif
## Nick    VERB  X
##            0  5
##   ADJ      0  0
##   ADP      0  0
##   ADV      0  0
##   AUX      1  0
##   CONJ     0  0
##   DET      0  0
##   NOUN     0  0
##   NUM      0  0
##   PART     0  0
##   PRON     0  0
##   PROPN    0  0
##   PUNCT    0  0
##   SCONJ    0  0
##   VERB     8  0
##   X        0  0
## [1] "Agreement rate"
## [1] 0.7333333
## [1] "Cohen's Kappa"
## [1] 0.7043741
```

Nick and Arif have fairly good agreement rate. The majority of their disagreement comes from the fact that Arif omitted tags where he was completely unable to decipher the part of speech and he put in X for blank words where Nick simply left these blank. On the majority of tags they do have good agreement with confusion coming on the objectively confusing words. For example, the word le/la/l preceding verbs was a point of contention with Arif labeling it a DET and Nick labeling it a PRON. This makes sense since in those situations le/la/l takes the place of the object, but normally le/la/l is a determiner. Another agreement issue, which is common to many of the annotation differences, is what words are PROPNs. With organization names and names containing numbers there is much room for disagreement about which tag is correct. Other disagreements were minor such as AUX vs VERB and CONJ vs SCONJ. Some amount of error can certainly be attributed to the fact that neither annotator is a native speaker. Overall, by Landis and Kochs guidelines the pair had substantial agreement and by Fleiss standards the pair had almost excellent agreement.

Nick vs Max

```
POS_cm("Nick", "Max")

## [1] "Confusion Matrix"
##        Max
## Nick       ADJ ADP ADV AUX CONJ DET NOUN NUM PART PRON PROPN PUNCT SCONJ
##          5   0   0   0   0    0   0    0   0    0    0     0     0     0
##   ADJ    0   4   0   0   0    0   0    0   0    0    2     0     0     0
##   ADP    0   0  12   0   0    0   0    0   0    0    0     0     0     0
##   ADV    0   0   0   0   0    0   0    0   0    1    0     0     0     0
##   AUX    0   0   0   0   4    0   0    0   0    0    0     0     0     0
##   CONJ   0   0   0   0   0    1   0    0   0    0    0     0     0     1
##   DET    0   0   1   0   0    0  11    0   1    0    2     0     0     0
##   NOUN   0   0   0   0   0    0   0   17   0    0    0     0     0     0
##   NUM    0   0   0   0   0    0   0    0   2    0    0     1     0     0
##   PART   0   0   1   0   0    0   0    0   0    2    0     0     0     0
##   PRON   0   0   0   0   0    0   0    0   0    0    7     0     0     0
##   PROPN  0   0   0   0   0    0   0    0   0    0    0     5     0     0
##   PUNCT  0   0   0   0   0    0   0    0   0    0    0     0    12     0
##   SCONJ  0   0   0   0   0    0   0    0   0    0    0     0     0     3
##   VERB   0   0   0   0   0    0   0    0   0    0    0     0     0     0
##        Max
## Nick    VERB
##            0
##   ADJ      0
##   ADP      0
##   ADV      0
##   AUX      0
##   CONJ     0
##   DET      0
##   NOUN     0
##   NUM      0
##   PART     0
##   PRON     0
##   PROPN    0
##   PUNCT    0
##   SCONJ    0
##   VERB    10
## [1] "Agreement rate"
## [1] 0.9047619
## [1] "Cohen's Kappa"
## [1] 0.8945254
```

Between Nick and Max there were six categories for disagreement, however each category only had minor disagreements. The biggest source of disagreement was in the tagging of PRON, where on four occasions, Max tagged the word as PRON and Nick did not. Nick tagged possessive pronouns such as sa or son as ADJ, while Max tagged them as PRON. Similar to the situation with Mohammad and Max, there was some disagreement in groupings of le/la/l' + verb, where Max always tagged the le/la/l' as PRON, while Nick sometimes tagged this as DET (and in other contexts as PRON). The two also had a disagreement regarding SCONJ and CONJ in one instance, but both still saw the word as some kind of conjunction (differing only in which type of conjunction). There was an instance of disagreement for names, in which Nick considered an ADP within a name to actually be a PART, while Max considered it an ADP. Nick noted that this decision was influenced by the fact that the ADP within the name should not be parsed without the surrounding PROPN was thus a PART. There was also a disagreement in tagging alors, where Max tagged it as PART and Nick tagged it as ADV. Max noted that he was unsure of this tag, and given Mohammad's tagging of alors as ADV, it is likely that Max was incorrect in this tagging. Despite these many minor disagreements, Max and Nick had the highest agreement rate, at 90.476%, with a similarly high Cohen's Kappa suggesting that the matching between Nick and Max is substantial. This could possibly be attributed to both participants having similar backgrounds in French and English.

Arif vs Max

```
POS_cm("Arif", "Max")

## [1] "Confusion Matrix"
##         Max
## Arif       ADJ ADP AUX CONJ DET NOUN NUM PART PRON PROPN PUNCT SCONJ VERB
##          0   0   0   1    0   0    1   0    1    0     0     0     0    1
##   ADJ    0   4   0   0    0   0    0   0    0    0     1     0     0    0
##   ADP    0   0  13   0    0   0    0   0    0    0     0     0     0    0
##   AUX    0   0   0   2    0   0    0   0    0    0     0     0     0    0
##   CONJ   0   0   0   0    1   0    1   0    0    0     0     0     1    0
##   DET    0   0   1   0    0  11    0   2    0    4     0     0     0    1
##   NOUN   0   0   0   0    0   0   13   0    0    0     1     0     0    0
##   NUM    0   0   0   0    0   0    1   1    0    0     0     0     0    0
##   PART   0   0   0   0    0   0    0   0    2    0     0     0     0    0
##   PRON   0   0   0   0    0   0    0   0    0    7     0     0     2    0
##   PROPN  0   0   0   0    0   0    1   0    0    0     4     0     0    0
##   PUNCT  0   0   0   0    0   0    0   0    0    0     0    12     0    0
##   SCONJ  0   0   0   0    0   0    0   0    0    0     0     0     1    0
##   VERB   0   0   0   1    0   0    0   0    0    0     0     0     0    8
##   X      5   0   0   0    0   0    0   0    0    0     0     0     0    0
##         Max
## Arif     X
##          0
##   ADJ    0
##   ADP    0
##   AUX    0
##   CONJ   0
##   DET    0
##   NOUN   0
##   NUM    0
##   PART   0
##   PRON   0
##   PROPN  0
##   PUNCT  0
##   SCONJ  0
##   VERB   0
##    X       0
## [1] "Agreement rate"
## [1] 0.752381
## [1] "Cohen's Kappa"
## [1] 0.7260686
```

Like the case with Mohammad vs Arif, Arif used X values for blanks words whereas Max Omitted them. The two categories with majority disagreements were DET and PRON. While a majority of DET tags were agreed upon, Max disagreed and labels some of these as PRON. This could be do to cases where le/la/l was used in place of an object, but the POS tag guide generally states that this is a DET. One of the starkest difference is the agreement of tags that are SCONJ. Max and Arif disagree in most cases. This could be due to the difference in French between CONJ and SCONJ. Based on Landis and Koch interpretation of κ, the matching between Arif and Max is substantial. The main reason for the differences could be difference in lnaguage exposure between Arif and Max. Based off of discussions during the tagging, Max has had years of study in the French language in addition to a fluent Parent.

The Java source code used to calculate Cohen's K and raw agreement:

```
import java.io.File;
import java.io.FileInputStream;
import java.io.UnsupportedEncodingException;
import java.util.ArrayList;
import java.util.Arrays;

/**
 * Created by yektaie on 11/3/16.
 */
public class Program {
    private static String[] TAGS = {"", ""};
    private static final String FILE_PATH = "/Volumes/Files/Georgetown/Natural Language
Processing/A2/files/%s.txt";
    private static final String[] annotators = {"Ali", "Arif", "Nick", "Max",
"Ali-Corrected"};

    public static void main(String[] args) {
        loadTags();

        for (int i = 0; i < annotators.length; i++) {
            for (int j = 0; j < i; j++) {
                String a1 = annotators[i];
                String a2 = annotators[j];

                ConfusionResult confusion = computeCohenAndRaw(a1, a2);
                System.out.println(String.format("%s - %s -> κ = %.5f, r = %.5f", a1, a2,
confusion.cohen, confusion.raw));
            }
        }
    }

    private static ConfusionResult computeCohenAndRaw(String a1, String a2) {
        String[] ts_1 = readTextFile(String.format(FILE_PATH, a1)).replace("\r",
"").split("\n");
        String[] ts_2 = readTextFile(String.format(FILE_PATH, a2)).replace("\r",
"").split("\n");

        int[][] matrix = new int[TAGS.length + 1][];
        for (int i = 0; i < matrix.length; i++) {
            matrix[i] = new int[TAGS.length + 1];
        }

        for (int i = 0; i < ts_1.length; i++) {
            int a1i = getIndex(TAGS, ts_1[i]);
            int a2i = getIndex(TAGS, ts_2[i]);

            matrix[a1i][a2i]++;
            matrix[a1i][TAGS.length]++;
            matrix[TAGS.length][a2i]++;
        }

        double AO = 0;
        for (int i = 0; i < matrix.length; i++) {
            AO += matrix[i][i];
        }

        AO /= ts_1.length;
        double AC = 0;

        for (int i = 0; i < TAGS.length; i++) {
            AC += (matrix[i][TAGS.length] * matrix[TAGS.length][i]);
        }

        AC /= (ts_1.length * ts_1.length);

        ConfusionResult result = new ConfusionResult();
        result.cohen = (AO - AC) / (1 - AC);
        result.raw = AO;

        return result;
    }

    private static int getIndex(String[] tags, String s) {
        int result = -1;

        for (int i = 0; i < tags.length && result == -1; i++) {
            if (tags[i].equals(s)) {
                result = i;
            }
        }

        return result;
    }

    private static ArrayList<String> loadTags() {
        ArrayList<String> tags = new ArrayList<>();
        addTags(tags, "Ali");
        addTags(tags, "Arif");
        addTags(tags, "Max");
        addTags(tags, "Nick");

        TAGS = toArray(tags);
        return tags;
    }

    private static String[] toArray(ArrayList<String> tags) {
        String[] result = new String[tags.size()];

        for (int i = 0; i < result.length; i++) {
            result[i] = tags.get(i);
        }

        Arrays.sort(result);

        return result;
    }

    private static void addTags(ArrayList<String> tags, String fileName) {
        String[] ts = readTextFile(String.format(FILE_PATH, fileName)).split("\n");
        for (String tag : ts) {
            if (!tags.contains(tag.trim())) {
                tags.add(tag.trim());
            }
        }
    }

    private static String readTextFile(String path) {
        byte[] result = null;
        try {
            FileInputStream fs = new FileInputStream(new File(path));
            int length = fs.available();
            byte[] content = new byte[length];
            fs.read(content);

            result = content;

            fs.close();
        } catch (Exception e) {
            // e.printStackTrace();
        }

        try {
            return new String(result, "UTF-8");
        } catch (UnsupportedEncodingException e) {
            e.printStackTrace();
        }

        return null;
    }

}

class ConfusionResult {
    public double cohen;
    public double raw;
}
```

The result of this program is as follow:

```
Arif - Ali -> κ = 0.71285, r = 0.74286
Nick - Ali -> κ = 0.86151, r = 0.87619
Nick - Arif -> κ = 0.70437, r = 0.73333
Max - Ali -> κ = 0.87242, r = 0.88571
Max - Arif -> κ = 0.72607, r = 0.75238
Max - Nick -> κ = 0.89453, r = 0.90476
Ali-Corrected - Ali -> κ = 0.94655, r = 0.95238
Ali-Corrected - Arif -> κ = 0.72474, r = 0.75238
Ali-Corrected - Nick -> κ = 0.86234, r = 0.87619
Ali-Corrected - Max -> κ = 0.88370, r = 0.89524
```

In the results, κ is Cohen’s Kappa and r is the Agreement Rate.
Note that there are two Alis. One is the first one Ali have annotated. Ali did not pay attention that there is proper noun tag. So Ali tagged all proper nouns as “NOUN”. After Arif pointed that out, Ali have corrected the error, and considered it as another annotator.

More Details about the differences:

The lines which at least one of the annotators of this group had annotated differently is shown in table 1.

Lines 6 and 62 [découvre l'une d'elles]: In French, “une” have multiple senses. It can mean one (feminine) or a (feminine). Among the annotators, only Arif believed “une” means “a” not “one”.

Line 8 [découvre l'une d'elles]: Among the annotators, only Arif believed “elles” have a DET part of speech.

Line 11 [il lui donne pour]: Like “une”, “lui” have different senses. One of them is a possessive determiner while the other is the pronoun “il” used as object. In this case, Max believed “lui” is a pronoun.

Line 16 [il lui donne pour nom un zéro]: “zéro” is the number zero. It’s usage in this sentence does not refer to the number zero, but zero as a name for a planet, based on the previous context.

Lines 31, 54, 75 and 93: Among the annotators, only Arif tagged the white spaces as an X part of speech. Ali have not tagged it anything, because an empty token does not have any meaning.

Line 20 [Il l' appelle par example]: Max believe “l’” is a pronoun here, while other annotators considered it as a determiner.

Line 27 [l' astéroide 3251]: Corrected after the first time. Ali believe it is part of name, so it should be tagged as proper noun.

Line 28 [l' astéroide 3251]: See explanation for lines 6 and 27.

Line 34 [J' ai de sérieuses raisons]: Only Nick believed “de” to be a determiner, while other annotators tagged it as ADP. Even after evaluating my result with others, ali still think it to be ADP not determiner. Since in French, if a noun is plural, the determiner should also be plural. “De” is single, but “sérieuses raisons” is plural.

Line 39 [de croire que la planète]: “que” can be both SCONJ and PRON. In fact, the documentation uses examples to show both of its usage. Ali think annotators tagged it based on the meaning they understand from the sentence.

Line 43 [la planète d' ou venait le ...]: Like line 6, “ou” have different senses. It can mean “where” or it can mean “or”. Based on the sense the annotator understand from it, it can be either CONJ or SCONJ.

Line 47 [la planète d' ou venait le petit prince]: See line 27 explanation. Ali think Arif is right on this word and the other annotator including myself are wrong.

Line 50, 51 and 52 [l' astéroide B 612]: See explanation for lines 16, 27 and 28.

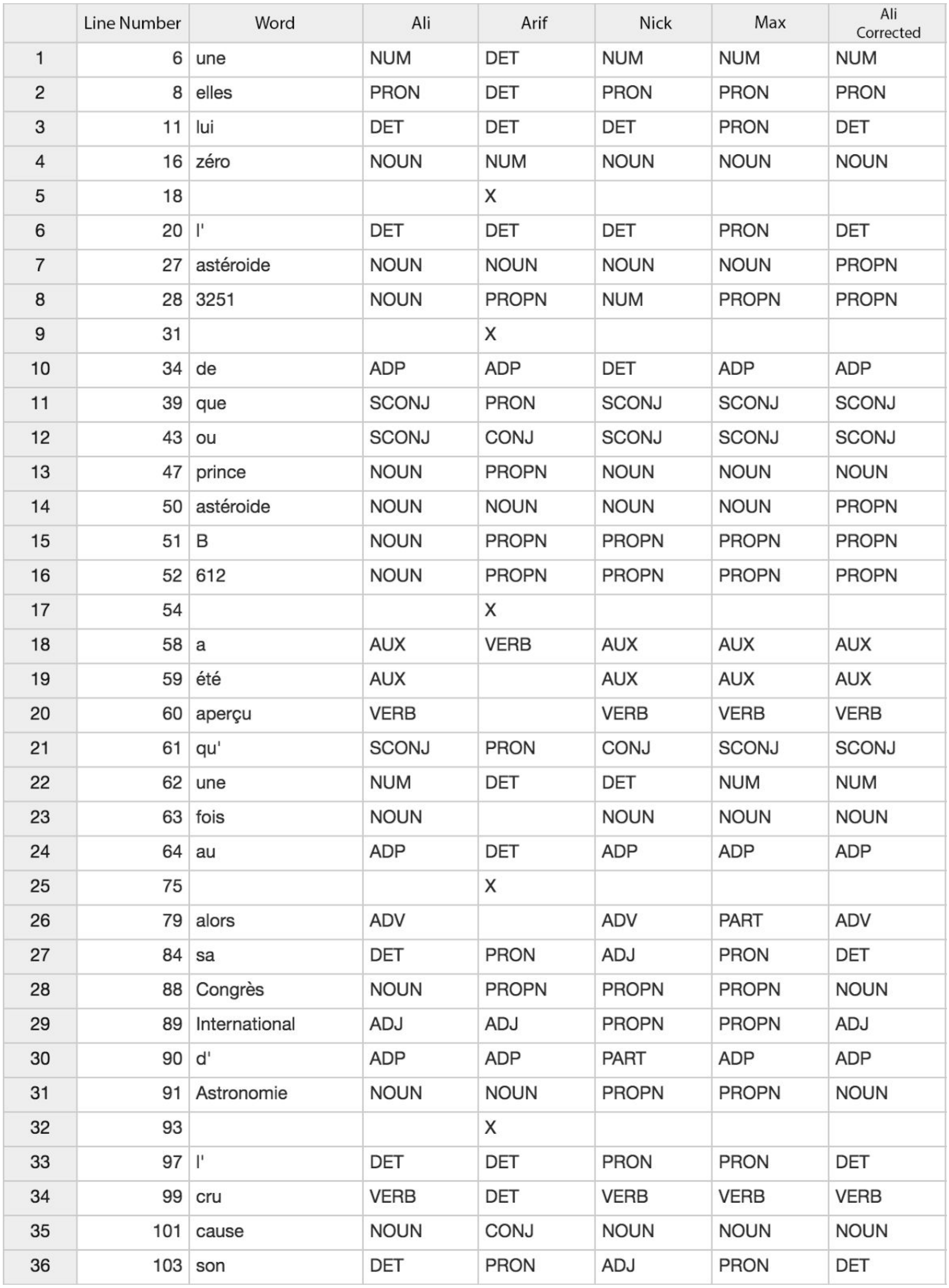
Line 58, 59 and 60 [a été aperçu qu' une fois]: All the annotators agrees that “a été aperçu” is made of 3 verb token. The difference is that is it a VERB or AUX.

Line 61: See explanation for line 39.

Line 64 [une fois au télescope]: “voire au télescope” means watching from behind the lenses of a telescope. That is why all annotator tagged it as ADP. Arif believed “au” is a determiner for “télescope”.

Line 79 [Il avait fait alors une grande]: In this case, almost every annotator disagree with the others. Ali and Nick tagged it as ADV while Arif left it blank. Max tagged it as PART.

Line 84 [démonstration de sa découverte]: Just like line 11, “sa” have different senses. Each annotator had a different opinion on its meaning. The tags includes DET, PRON and ADJ.



	Line Number	Word	Ali	Arif	Nick	Max	Ali Corrected
1	6	une	NUM	DET	NUM	NUM	NUM
2	8	elles	PRON	DET	PRON	PRON	PRON
3	11	lui	DET	DET	DET	PRON	DET
4	16	zéro	NOUN	NUM	NOUN	NOUN	NOUN
5	18			X			
6	20	l'	DET	DET	DET	PRON	DET
7	27	astéroide	NOUN	NOUN	NOUN	NOUN	PROPN
8	28	3251	NOUN	PROPN	NUM	PROPN	PROPN
9	31			X			
10	34	de	ADP	ADP	DET	ADP	ADP
11	39	que	SCONJ	PRON	SCONJ	SCONJ	SCONJ
12	43	ou	SCONJ	CONJ	SCONJ	SCONJ	SCONJ
13	47	prince	NOUN	PROPN	NOUN	NOUN	NOUN
14	50	astéroide	NOUN	NOUN	NOUN	NOUN	PROPN
15	51	B	NOUN	PROPN	PROPN	PROPN	PROPN
16	52	612	NOUN	PROPN	PROPN	PROPN	PROPN
17	54			X			
18	58	a	AUX	VERB	AUX	AUX	AUX
19	59	été	AUX		AUX	AUX	AUX
20	60	aperçu	VERB		VERB	VERB	VERB
21	61	qu'	SCONJ	PRON	CONJ	SCONJ	SCONJ
22	62	une	NUM	DET	DET	NUM	NUM
23	63	fois	NOUN		NOUN	NOUN	NOUN
24	64	au	ADP	DET	ADP	ADP	ADP
25	75			X			
26	79	alors	ADV		ADV	PART	ADV
27	84	sa	DET	PRON	ADJ	PRON	DET
28	88	Congrès	NOUN	PROPN	PROPN	PROPN	NOUN
29	89	International	ADJ	ADJ	PROPN	PROPN	ADJ
30	90	d'	ADP	ADP	PART	ADP	ADP
31	91	Astronomie	NOUN	NOUN	PROPN	PROPN	NOUN
32	93			X			
33	97	l'	DET	DET	PRON	PRON	DET
34	99	cru	VERB	DET	VERB	VERB	VERB
35	101	cause	NOUN	CONJ	NOUN	NOUN	NOUN
36	103	son	DET	PRON	ADJ	PRON	DET

Table 1: The words annotated differently by the group of annotators