Ressource multi-niveaux annotée pour la pathologie mentale

Projet Tutoré de Master Sciences de la Cognition et Applications

Kenny RIVALIN et Nathalie WITTMANN
Année 2013/2014

Encadrant : Maxime AMBLARD,

Co-encadrants : Karën FORT, Michel MUSIOL, Manuel REBUSCHI

Ressource multi-niveaux annotée pour la pathologie mentale

Projet Tutoré : Master Sciences de la Cognition et Applications

Kenny RIVALIN et Nathalie WITTMANN

Encadrant : Maxime AMBLARD

Co-encadrants : Karen FORT, Michel MUSIOL, Manuel REBUSCHI

2013-2014

Remerciements

Nous souhaitons remercier Maxime Amblard, notre encadrant principal, qui nous a guidé et conseillé tout au long du projet.

Nous voulons également remercier Karen Fort, Michel Musiol et Manuel Rebuschi pour leurs conseils.

Nous remercions également les quelques personnes qui ont bien voulu prendre de leur temps pour effectuer le travail d'annotation.

Table des matières

I Présentation du sujet

Le projet de recherche SLAM vise à étudier des conversations de patients schizophrène et, grâce à l'outil informatique, d'automatiser certaines étapes de l'analyse de leur discours.

La schizophrénie est une pathologie qui reste par certains points encore mal connue et mal appréhendée, notamment car les symptômes qui la caractérisent ne sont pas propres à cette pathologie : il est difficile de trouver des symptômes ou des caractéristiques de maladies communes à toutes les personnes atteintes.

De plus, les actuels instruments de mesure tel que la psychométrie sont assez mal adaptés à la mesure des troubles de la pensée et du langage chez les schizophrènes car ils ne tiennent pas suffisamment compte du contexte et de la dynamique qui entoure la communication (Musiol, Verhaegen, 2008) .

Il est cependant admis par la communauté scientifique que l'un des éléments communs aux patients schizophrènes est une anomalie des processus de la pensée ce qui entraine également des troubles du langage (Amblard, Musiol, Rebuschi, 2011).

C'est pourquoi le projet de recherche SLAM vise à étudier des conversations de patients schizophrènes et cela notamment :

- En se basant sur des théories du discours : la S-DRT (qui est une extension de la DRT : Discourse Representation Theory. La S-DRT permet de retranscrire les processus présent dans le dialogue en formalisant le langage et en s'appuyant également sur l'ensemble de la conversation).

- En mettant en évidence l'existence de tics de langages non habituels chez la population générale qui sont récurrent chez les patients schizophrènes.

Pour ce faire, il existe un corpus très important recueillant des dialogues entre personnes schizophrènes ou personnes témoins et psychologues. Ces données ont été enregistrées à Lyon et sont disponibles dans un fichier au format texte. Cependant les interventions ont été replacées de manière aléatoire pour des raisons de confidentialité.

Ce projet de recherche se base donc sur des notions pluridisciplinaires concernant l'informatique, la linguistique, la psychologie et la philosophie.

Ce projet tutoré s'inscrit dans le cadre du projet SLAM car son but est de participer à la mise en place de deux campagnes d'annotations concernant ces discours.

Les études d'analyse du langage se font sur différents niveaux :

- Phonologique : étudie le son
- Morphologique : étudie le mot à un niveau grammatical
- Syntaxique : étudie les mots au sein d'une énoncé
- Sémantique : étudie le sens des énoncés
- Pragmatique : étudie le sens en contexte

Dans un premier temps, il s'agit de mener une campagne d'annotation en morphosyntaxe afin d'analyser le texte ciblé à un niveau grammatical. On pourra ensuite extraire des informations statistiques sur les résultats obtenus, comme par exemple le pourcentage d'hésitations à l'oral d'un patient schizophrène par rapport à celui d'un patient témoin. La différence entre ces deux valeurs pourrait être significative, ce qui signifiera la mise en évidence d'une caractéristique propre à la schizophrénie. Le niveau d'analyse du langage de cette campagne est donc morphologique et syntaxique.

Dans un second temps, nous ferons une campagne d'annotation permettant d'identifier des discontinuités décisives. Les discontinuités décisives sont des ruptures dans le dialogue et sont des caractéristiques de la schizophrénie. La méthodologie consiste à faire annoter des extraits de corpus de transcriptions de dialogues entre schizophrène et psychologue par des annotateurs volontaires à l'aide d'un guide d'annotation rédigé au préalable par nos soins. Le but recherché est de savoir si ces annotations mettront en évidence ces troubles du langage, caractéristiques de la schizophrénie. Le niveau d'analyse du langage de cette campagne est de niveau pragmatique.

II Travail réalisé

Cette partie va présenter les différentes étapes de notre travail.
Etant donné qu'il nous a été confié deux travaux distincts, deux parties étaient nécessaires pour les traiter. Tout d'abord une partie concernant l'annotation en morpho syntaxe, c'est-à-dire une annotation automatique via un logiciel de traitement automatique de langue (MElt) puis une deuxième partie concernant l'annotation par des annotateurs « humains » sur les discontinuités décisives.

Partie 1 : L'annotation en morpho syntaxe

Le corpus sur lequel a été effectué l'analyse en morphosyntaxe rassemble en réalité 41 corpus en tout et pour tout. Parmi eux :

- 18 sont ceux de patients schizophrènes dont :
  - 15 sont des hommes
  - 3 sont des femmes
- 23 sont ceux de témoins dont :
  - 15 sont des hommes
  - 8 sont des femmes

Le nombre d'années d'études de différence entre les témoins et les patients n'est pas très élevé. De même que le QI des témoins est légèrement supérieur à celui des patients, mais les deux groupes sont relativement bien appareillés.
Ces éléments permettent de limiter les biais. Cependant il existe toujours éventuellement celui du genre : il y a plus d'hommes que de femmes. Il est néanmoins difficile de contourner le problème, étant donné qu'il est très difficile d'accéder aux transcriptions de discours entre patient et psychologue.

A - Utilisation du logiciel MElt

MElt est un système d'étiquetage séquentiel de textes très complet, c'est-à-dire qu'il permet de « tagger » un corpus : il suffit d'entrer une commande unix comprenant le nom du fichier que l'on souhaite tagger ainsi que MElt pour que le programme se lance et tag automatiquement.

MElt est utilisable pour de nombreuses langues et possède plusieurs outils comme un Lemmatiseur (option MElt –L) ou encore un ensemble de scripts (découpage de texte, tokenisation, reconnaissance automatique de certaines entités type URL, adresse mail, etc… entre autre).

Un tagging est un étiquetage grammatical (POS : Part of Speech Tagging). C'est un processus permettant d'associer à chaque mot d'un corpus une information grammaticale en fonction du contexte dans lequel ce mot est placé.

Exemple de l'exécution du logiciel MElt via une commande unix :

```
echo "Argan, qui est le malade imaginaire, s'apprête à marier sa fille." | MElt


  TAGGER: Loading tag dictionary...
  TAGGER: Loading tag dictionary: done
  TAGGER: Loading external lexicon...
  TAGGER: Loading external lexicon: done
  TAGGER: Loading model from /usr/local/share/melt/fr...
  TAGGER: Loading model from /usr/local/share/melt/fr: done
  TAGGER: POS Tagging...
  TAGGER: POS Tagging: done (in 0.0333611965179 sec).
Argan,/NPP qui/PROREL est/V le/DET malade/ADJ imaginaire,/ET s'apprête/ET à/P
marier/VINF sa/DET fille./NC
```

On constate qu'après chaque mot est apparu un « / » suivit d'une suite de quelques lettres en majuscule. Il s'agit d'un tag qui permet de définir la catégorie grammaticale des mots dans

leur contexte. Ici, /NC signifiera qu'il s'agit d'un nom commun, /ADJ d'un adjectif, /V d'un verbe, etc.

Voici la liste exhaustive des TAGS produits par MElt :

ADJ	adjective
ADJWH	interrogative adjective
ADV	adverb
ADVWH	interrogative adverb
CC	coordination conjunction
CLO	object clitic pronoun
CLR	reflexive clitic pronoun
CLS	subject clitic pronoun
CS	subordination conjunction
DET	determiner
DETWH	interrogative determiner
ET	foreign word
I	interjection
NC	common noun
NPP	proper noun
P	preposition
P+D	preposition+determiner amalgam
P+PRO	prepositon+pronoun amalgam
PONCT	punctuation mark
PREF	prefix
PRO	full pronoun
PROREL	relative pronoun
PROWH	interrogative pronoun
V	indicative or conditional verb form
VIMP	imperative verb form
VINF	infinitive verb form
VPP	past participle
VPR	present participle
VS	subjunctive verb form

La commande utilisée dans notre projet est *cat « nomdufichier.txt » | MElt > fichierdesortie.txt*

« > » Indique une redirection dans un fichier : le résultat de la commande entrée sera placée dans le fichier suivant cette instruction

« | » Un pipe permet d'enchaîner l'exécution de plusieurs commandes, ici « cat » et « MElt »

« Cat » affiche le contenu du fichier suivant cette commande

Le résultat de la commande est donc une version taggée du fichier mis en entrée. Ce résultat se trouvera dans « fichierdesortie.txt »

Dans le cadre de ce projet, MElt nous permet donc de tagger entièrement un corpus en POS.

Le corpus que nous utilisons est un assemblement d'interventions issues de dialogues entre patients schizophrènes et psychologues de Lyon. Pour des soucis de confidentialité, les interventions ont été mélangées aléatoirement. Certains des « Patients » sont en réalité des patients témoins qui permettent d'avoir un élément de comparaison avec les données des patients schizophrènes, en plus des interventions des psychologues. Il sera ainsi possible d'extraire des statistiques sur le corpus en confrontant les résultats des différents profils.

Pour cela, nous avons créé un programme en Python appelé Analyse_Corpus.py. Ce programme utilise une bibliothèque d'expressions régulières pour certains traitements linguistiques du corpus ainsi qu'une de gestion de codecs : les fichiers d'entrée et de sortie sont tous codés en UTF8 (encodage universel).

Il y a 7281 interventions (lignes) de témoin, 3788 de patient et 11454 de psychologue. C'est une différence non négligeable pouvant fausser les statistiques et entrainer des conclusions erronées si ce paramètre n'est pas pris en compte.

Lorsque que l'on récupère l'intégralité des mots d'un corpus, les plus utilisés ne sont souvent pas pertinents car dénués de sens. Ce sont par exemples des mots comme « quel, que, qui, moins, … ». Ces mots sont appelés « stopwords » et il convient de ne pas les prendre en compte pour effectuer une analyse pertinente. Pour ce faire, nous avons créé une liste stopwords et le programme vérifie si les mots trouvés dans le texte appartiennent à cette liste. Si c'est le cas, ils sont simplement écartés des analyses suivantes.

La fonction principale du programme prend un fichier.txt en paramètre. Dans un premier temps, cette fonction extrait toutes les interventions des patients schizophrènes et les place dans un fichier appelé « Corpus_patient.txt ». Idem pour les interventions des témoins, placées dans le fichier « Corpus_temoin.txt » et celles du psychologue dans le fichier « Corpus_psy.txt

Ensuite, le programme va tagger les fichiers précédemment créés grâce à MElt en utilisant respectivement les commandes suivantes :

cat Corpus_patient.txt | MElt > Corpus_patient_tag.txt
cat Corpus_temoin.txt | MElt > Corpus_temoin_tag.txt
cat Corpus_psy.txt | MElt > Corpus_psy_tag.txt

Le programme effectue aussi une analyse de lexème grâce à l'option MElt – L et créé à nouveau deux fichiers : l'un comporte le texte avec les lexèmes des interventions des patients et l'autre comporte la version témoin.

cat Corpus_patient.txt | MElt -L> Corpus_patient_lexeme.txt
cat Corpus_temoin.txt | MElt -L> Corpus_temoin_lexeme.txt
cat Corpus_psy.txt | MElt -L> Corpus_psy_lexeme.txt

Il définit ensuite plusieurs dictionnaires, pour chaque intervenant :

- pour les tags des fichiers patients et témoins
- pour les mots des fichiers patients et témoins
- pour les lexèmes des fichiers patients et témoins

Ces dictionnaires sont ensuite « triés » par nombre d'occurrence de manière décroissante. Pour ceux comportant les mots et lexèmes, ils sont mis en colonnes dans de nouveaux fichiers de sortie qui permettront d'effectuer une analyse de comparaison « manuelle » en comparant les mots les plus utilisés.

Il existe 3 fichiers de sortie :

- Tableau.txt : fichier de sortie principal qui est un tableau à deux entrées (transformé en tableau à une entrée pour faciliter l'analyser de donnée) comportant différentes valeurs calculées par le programme. Il y a 20 colonnes dont 10 pour les tags les plus utilisés.
- Frequence_mots.txt : Il s'agit du fichier comportant la fréquence des mots des 3 intervenants, triés par ordre décroissant
- Frequence_lexemes.txt : Il s'agit du fichier comportant la fréquence des lexèmes des 3 intervenants, triés par ordre décroissant

Pour plus d'informations sur le code et les fonctions, l'intégralité du code du programme est disponible en annexe.

Exemple du tableau créé :

Témoin	Patient	Mots	Lignes	Lexèmes	NbTAG	ADV	ADJ	NC	DET
0	1	3987	3017	3790	5830	5839	3547	2578	1580
1	0	4729	4826	7938	9866	8938	4976	3108	1973

B - Difficultés rencontrées

Lors du développement de ce programme, un des problèmes majeur a été l'encoding de toutes les entrées et toutes les sorties nécessaires en UTF8, qui est considéré comme l'encodage universel, celui posant le moins de problème.

Dans le corpus sur lequel nous avons travaillé, il existe 11 phrases non standardisées (suite à une typo peut être) qu'il a fallu prendre en compte.

Il fut par ailleurs difficile d'avoir assez de recul pour déterminer quelles données extraire du programme pour effectuer une analyse statistique pertinente.

C - Analyse de données sur les résultats du part of speech

Psy	Temoin	Patient	Mots	MotsNoS	Lignes	Lexemes	LexemesNoS	NbTAG	/ADV	/NC	/P	/V	/DET	/CC	/CLS	/ADJ	/PRO	/VPP
0	0	1	3380	3282	3788	5052	4995	44364	7314	5832	4894	4760	3533	2708	2419	2294	1625	1530
0	1	0	4809	4703	7281	7705	7647	73404	16432	10044	7735	6652	6179	4721	3583	3102	3051	2525
1	0	0	5779	5669	11454	10407	10348	130804	20806	17561	15025	14841	10698	7842	6782	6181	5349	4298

Figure 1 : Résultat obtenu par MElt

La figure ci-dessus donne le résultat en sortie de l'analyse de corpus via le programme qui a été créé.

Nous avons décidé de faire une analyse de donnée pour savoir si une éventuelle corrélation existait entre le fait d'être un patient schizophrène et d'utiliser préférentiellement des adverbes, verbes, sujet… pour s'exprimer.

Nous avons donc fait quelques modifications minimes sur ce tableau pour travailler sur la répartition des Tags :

	/ADV	/NC	/P	/V	/DET	/CC	/CLS	/ADJ	/PRO	/VPP
Patient	7314	5832	4894	4760	3533	2708	2419	2294	1625	1530
Temoin	16432	10044	7735	6652	6179	4721	3583	3102	3051	2525
Psy	20806	17561	15025	14841	10698	7842	6782	6181	5349	4298

Figure 2 : Tableau crée pour analyse dans R

Nous allons donc, via une analyse factorielle des correspondances étudier la corrélation entre les catégories grammaticale les plus utilisées et le type de population s'exprimant dans le corpus.

Le programme nous a permis d'obtenir les 10 catégories grammaticales les plus utilisées qui sont donc : ADV pour adverbe, NC pour nom commun, P pour préposition, V pour verbe, DET pour déterminant, CC pour conjonction de coordination, CLS pour les pronoms clitiques sujets, ADJ pour adjectif, PRO pour pronom et enfin VPP pour participe passé.

Pour effectuer l'analyse, nous avons utilisé le logiciel R qui est un logiciel d'analyse statistique.

Nous avons donc rentré le tableau en figure () et effectué un calcul des effectifs du tableau présenté si dessus sous l'hypothèse de l'indépendance :

```
> tableau
          ADV    NC     P     V   DET   CC  CLS  ADJ  PRO  VPP
Patient  7314  5832  4894  4760  3533 2708 2419 2294 1625 1530
Temoin  16432 10044  7735  6652  6179 4721 3583 3102 3051 2525
Psy     20806 17561 15025 14841 10698 7842 6782 6181 5349 4298
```

Figure 3 : Tableau des données dans R

```
> theorique
             ADV        NC         P         V       DET       CC      CLS      ADJ      PRO      VPP
Patient  7818.567  5867.962  4853.085  4607.220  3581.814 2679.955 2243.503 2031.683 1759.318 1465.894
Temoin  13562.436 10178.828  8418.379  7991.889  6213.174 4648.769 3891.681 3524.249 3051.792 2542.805
Psy     23170.997 17390.210 14382.536 13653.892 10615.013 7942.276 6648.815 6021.068 5213.890 4344.302
```

Figure 4 : Tableau des effectifs théoriques dans le cas d'une indépendance

On remarque qu'il n'y a pas de différences très grandes entre le tableau des effectifs théoriques et celui de données que nous avons eu de MElt.

Ce que pourrait donc signifier qu'il n'y a pas de phénomènes de dépendance : un patient n'utilise pas préférentiellement tel ou tel autre catégorie grammaticale.

Pour cette analyse, nous avons rencontré quelques difficultés : en effet le corpus étant très grand, nous avons eu des difficultés à extrapoler l'interprétation du test du Chi2 pour savoir s'il était significatif (par rapport aux exemples que nous avons vu en cours d'analyse de données). En effet un corpus très grand donne un Chi2 significatif alors que les données ne le sont pas forcément, il aurait fallu préciser en faisant d'autres tests, mais par manque de connaissances dans le domaine d'analyses de données et par manque de temps, nous n'avons pas pu effectuer de tests supplémentaires.

```
        Pearson's Chi-squared test

data:  tableau
X-squared = 1453.166, df = 18, p-value < 2.2e-16
```

Figure 5 : Test du Chi2

Avec p-value < 0.05, le Chi2 est normalement significatif, ce qui voudrait donc dire que notre analyse de données est significative.

Cela correspond à un seuil de 5% (on fixe généralement ce seuil à 5%) : ce qui veut dire qu'il y a moins de 5% de chance que les données analysées soient dû au hasard.

Nous n'avons également pas pu obtenir un graphique de la projection de ces données, le logiciel R nous mettais un message d'erreur que nous n'avons pas réussis à résoudre, ni à contourner.

Psy	Temoin	Patient	Mots	MotsNoS	Lignes	Lexemes	LexemesNoS	NbTAG
0	0	1	3386	3282	3788	5052	4995	44364
0	1	0	4809	4703	7281	7705	7647	75404
1	0	0	5779	5669	11454	10407	10348	130894

Figure 6 : Tableau de données

Légende :

NbTAG : La somme de tous les Tags existants : c'est-à-dire le nombre de mots

Mots : Le nombre de mots différents

MotsNoS : Le nombre de mots différents sans les stopwords

Lignes : Le nombre de lignes

Lexemes : Le nombre de lexemes

LexemesNoS : Le nombre de lexemes sans les stopwords

En soustrayant les mots différents aux mots différents sans les stopwords, on obtient des valeurs très proches pour les 3 profils. On en conclue que les schizophrènes n'utilisent pas plus de mots « inutiles », dénués de sens. On obtient le même résultat avec les lexèmes.

En divisant le nombre de mots par le nombre de lignes on obtient le nombre de mots par ligne. Encore ici, les valeurs sont quasiment identiques pour les 3 profils : ils ont des phrases de même longueur.

En comparant les lexèmes aux nombres de mots, on obtient des valeurs très proches pour les patients témoins et les patients schizophrènes alors que le psychologue possède un nombre de lexèmes proportionnellement plus élevé. On peut en conclure que le psychologue utilise un vocabulaire un peu plus varié mais on ne peut rien conclure à propose de la personne schizophrénique.

On constate qu'avec les données obtenues on ne peut pas conclure sur une caractéristique du discours propre aux schizophrènes.

Partie 2 : L'annotation en pragmatique

L'autre campagne que nous avons été amenées à conduire concerne l'annotation de transcription d'entretien entre schizophrène et psychologue auprès de personnes « physique ».

Cette campagne d'annotation se base notamment sur le fait que le discours d'un schizophrène contiendrait des « discontinuités » ou ruptures dans le discours. (Musiol et Trognon, 2000 ; Musiol et Verhaegen, 2002 ; Verhaegen, 2007). Pour notre campagne d'annotation, nous nous intéresserons particulièrement aux discontinuités décisives.

Selon la thèse de Verhaegen (2007), Il existe plusieurs types de discontinuités décisives dans le discours du patient schizophrène, et qui serait des preuves manifestes de la pathologie.

- Tout d'abord « le débrayage conversationnel » : ce type de discontinuité se caractérise par le fait que le patient schizophrène entame un changement de sujet, sans indication de ce changement, alors qu'il a pourtant initié le sujet.

- Ensuite « les séquences à double discontinuité réactive » cela se caractérise par une cohérence deux à deux des interventions, mais si on analyse sur son ensemble le discours il y a une incohérence entre la première intervention et la troisième (tout en ayant une cohérence entre les deux premières interventions et les deux dernières).

- Enfin, « la défectuosité de l'initiative conversationnelle » : dans ce type de discontinuité, le schizophrène ne respecte pas les contraintes de communication usuelles (dans un dialogue, on répond à ce qui est dit précédemment et on continue le dialogue, ici ces contraintes ne seraient pas respectées).

Ces différents types de discontinuité peuvent être étudiés grâce à la création d'arbres hiérarchiques de ce type :

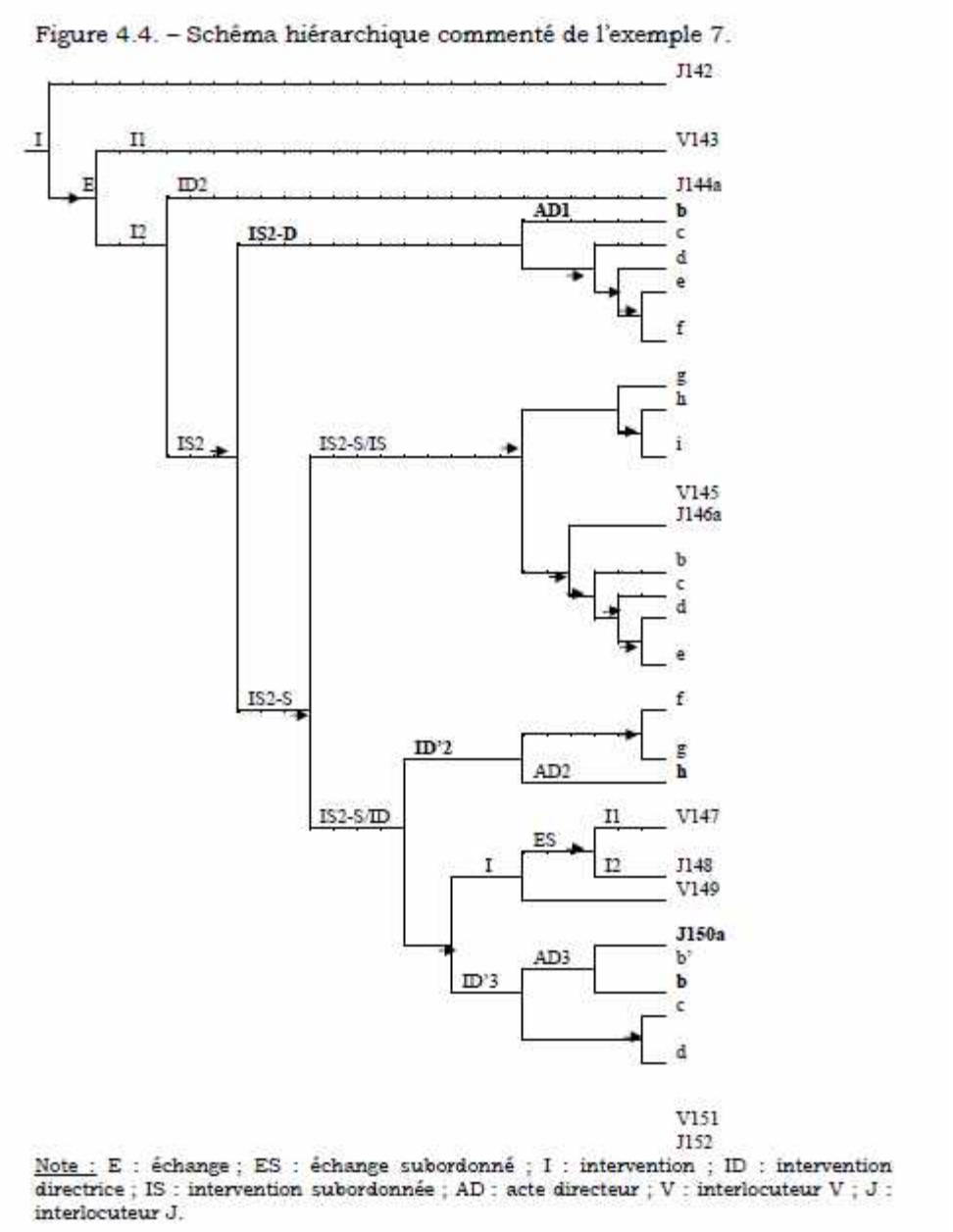


Figure 7 : Figure issue de la thèse de Verhaegen (2007), de type « défectuosité de l'initiative conversationnelle », qui est un bon exemple d'arbre hiérarchique

L'idée serait donc de faire effectuer aux annotateurs la tâche de créer des arbres hiérarchiques. Pour cela il est nécessaire de passer par un certain nombre d'étapes pour mener la campagne d'annotation.

A - Elaboration de la campagne d'annotation

Pour élaborer la campagne d'annotation, il est nécessaire de l'organiser en passant par plusieurs étapes distinctes :

En référence à la thèse de Karen Fort – (« Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus », 2012) - il est important de consacrer une partie du temps à un travail préparatoire. La rédaction du guide en fait partie, même si celui-ci n'est pas définitif et sera encore modifié durant la campagne elle-même.

Il faut également déterminer la population cible ainsi que le nombre de personnes auquel il faudra faire annoter les corpus.

Dans un premier temps il a été envisagé de faire passer les annotations à un groupe d'annotateur expert : des psychologues ou des étudiants en psychologie, plus à même peut-être de travailler et d'être familiarisé aux corpus qui seront présenté. Mais également à des linguistes ou des étudiants en linguistique, qui sont d'avantages familiarisé aux arbres hiérarchiques : « Sélectionner des annotateurs correctement formés donne de meilleur résultats » (Fort, 2012).

Ensuite, le nombre de personne pour faire une annotation doit être au minimum de deux (pour pouvoir constater une convergence des résultats ou non). Cependant un peu plus de personnes (3 ou 4) permettent de confirmer, et d'avantage de personnes permet de limiter le hasard.

Une fois ces constatations établies, nous avions donc pour intention de départ de faire passer les corpus qui nous ont été confié à environ 5 étudiants en psychologie ou en linguistique.

B - Rédaction du guide d'annotation

Une deuxième étape importante avant de commencer la campagne d'annotation en tant que telle a été la rédaction du guide d'annotation.

Cela constitue une phase essentielle, car c'est ce qui va faire office de référentiel pour l'annotateur. Il doit donc contenir le bon nombre d'indications sur ce qu'est l'annotation, dans quel but elle est réalisée, l'origine des corpus… sans être trop exhaustif pour ne pas ennuyer ou perdre l'annotateur avant même le début de la tâche.
Mais surtout elle doit contenir la démarche à suivre pour que l'annotateur effectue le travail qui lui est demandé.

Le guide doit pouvoir être compréhensible et assez accessible pour que l'annotateur puisse effectuer l'annotation sans forcément avoir des connaissances au préalable, que ce soit de la tâche (créer des arbres hiérarchiques) ou du sujet (les discontinuités décisives).

Ce fut l'une des grosses difficultés de cette partie : la théorie n'est pas forcément des plus accessibles, et il faut cependant l'assimiler suffisamment bien pour pouvoir la restituer de manière à ce qu'elle soit facilement compréhensible.

Il a donc été essentiel de se concentrer sur les points clés de la construction de l'arbre, d'expliquer les éléments nécessaires autant que possible sans noyer le lecteur en le surchargeant d'informations. Une première version a donc été écrite en tenant compte de toutes ces contraintes.

Cependant comme évoqué dans la partie précédente, le guide n'est à ce moment-là pas définitif. Il sera retouché au cours de la campagne au grès des remarques et des difficultés rencontrées.
La version « finale » du guide d'annotation est disponible en annexe.

C - Mise en place de la campagne d'annotation

Une fois ces différents éléments décidés et le guide d'annotation rédigé, nous avons donc entamé la campagne d'annotation en tant que tel.

Nous avions donc comme intention de départ de recruter environ 5 étudiants en psychologie ou en linguistique pour faire le travail d'annotation. Nous avions 4 extraits différents, certaines contenait une discontinuité décisive, d'autres n'en comptais pas du tout. Ils se composaient chacun de 45 à 60 interventions (répliques) en tout, ce qui représente un travail conséquent pour l'annotateur, venant du corpus de Lyon.
A cause de ces contraintes notamment, nous n'avons pas pu trouver comme nous l'avions voulu des annotateurs dans la population cité.

Nous avons cependant réussis à trouver des annotateurs pour effectuer le travail : il s'agit d'étudiant, de filières indifférenciées.
Cependant là encore des difficultés sont intervenues. Après le premier extrait annoté, les annotateurs se sont pour beaucoup découragés pour faire la suite des extraits.
Et en l'occurrence certains n'ont pas réussis à aller au bout du premier extrait.

Plusieurs choses peuvent expliquer cela :

- Les extraits à annoter étaient particulièrement long, le premier contient 60 interventions, le temps nécessaire pour annoter variait entre 1h et 2h en fonction de l'annotateur. Les extraits suivant était sensiblement plus court à annoter (on peut penser qu'une certaine habitude a été mise en place, ce qui a facilité et raccourcis le temps passé par l'annotateur sur les extraits suivants, mais aussi tout simplement car ils étaient plus court).
- Le guide d'annotation était sans doute perfectible. Il aurait nécessité peut-être plus d'explications (ou moins) ou d'être clarifié sur plusieurs points, chose qui a essayé d'être faite au cours des quelques retouches effectuées au cours de la campagne.
- L'annotation qui était effectuée était d'un niveau d'abstraction élevée : il ne s'agissait pas d'un simple travail ou la personne doit répondre à une simple question (elle sait où elle ne sait pas) ou encore ou il s'agit de donner un avis sur une échelle (notion de ressentie personnel, automatisme). Dans notre cas la personne devait effectuer un

travail conséquent nécessitant une concentration soutenue durant une longue période, ce qui peut expliquer que certains, par manque de motivation peut-être, n'ont pas été au bout des annotations.

D - Résultats de la campagne d'annotation

Au final 8 personnes se sont prêtés à l'expérience. Sur ce nombre :

- 2 personnes ont tenté d'annoter et n'ont finalement pas été au bout d'un extrait
- 1 personne a annoté l'extrait 2, mais nous avons décidé de ne pas le retenir, la structure de l'arbre étant très différentes des autres productions
- 5 personnes ont pu annoter le premier extrait (contenant une discontinuité). Parmi ces personnes :
    - 1 annotateur a fait en plus l'extrait 2
    - 1 a annoté en plus l'extrait 3

Les productions de ces 5 annotateurs sont disponibles en annexe de ce rapport (avec en haut à droite : a pour annotateur ; e pour l'extrait ; p pour le numéro de page).

Nous nous sommes donc concentrés sur l'analyse du premier extrait. Les autres extraits ayant eu trop peu d'annotation n'aurait pas été aussi pertinent à interpréter.

De façon générale, les arbres de ce premier extrait ont une certaine structure similaire : en effet ils contiennent tous 6 Interventions Directrices (sauf dans l'arbre de l'annotateur 3 qui en contient 7). Les annotateurs ont donc presque tous repéré plus ou moins les même blocs de sens (une Intervention Directrice représentant en gros une idée résumable facilement dans un dialogue, elle est généralement composé de plusieurs tours de parole).

Cependant lorsque nous allons plus dans le détail, les branches des arbres sont très différentes d'un annotateur à l'autre, les répliques marquants les débuts et fin d'Intervention Directrices sont différentes pour chaque annotateur ou presque. Seul la réplique 42 est considéré par 4 annotateurs sur 5 comme étant un début d'Intervention Directrice.

Intéressons-nous maintenant à ce pour quoi nous avons fait ces arbres : est-ce que ce travail d'annotation a permis de repérer une discontinuité ?

Dans cet extrait, une discontinuité est présente à la ligne 25.

En regardant les arbres, nous remarquons que seulement l'annotateur 1 met clairement en évidence cette discontinuité :

La personne a scindé en trois (a, b et c) la réplique 25 du patient. Pour l'annotateur, les répliques Pa a et Pa b appartiennent à l'Intervention Directrice 2 (I2) tandis que la réplique Pa 25 c est affilié à l'Intervention Directrice 3 (I3).
De par son arbre, l'annotateur 1 a donc repéré cette discontinuité.

Pour les autres annotateurs, on peut constater qu'autour de la réplique 25 il y a une construction plus complexe pour les annotateurs 2, 4 et 5, mais rien ne peut laisser penser qu'il y a eu un repérage de la discontinuité (ces « constructions plus complexe » sont peut-être présentes à d'autres endroit, alors qu'il n'y a aucune discontinuité existantes).

E - Améliorations envisageables

Une amélioration qui pourrait être envisagé pour ce type de travail d'annotation serait de créer une interface, ce qui aurait plusieurs avantages :

- Faciliter pour l'annotateur le travail. Même si cela n'enlève pas la charge cognitive d'une annotation de ce type, cela permettrait de rendre l'annotation plus ludique et moins rébarbative et peut-être plus rapide également.

- Cela faciliterait l'interprétation des arbres obtenus. Et cela permettrait d'appliquer des algorithmes de façon automatique pour obtenir des conclusions plus poussées.

Ces quelques améliorations permettraient de rendre plus « simple » l'annotation, de la faire éventuellement à plus grande échelle et de permettre une interprétation plus rapide et de diversifier les types de résultats que l'on pourrait obtenir à partir d'un seul arbre (comparer les structures autour d'une rupture décisive, les comparer au reste du corpus ect).

F - Bilan de la campagne d'annotation

Cette campagne d'annotation a donc pu être mené à son terme malgré plusieurs obstacles et un moins grand nombre de données recueillies que ce qui avait été prévu et espéré au départ.

Finalement, le fait que ce soit des étudiants de cursus indifférencié plutôt que d'étudiants en psychologie ou en linguistique n'a pas été gênant : au final les difficultés ont plus été d'ordre de temps ou de motivation de la part de l'annotateur, plutôt qu'une incompréhension de la tâche d'annotation, ou de la théorie derrière la tâche.

Ce qui est encourageant, car cela peut vouloir dire qu'à priori n'importe qui (quel que soit les prérequis antérieur de la personne dans le domaine) peut effectuer une tâche d'annotation de ce type, ce qui permet de faciliter la recherche d'annotateur.

Au final, dans notre campagne d'annotation, une seule personne (sur 5 annotateurs) a pu clairement mettre en évidence une discontinuité présente dans le corpus. Ce qui est encourageant (car cela veut dire que c'est possible), mais ce n'est pas encore suffisant.
En améliorant peut être certains éléments (avec la création d'une interface, comme évoqué dans la partie « Amélioration ») ou en clarifiant peut-être d'avantage le but de la tâche dans le Guide d'Annotation, peut être que le résultat en sera amélioré.

III Conclusion

Notre projet tutoré avait pour objectif d'effectuer un travail d'annotation en morphosyntaxe et un travail d'annotation en discontinuité décisive.

Bien que dans les deux cas il s'agisse d'annotation, les deux travaux sont d'une nature différente :

L'annotation en morphosyntaxe a été réalisée sur un énorme corpus issu de dialogues entre psychologues et schizophrènes ayant eu lieu à Lyon. Afin d'extraire des données de ce corpus, nous avons utilisé un système d'étiquetage pour tagger l'intégralité du texte en Part Of Speech. Puis nous avons créé un programme permettant d'obtenir des fichiers possédant des données exploitables (nombre de mots utilisés, fréquence des mots, catégorie grammaticale la plus utilisée, etc…). Ainsi, nous avons finalement pu faire une analyse de données et arriver à la conclusion qu'il n'y a pas de profil se dégageant. La théorie d'un discours propre aux schizophrènes n'est pas vérifiée ici.

L'annotation en discontinuité décisive a nécessité un travail préparatoire avant de lancer la campagne en elle-même, et l'annotation s'est faite par des annotateurs humains et les résultats étaient donc dépendants de la préparation de la campagne et du travail des annotateurs en eux-mêmes.

Dans les deux cas, les campagnes avaient pour but de systématiser la mise en évidences des caractéristiques de la schizophrénie mais elles ont sollicitées différentes compétences étant donné que les travails ont été différents.
La pluridisciplinarité de notre formation a été un avantage pour la réalisation du projet. En effet nous avons mobilisé notamment des compétences en informatique, linguistique, psychologie, statistique.

Ce projet tutoré nous a permis d'approfondir certaines notions vues en cours, mais également de se rendre compte des réalités du monde de la recherche comme les contraintes : notamment de la difficulté d'obtenir des corpus de patients schizophrènes, pour des raisons légales, ce qui complexifie le problème déjà posé.

IV Bibliographie

- Maxime Amblard, Michel Musiol, and Manuel Rebuschi.
  Une analyse basée sur la S-DRT pour la modélisation de dialogues pathologiques. In Traitement Automatique des Langues Naturelles - TALN 2011, page 6, Montpellier, France, June 2011.

- Maxime Amblard, Michel Musiol, and Manuel Rebuschi.
  Schizophrénie et Langage : Analyse et modélisation. De l'utilisation des modèles formels en pragmatique pour la modélisation de discours pathologiques. In Congrès MSH 2012, Caen, France, December 2012.

- Christophe Benzitoun, Karën Fort, and Benoît Sagot.
  TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In Traitement Automatique des Langues Naturelles (TALN), pages 99–112, Grenoble, France, 2012.

- Karën Fort.
  Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus. PhD thesis, Université Paris XIII, LIPN, INIST-CNRS, December 2012.

- Frédéric Verhaegen.
  Psychopathologie cognitive des processus intentionnels schizophréniques dans l'interaction verbale. PhD thesis, Université Nancy 2

- Michel Musiol, Frédéric Verhaegen.
  Appréhension et catégorisation de l'expression de la symptomatologie schizophrénique dans l'interaction verbale. Février 2008

- Michel Musiol, Manuel Rebuschi.
  « La rationalité de l'incohérence en conversation schizophrène (Analyse pragmatique conversationnelle et sémantique formelle) » Mars 2006

- . Pascal Denis and Benoît Sagot (2012).
  Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. In *Language Resources and Evaluation* 46:4, pp. 721-736, DOI 10.1007/s10579-012-9193-0.]

V Annexes

Guide d'annotation pour l'identification des discontinuités décisives dans la transcription des entretiens entre schizophrène et psychologue

Personnes : M. Amblard, M. Musiol, K. Fort, M. Rebuschi, N. Wittmann, K. Rivalin

Gestionnaire de la campagne : N. Wittmann, K. Rivalin

1. Contexte

Dans le cadre du projet SLAM, des études sont conduites sur des entretiens entre des patients atteints de schizophrénie et des psychologues.
Les corpus que nous allons étudier sont constitués des transcriptions de ces entretiens. Le type d'annotation à réaliser sur ces corpus est effectué dans le but de mettre en avant certains éléments caractéristiques de la schizophrénie.

2. Objectif de la transcription

Les chercheurs accordent une place non négligeable aux troubles de la pensée (et donc du langage) en ce qui concerne la schizophrénie.

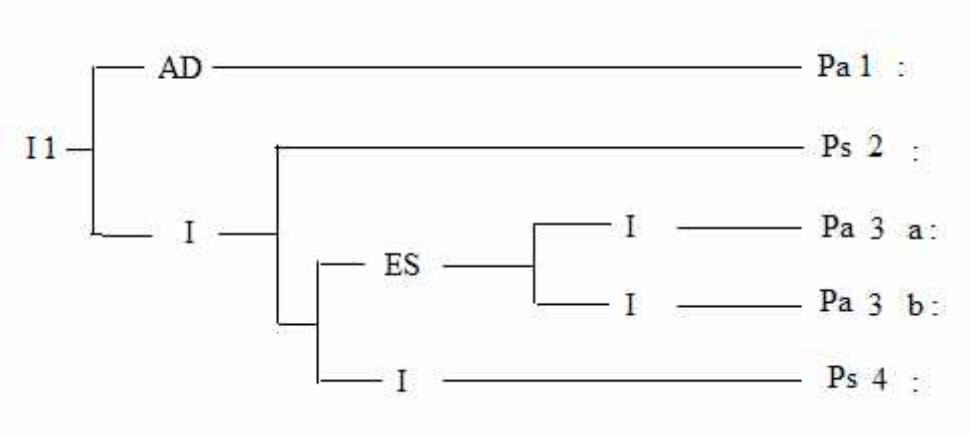
Une des caractéristiques de la schizophrénie serait la présence dans le discours du patient de « discontinuité » : c'est-à-dire une sorte de « rupture » dans le discours entre ce qui est dit précédemment et ce qui est dit par le patient.
Pour mettre en évidence ces discontinuités, ces ruptures dans le dialogue, un moyen serait de créer des arbres hiérarchiques. Il s'agit d'arbres qui décrivent le cheminement du discours.
La façon de créer et de construire un arbre à partir d'un dialogue de transcription entre patient schizophrène et psychologue est décrite dans la partie 4 (ci-après).

3. Identification des données

Ce sont des transcriptions (conversion de données oral en données écrites) issues de discours entre patients schizophrènes et psychologues, venant du CH de Vinatier à Lyon.

4. Annotation : Démarche à suivre

Un arbre hiérarchique permet de détailler un dialogue. Dans notre cas il se fera entre deux personnes : un patient schizophrène et un psychologue (Pa pour patient et Ps pour psychologue).
Ce guide a donc pour but de décrire la façon de procéder pour créer un arbre.
Les transcriptions sont données sous forme de corpus numéroté pour faciliter le repérage, et les lignes du psychologue sont mises en italique.
Le but de l'annotateur est de reprendre les corpus qui lui seront confié et de créer des arbres finals de ce type :



Il est à noter qu'il n'y a pas de bonnes ni de mauvaises réponses, l'annotateur devra simplement suivre le guide qui explique pas à pas la façon de créer l'arbre.
La marche à suivre est la suivante :

Dans un premier temps, chaque intervention de chacun des deux intervenants doivent être placées les unes à la suite des autres dans l'ordre (on change de ligne à chaque fois que c'est l'autre locuteur qui prend la parole).

——— Ps 4 : En quelle année êtes vous né ?

——— Pa 5 : En 1975

——— Ps 6 : Dans quel ville ?

Parfois, au sein de l'intervention du patient ou du psychologue, deux idées différentes seront exprimées, ou bien la personne va répondre à une intervention précédente et relancer la discussion sur autre chose…
Dans ce cas l'annotateur pourra juger nécessaire de scinder l'intervention en autant de sous partie qu'il sera nécessaire pour lui : a, b, c…

——— Pa 15 a Je crois que vous, avez raison...

——— Pa 15 b Mais pensez-vous que je devrait...

Ici c'est la même personne qui parle (le patient), mais il semble nécessaire de séparer l'énoncé (dans la première partie il répond à ce qui est évoqué précédemment, dans l'autre partie il relance la discussion).
Nous avons donc ainsi jusque-là une suite d'interventions, il faut donc maintenant les identifier plus précisément et les rassembler pour former l'arbre en lui-même.

➔ Pour cela il faut identifier le type de déclaration dont il s'agit :

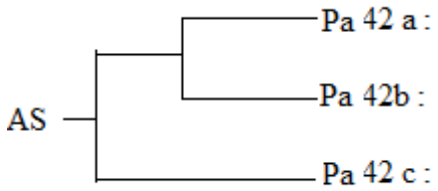
- AD : Acte Directeur : une déclaration qui a pour but « d'orienter » le discours dans une certaine direction : « Que faites-vous quand vous êtes chez vous ? », « Ou allez-vous faire vos courses ? ». Ou bien encore la réponse à une question qui va donner une nouvelle orientation à la conversation : « Je pense que ce que je préfère d'avantage que l'écriture, c'est encore le cinéma, j'aime beaucoup… ».
  Il est à noter qu'un Acte Directeur a pour but d'orienter la conversation dans une nouvelle direction, cela ne veut pas dire que l'Acte Directeur y parviendra toujours. Un AD peut lancer un sujet, mais l'autre interlocuteur ne rebondira pas dessus.
  Il est à noter comme ceci :

—— AD —— Ps 8 :

- I : Intervention : des déclarations qui poursuivront le discours sans donner une nouvelle orientation particulière par rapport à la conversation précédemment engagée.

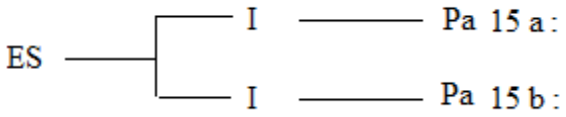
—— I ——Pa 8 :

- AS : Acte Subordonnée : ce sont des déclarations provenant d'une même personne, scindé en plusieurs sous interventions a, b, c.. Comme illustré si dessous :



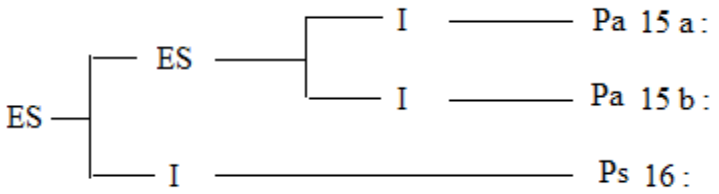
➔ Une fois que l'on a identifié les déclarations, on va tenter de les rassembler pour ainsi construire l'arbre en lui-même :

Deux Interventions (I) pourront être regroupées en Echange Subordonné (ES) (lorsque l'annotateur juge que les déclarations sont liées entre elles) :

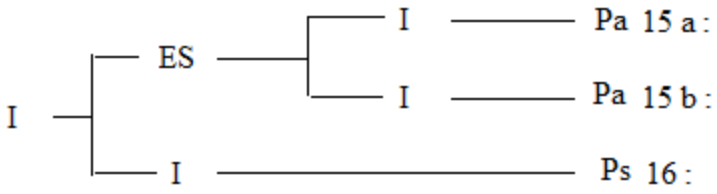


Par la suite, une Intervention (I) et un Echange Subordonné (ES) peuvent être rassemblé soit :

- en Echange Subordonné (ES) :



- En Intervention (I) :

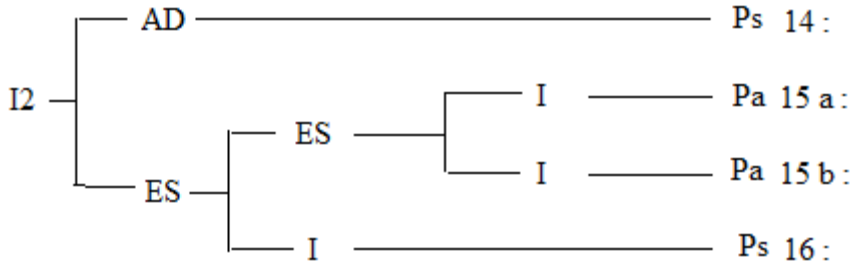


La différence entre les deux sera faites par l'annotateur : soit il considère qu'il s'agit d'avantage d'un échange d'idée (Echange) soit il s'agit d'avantage d'une déclaration dans la continuité (Intervention).

Si vous rencontrez des cas qui ne sont pas traité par le guide (comme l'association d'un AD avec un AS par exemple, associez le en fonction de ce qui vous semble être juste.

En remontant ainsi petit à petit, on va trouver ce qu'on appelle une Intervention Directrice (I1, I2, I3…), c'est-à-dire des interventions qui ont chacune une orientation précise : I1 va parler de tel sujet, I2 plutôt de tel autre.
Il s'agit donc d'un regroupement d'Interventions, d'Actes Directeurs, d'Echange Subordonné qui vont former un bout de dialogue qui signifie quelque chose, qui pourrait être résumé.

Cela peut donc donner un arbre tel que :

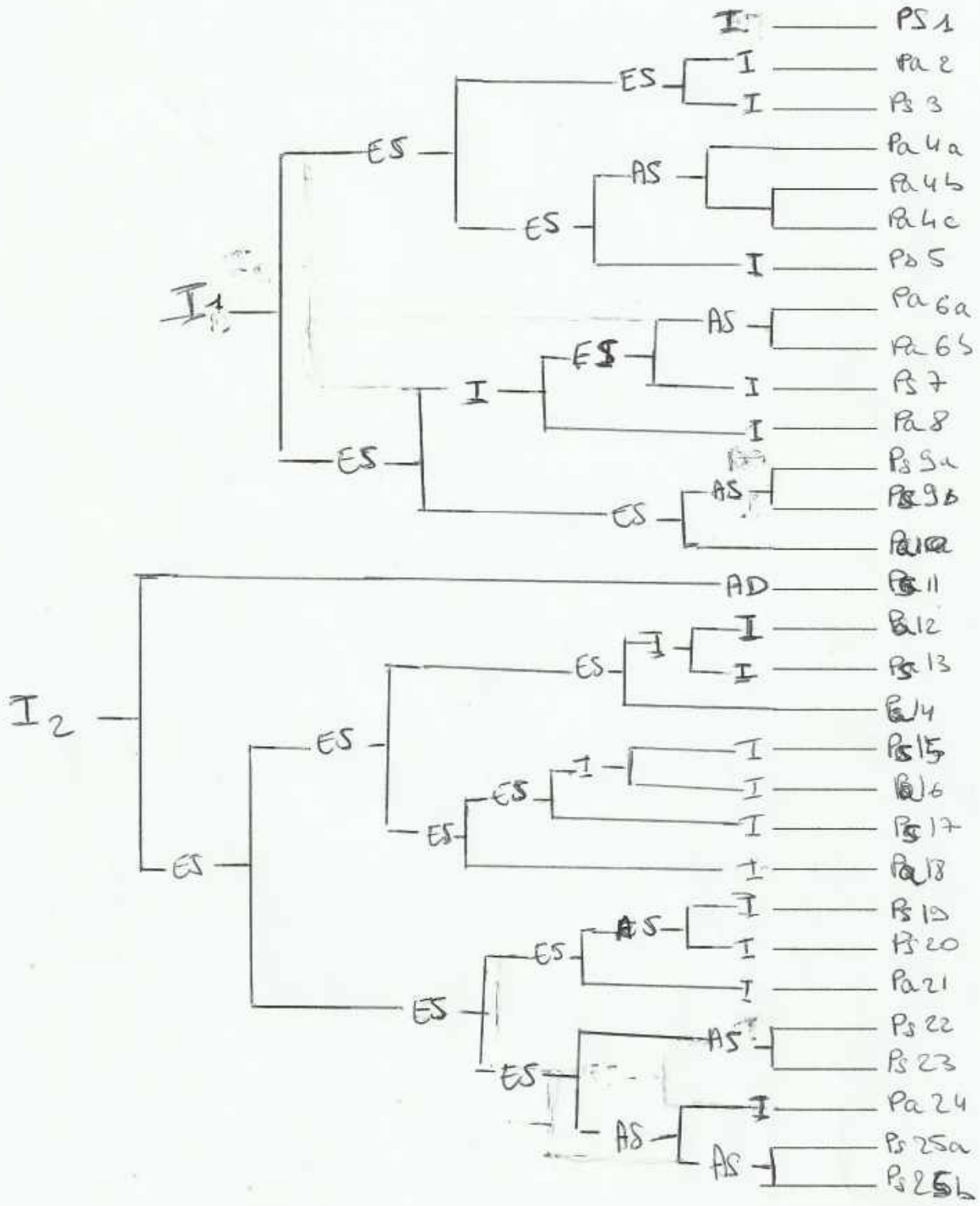


Les interventions les plus « liées » seront reliées en premier, et ainsi de suite jusqu'à arriver à l'Intervention Directrice.

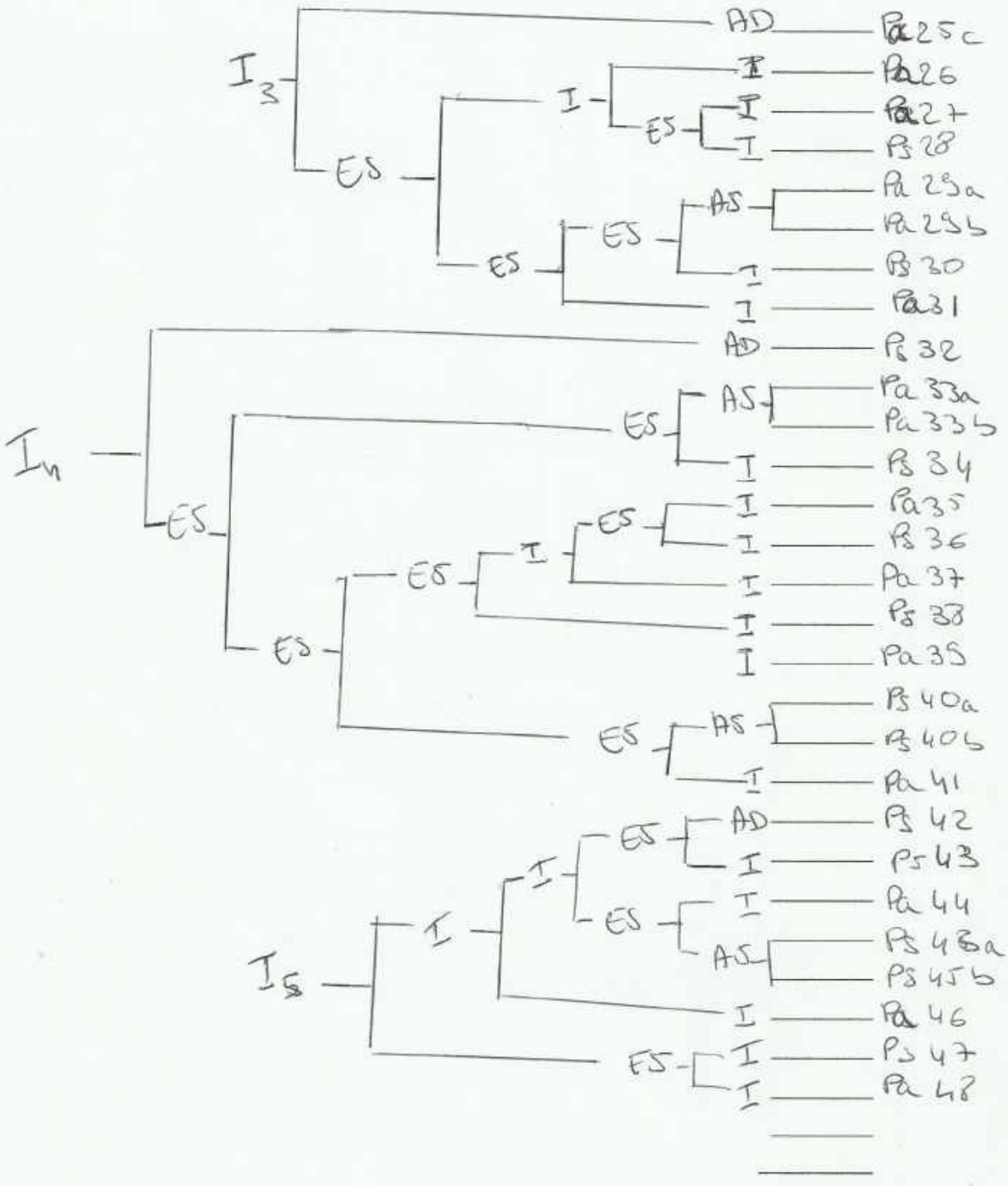
A la fin de l'étude d'un extrait, nous auront une suite d'Interventions Directrices, composées elles-mêmes de plusieurs tours de paroles (plusieurs Interventions, Actes Directeurs, Echanges subordonnée…).

Arbres hiérarchiques

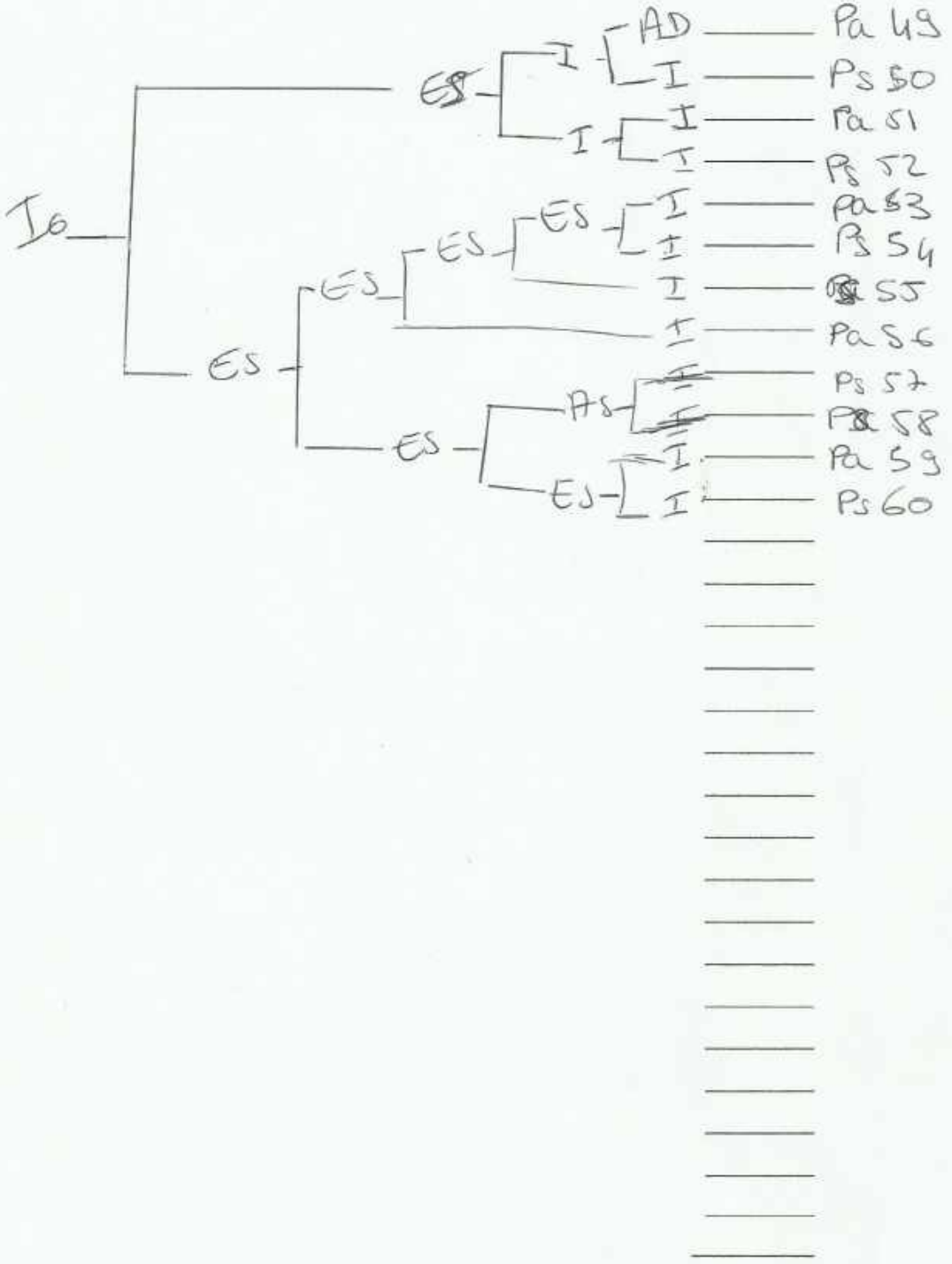
a:I , e:1, p1


I
ES
I
I
ES
AS
ES
I
AS
ES
I
I
ES
AS
I
ES
AD
I
I
ES
ES
ES
I
I
ES
I
ES
I
AS
I
I
ES
AS
ES
I
AS
AS
PS 1
Pa 2
Ps 3
Pa 4a
Pa 4b
Pa 4c
Ps 5
Pa 6a
Pa 6b
Ps 7
Pa 8
Ps 9a
Ps 9b
Pa 10a
Ps 11
Ps 12
Ps 13
Pa 14
Ps 15
Pa 16
Ps 17
Pa 18
Ps 19
Ps 20
Pa 21
Ps 22
Ps 23
Pa 24
Ps 25a
Ps 25b

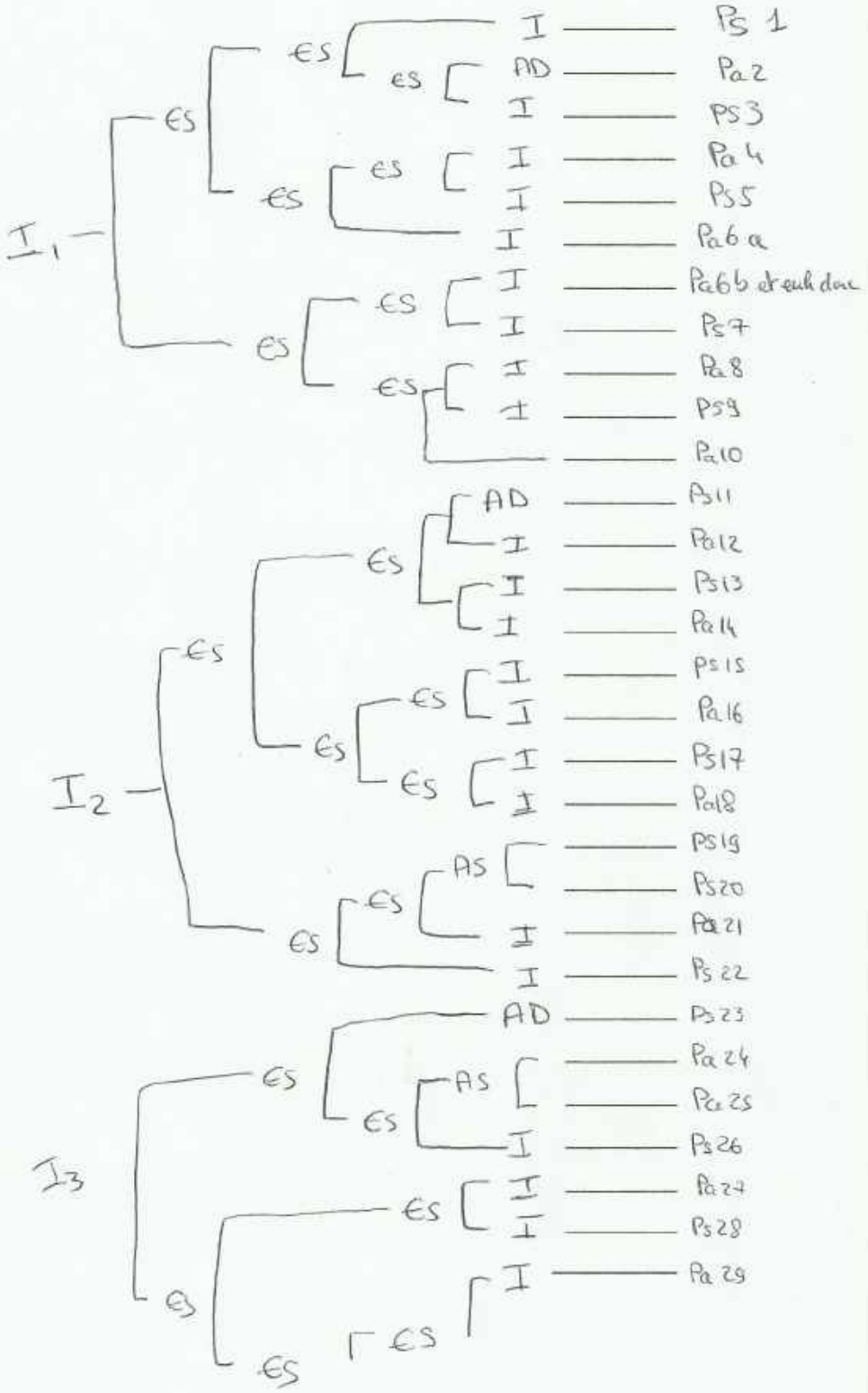
a:I, e:1, p2

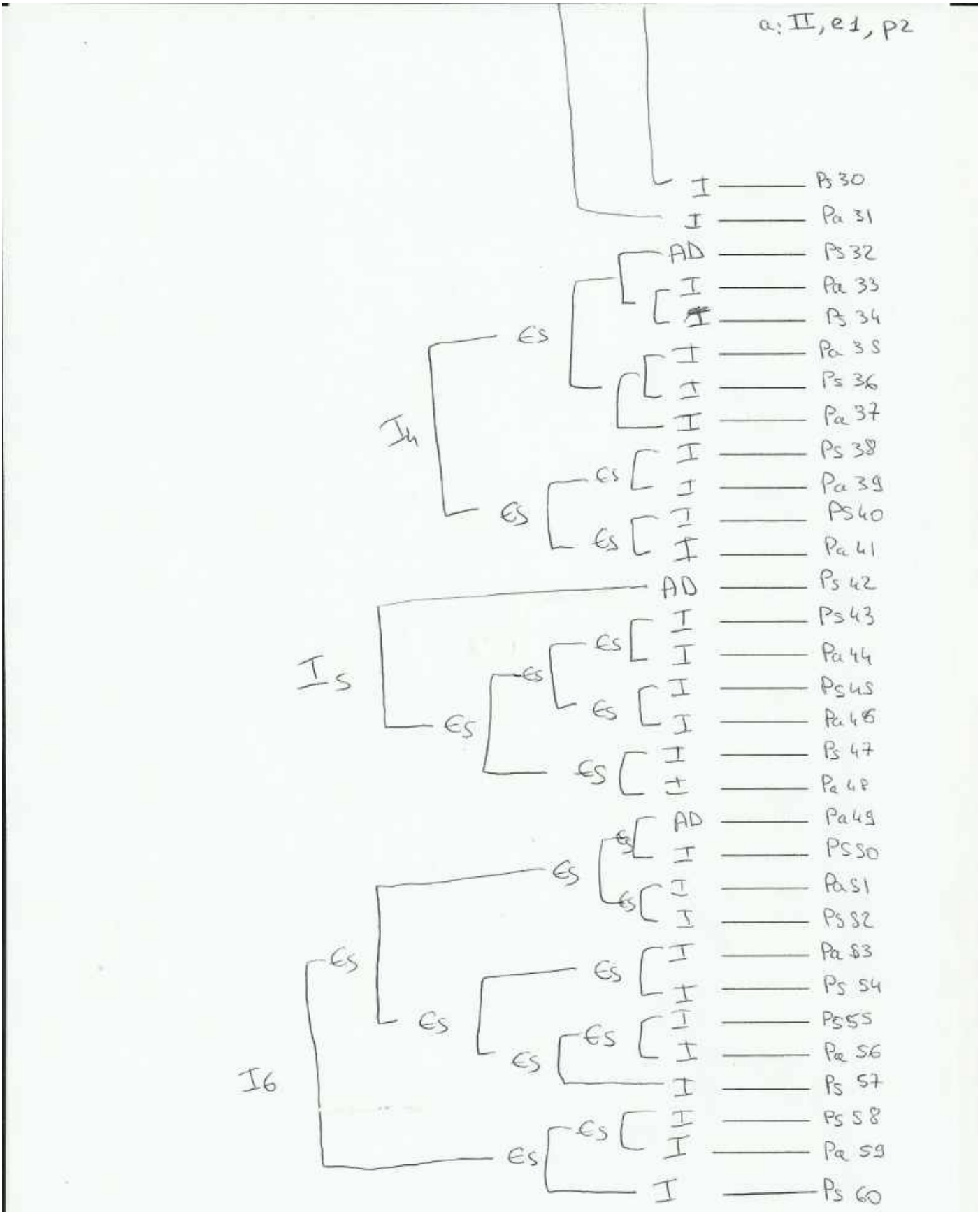

I_3
AD
Pa 25c
I
Pa 26
I
ES
I
Pa 27
I
Ps 28
ES
AS
Pa 29a
Pa 29b
ES
ES
I
Ps 30
I
Pa 31
I_4
AD
Ps 32
AS
Pa 33a
ES
Pa 33b
I
Ps 34
I
Pa 35
ES
I
Ps 36
ES
I
Pa 37
I
Ps 38
I
Pa 39
ES
ES
AS
Ps 40a
Ps 40b
I
Pa 41
I_5
AD
Ps 42
ES
I
I
Ps 43
I
Pa 44
ES
I
AS
Ps 45a
Ps 45b
I
Pa 46
I
Ps 47
ES
I
Pa 48

a:I, e:1, p3


Io
ES
I
AD
I
Pa 49
Ps 50
I
I
Pa 51
Ps 52
ES
ES
ES
I
I
Pa 53
Ps 54
ES
I
Ps 55
I
Pa 56
ES
Ps
I
Ps 57
I
Pa 58
ES
I
Pa 59
ES
I
Ps 60

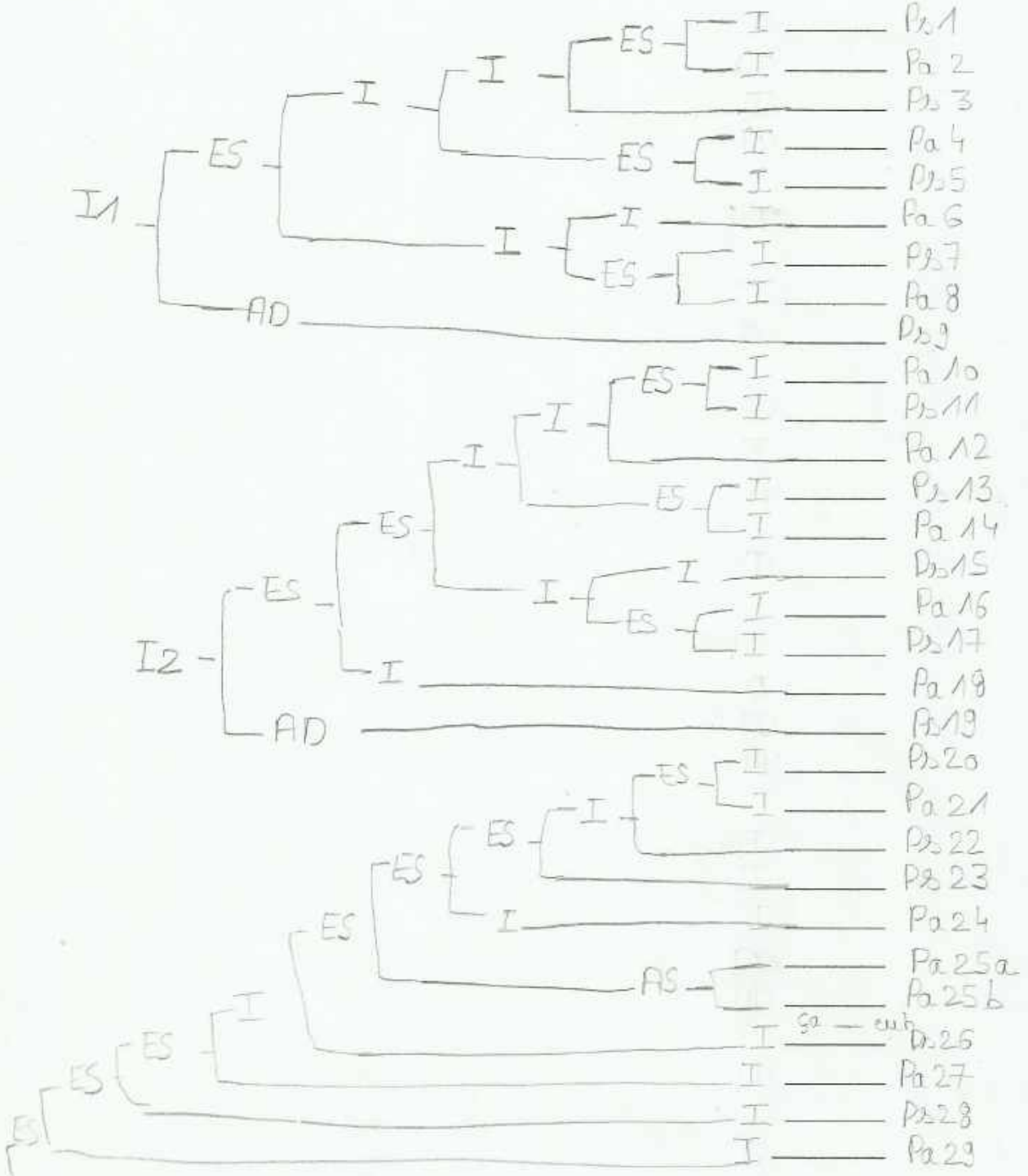
a:II, e1, p1


I1
I2
I3
ES
AD
AS
I
Ps 1
Pa 2
Ps 3
Pa 4
Ps 5
Pa 6 a
Pa 6 b et euh donc
Ps 7
Pa 8
Ps 9
Pa 10
Ps 11
Pa 12
Ps 13
Pa 14
Ps 15
Pa 16
Ps 17
Pa 18
Ps 19
Ps 20
Pa 21
Ps 22
Ps 23
Pa 24
Pa 25
Ps 26
Pa 27
Ps 28
Pa 29

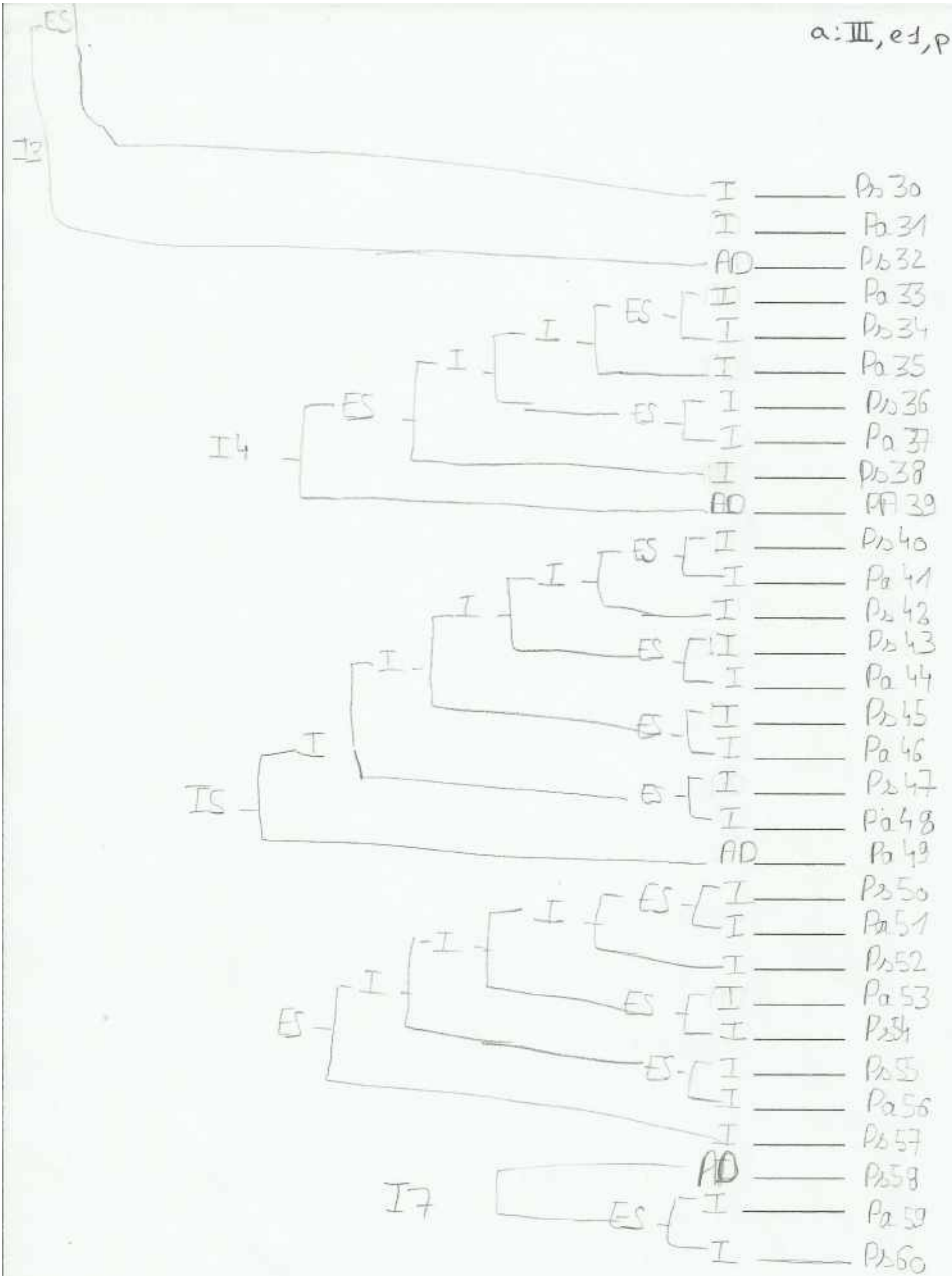

a: II, e1, p2
I
Ps 30
I
Pa 31
AD
Ps 32
I
Pa 33
I
Ps 34
Es
I
Pa 35
I
Ps 36
I
Pa 37
I
Ps 38
I
Pa 39
I
Ps 40
I
Pa 41
AD
Ps 42
I
Ps 43
I
Pa 44
I
Ps 45
I
Pa 46
I
Ps 47
I
Pa 48
AD
Pa 49
I
Ps 50
I
Pa 51
I
Ps 52
I
Pa 53
I
Ps 54
I
Ps 55
I
Pa 56
I
Ps 57
I
Ps 58
I
Pa 59
I
Ps 60
Es
I4
Es
Es
Is
Es
Es
Es
Es
Es
Es
Es
Es
Es
Es
Es
Es
Es
Es
Es
I6

(1)

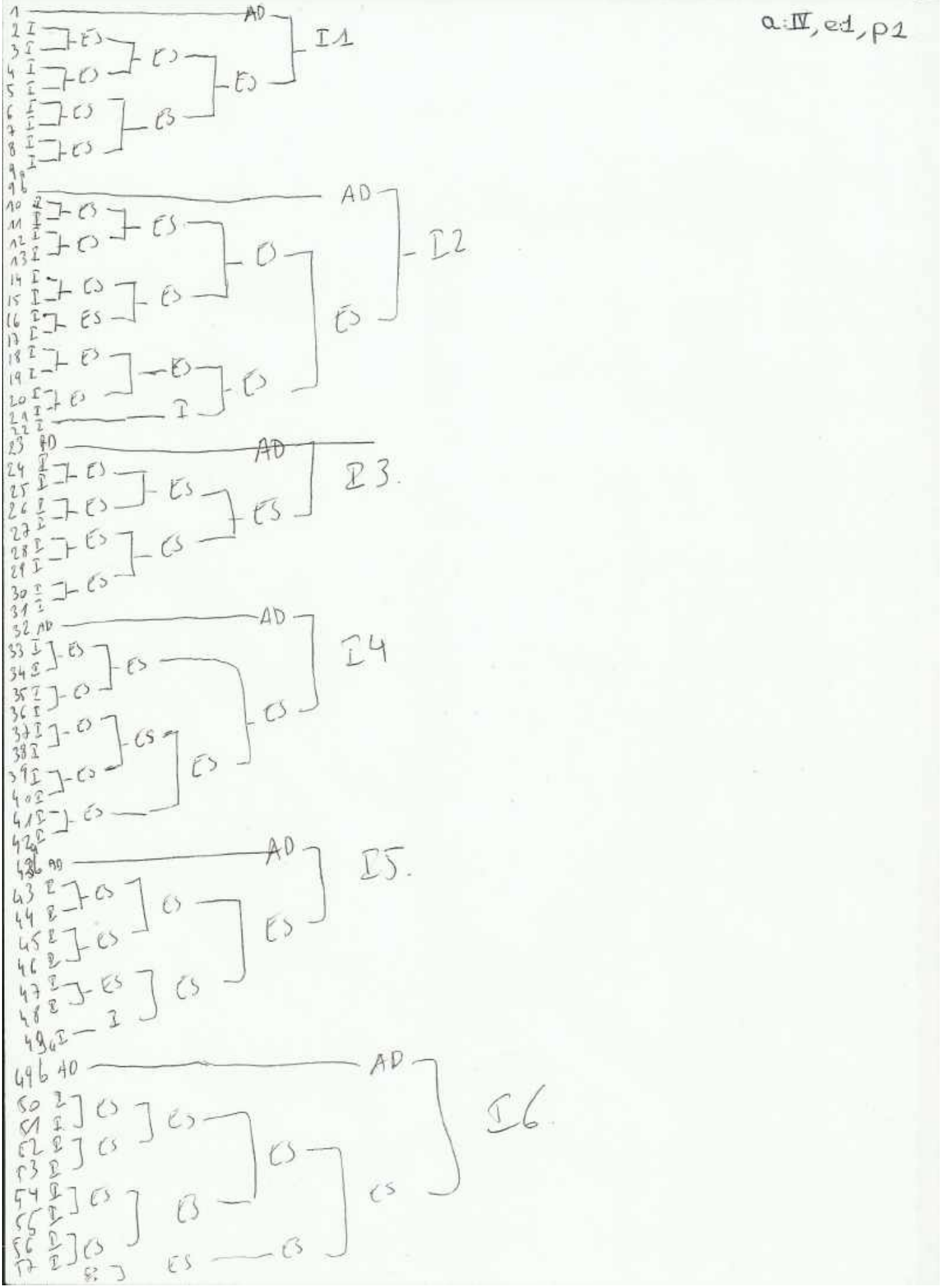
a: III, e1, p1


I1
I2
ES
I
AD
AS
Pa 1
Pa 2
Pa 3
Pa 4
Pa 5
Pa 6
Pa 7
Pa 8
Pa 9
Pa 10
Pa 11
Pa 12
Pa 13
Pa 14
Pa 15
Pa 16
Pa 17
Pa 18
Pa 19
Pa 20
Pa 21
Pa 22
Pa 23
Pa 24
Pa 25a
Pa 25b
Pa 26
Pa 27
Pa 28
Pa 29

a: III, e1, p2

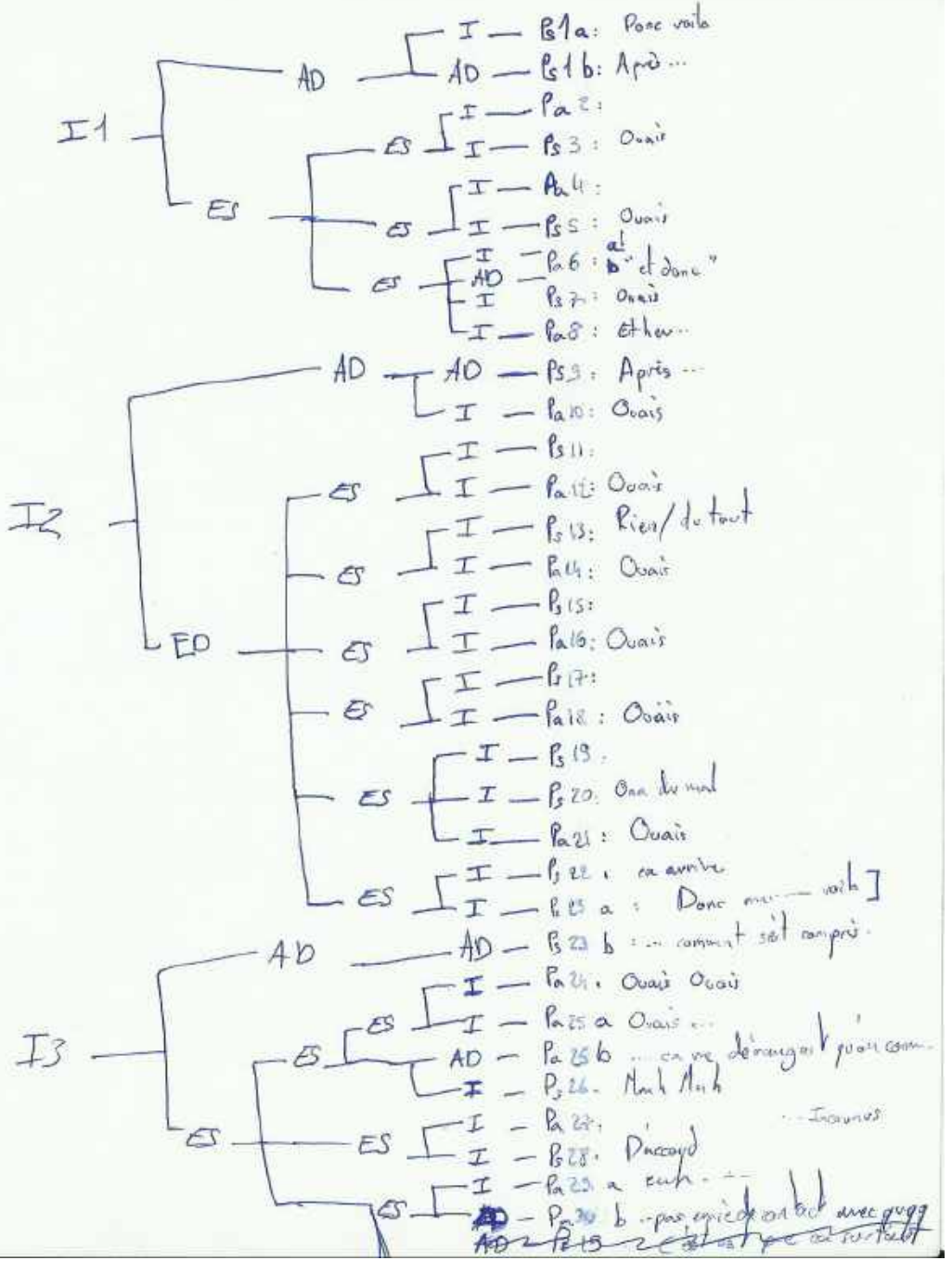

ES
I3
I — Pb 30
I — Pa 31
AD — Pb 32
ES
I — Pa 33
I — Pb 34
I — Pa 35
ES
I — Pb 36
I — Pa 37
I — Pb 38
AD — PA 39
I4
ES
I
I
ES
I — Pb 40
I — Pa 41
I — Pb 42
ES
I — Pb 43
I — Pa 44
ES
I — Pb 45
I — Pa 46
ES
I — Pb 47
I — Pa 48
AD — Pa 49
I5
I
I
I
ES
I — Pb 50
I — Pa 51
I — Pb 52
ES
I — Pa 53
I — Pb 54
ES
I — Pb 55
I — Pa 56
I — Pb 57
ES
I
I
AD — Pb 58
I7
ES
I — Pa 59
I — Pb 60

a:IV, e:1, p:1



Texte n°1.

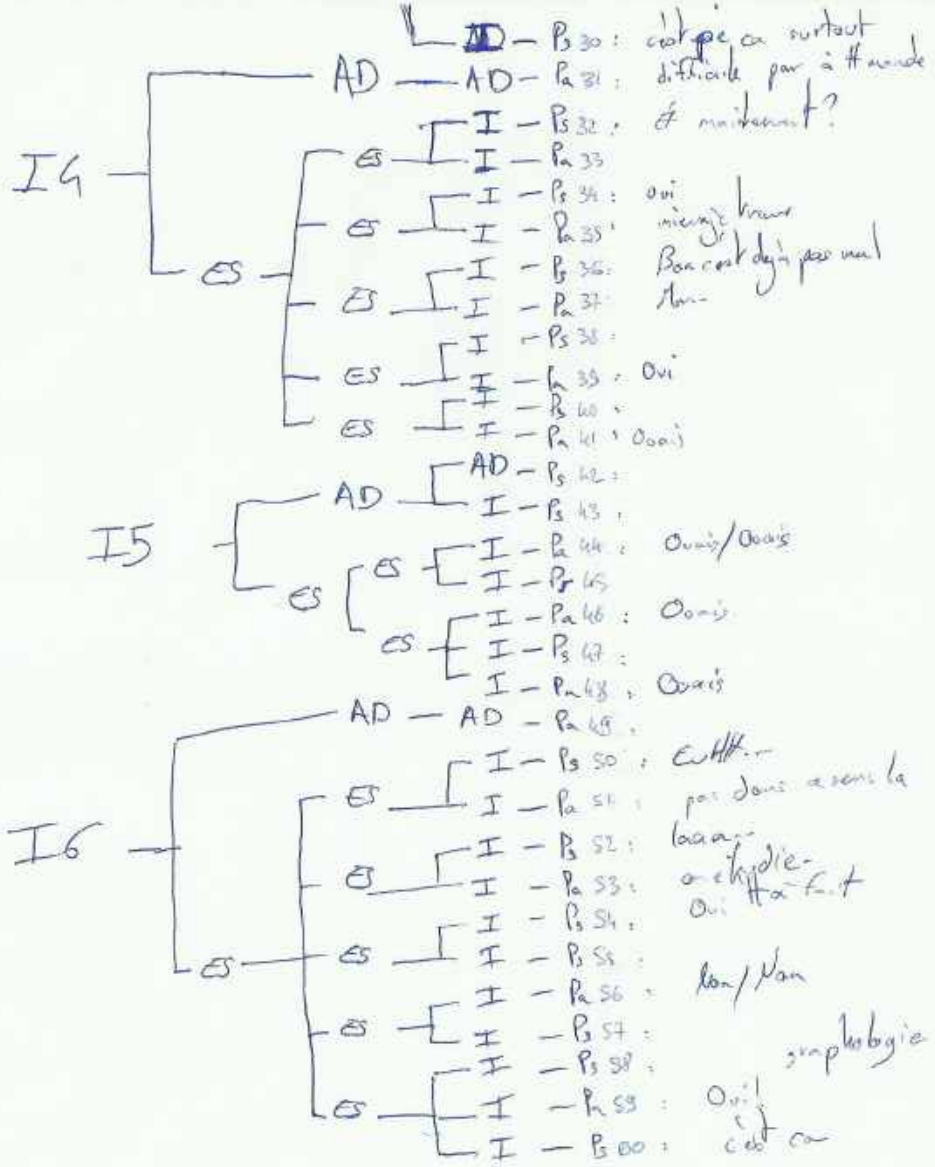
a:V, e:1, P1



Texte 01 suite

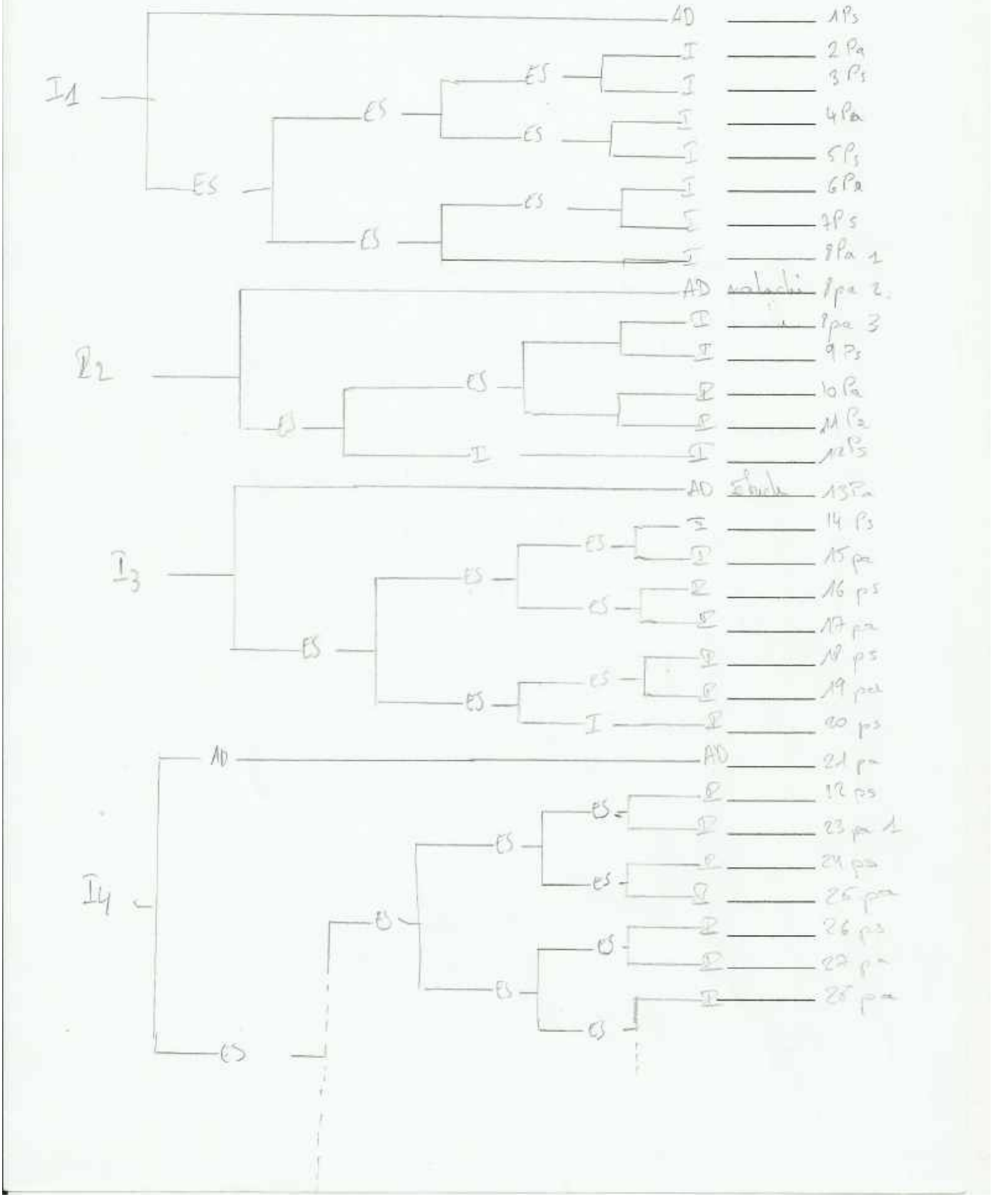
②

a I, e:1, p2

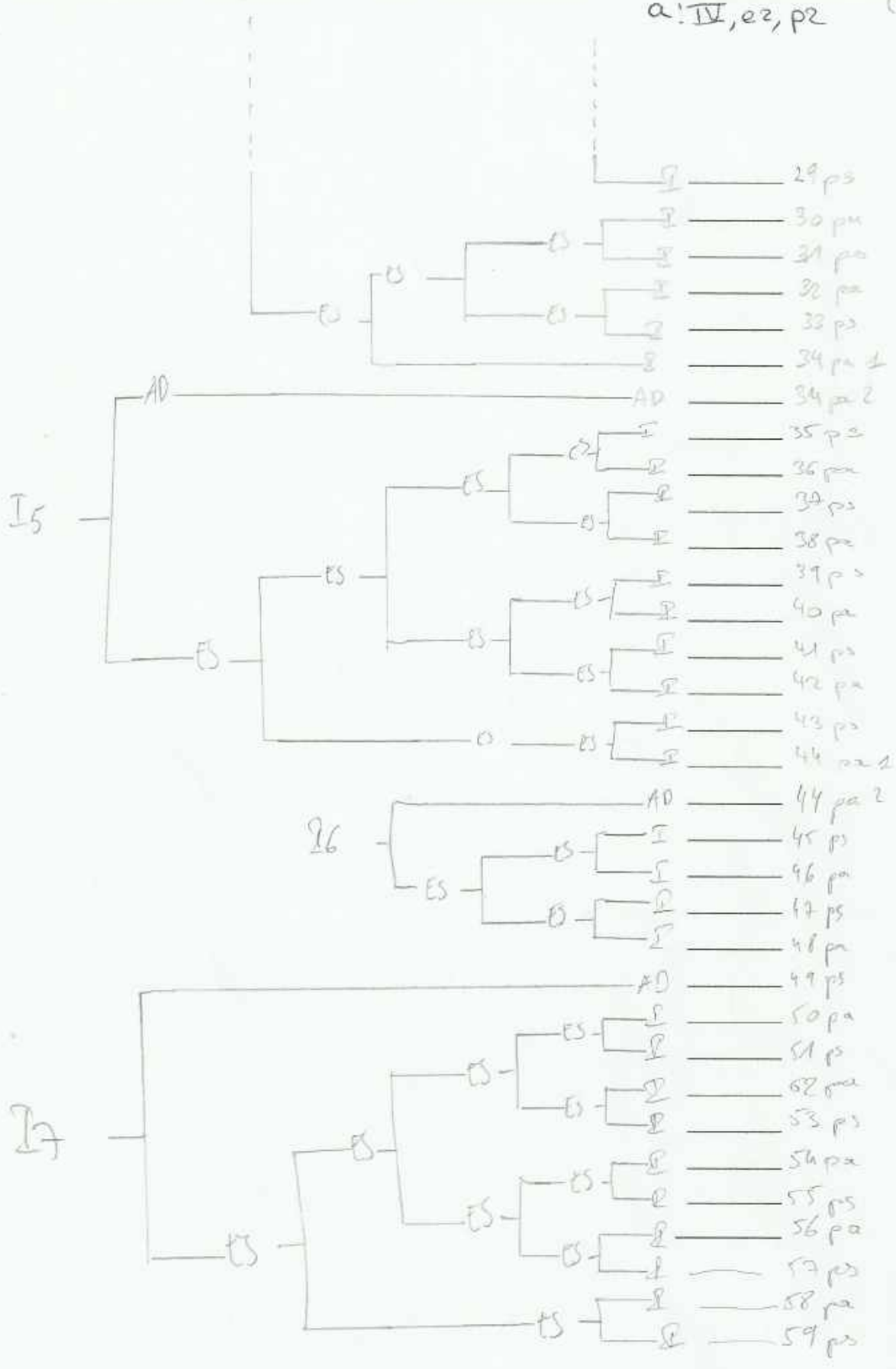


txt 2.

a : IV, e2, p1 (1)



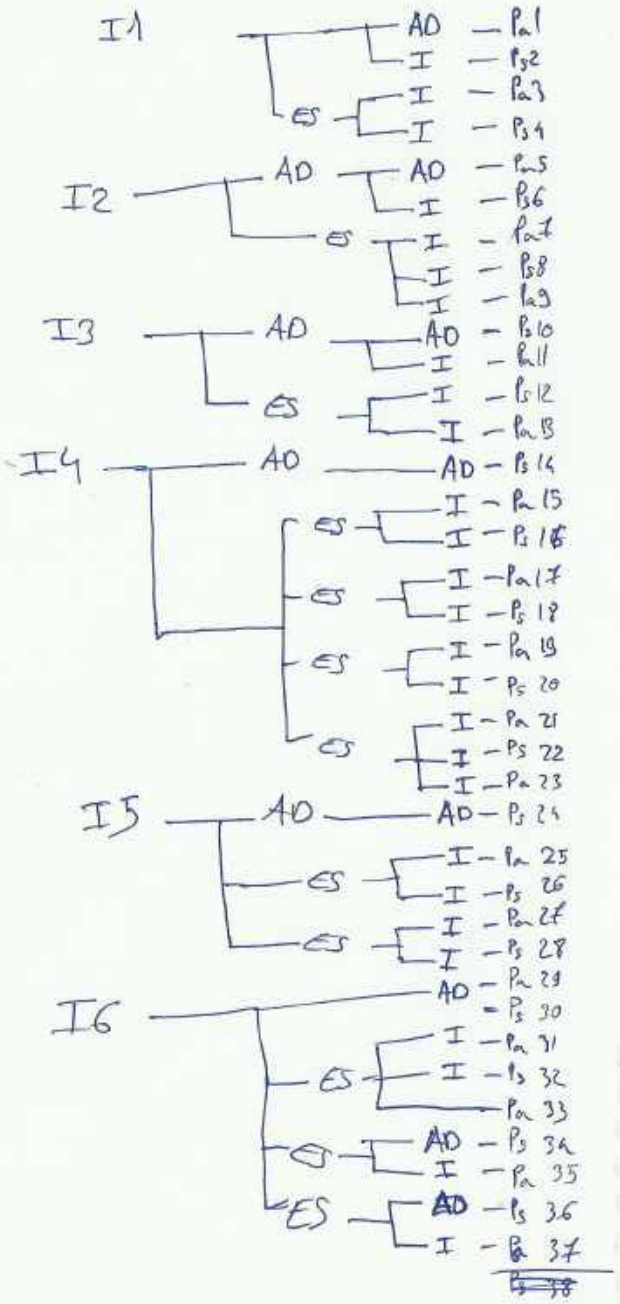
a: IV, e2, p2


29 ps
30 pm
31 pa
32 pa
33 ps
34 pa 1
AD
34 pa 2
35 ps
36 pm
39 ps
38 pa
37 ps
40 pa
41 ps
42 pa
43 ps
44 pa 1
I5
AD
44 pa 2
45 ps
46 pa
47 ps
48 pa
I6
AD
49 ps
50 pa
51 ps
52 pa
53 ps
54 pa
55 ps
56 pa
57 ps
58 pa
59 ps
I7

Texte n°3

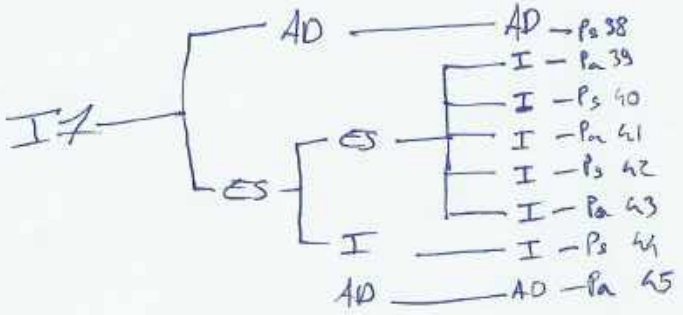
①

a. I, e. 3, p. 1



Texte n°3 :

a : I, e : 3, p2



Fichiers produit par le programme MElt : mots utilisés par les intervenants (25 premiers)

Patient	Témoin	Psychologue
('euh', 1038)	("c'est", 1791)	("c'est", 3933)
('de', 954)	('de', 1345)	('euh', 3315)
("c'est", 899)	('euh', 1299)	('de', 2862)
('à', 598)	('à', 803)	('à', 1705)
('un', 588)	('un', 777)	('un', 1656)
("j'ai", 504)	('euh...', 734)	('a', 1499)
('euh...', 502)	('a', 698)	('on', 1218)
('me', 490)	('on', 611)	('y', 1084)
('suis', 344)	('ouais', 583)	('euh...', 1022)
('ouais', 331)	('y', 494)	('coup', 838)
('on', 309)	("j'ai", 462)	('voilà, 823)
('a', 305)	('\nouais', 448)	('plus', 775)
('une', 242)	('me', 414)	('\nouais', 656)
('enfin', 230)	('enfin', 390)	('ouais', 641)
('puis', 230)	('\nMmh.', 354)	('enfin', 631)
('quoi', 221)	('oui', 352)	('faire', 589)
('plus', 220)	('\nhum', 345)	('une', 584)
('\n...', 211)	('une', 344)	('se', 517)
('bien', 210)	('plus', 337)	('va', 511)
('moi', 207)	('quoi', 307)	('après', 489)
('y', 199)	('\n...', 291)	('bien', 470)
('voilà', 196)	('moi', 288)	('faut', 438)
('ben', 185)	('voilà', 274)	('quoi', 432)

Fichiers produit par le programme MElt : lexèmes utilisés par les intervenants (25 premiers)

Patient	Témoin	Psychologue
('cln', 2014)	('cln', 2758)	('cln', 4761)
('le', 1257)	('le', 2269)	('le', 4398)
('un', 1083)	('un', 1513)	('*c'est', 3120)
('de', 921)	('*c'est', 1430)	('un', 3070)
('*euh', 880)	('de', 1301)	('de', 2787)
('être', 816)	('que', 1256)	('*euh', 2775)
('que', 797)	('avoir', 1210)	('avoir', 2702)
('cela', 791)	('cela', 1203)	('cela', 2647)
(‘*c'est’, 767)	('*euh', 1116)	('et', 2345)
('pas', 731)	('pas', 1108)	('pas', 2281)
('ou', 617)	('être', 1068)	('que', 2148)
('et', 617)	('*', 961)	('être', 1985)
('avoir', 600)	('et', 826)	('à', 1683)
('à', 592)	('faire', 795)	('*', 1653)
('en', 543)	('à', 792)	('en', 1610)
('cld', 497)	('en', 702)	('ou', 1458)
('faire', 492)	('ou', 673)	('faire', 1424)
(‘*l'ai’, 419)	('qui', 662)	('qui', 1346)
('son', 390)	('cll', 625)	('cll', 1332)
('qui', 362)	('mais', 582)	('*du', 1287)
('tout', 322)	('tout', 581)	('donc', 1136)
('mais', 316)	('cld', 548)	('tout', 1108)
('donc', 314)	(‘*j'ai’, 478)	('mais', 1103)
('aller', 298)	('*ouais', 463)	('ce', 1061)
(‘lui', 270)	('dire', 433)	('cld', 999)