UNIVERSITÉ DE PARIS-SORBONNE, PARIS 4
UFR DE SOCIOLOGIE ET D'INFORMATIQUE POUR LES SCIENCES HUMAINES (ISHA)

Construction de ressources langagières annotées par myriadisation pour le traitement automatique des langues peu dotées : le cas de l'alsacien

---

Mémoire
présenté par

Alice Millour

M2 recherche - Linguistique et informatique

dirigé par

Karën Fort

Soutenu le 17 mai 2016

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Ce projet n'aurait pas pu voir le jour sans l'aide de nombreuses personnes et je tiens à les remercier ici tout particulièrement.

Un grand merci à Karën Fort pour son soutien, sa disponibilité, ainsi que pour ses nombreux conseils tant méthodologiques que scientifiques.

Je remercie également Claude Montacié pour son soutien tout au long de ce travail ainsi que la confiance qu'il m'a accordée.

Merci à Delphine Bernhard et Lucie Steiblé de s'être montrées si réactives et de nous avoir fournies de nombreuses données à la base du développement de ce projet.
Un grand merci à toutes les personnes qui m'ont aidé à faire connaître mon travail en diffusant le lien de l'application, en particulier aux membres de l'OLCA et à Thierry Kranzer, président du FILAL.

Je remercie également particulièrement Emilie pour ses recommandations techniques et sa disponibilité ainsi qu'Andy pour son soutien, sa patience et sa présence à mes côtés.

Enfin, un grand merci aux utilisateurs de la plate-forme qui par leur motivation, leurs contributions et leurs retours ont permis de donner une vie et un sens à ce projet.

INTRODUCTION

Ce projet s'inscrit dans une démarche de conservation du patrimoine humain que constitue la diversité linguistique. Il vise à proposer et à évaluer l'efficacité d'un outil permettant l'intégration d'une langue peu dotée au domaine du Traitement Automatique des Langues (TAL) : l'alsacien.

L'existence et la disponibilité de ressources langagières spécifiques au TAL et en particulier de corpus annotés constitue la condition *sine qua non* de l'élaboration et de l'évaluation des outils de TAL et de TAP. Pour toute langue, une des premières étapes nécessaires à l'élaboration des outils lui permettant de s'intégrer dans les technologies actuelles est en effet – tokénisation mise à part – la constitution d'un corpus annoté en parties du discours ainsi que d'un lexique. Cette première brique permet ensuite de procéder à une analyse morphologique et morpho-syntaxique plus avancées, puis à une analyse syntaxique, et enfin sémantique.

Or, les campagnes d'annotation produisant des ressources pour le traitement automatique des langues sont longues et coûteuses, et requièrent la mobilisation de linguistes pendant plusieurs années. En témoignent les campagnes de 1993 du Penn Treebank (Marcus *et al.*, 1993) – annotation d'un corpus d'anglais américain en morpho-syntaxe de 4.5 millions de mots–, ou celle du Prague Dependancy Treebank –annotation d'un corpus tchèque de plus d'un million de tokens au niveaux syntaxe et morpho-syntaxe en 5 ans pour un coût de 600 000 dollars (Böhmová *et al.*, 2003).

Une approche permettant de réduire ces coûts humains et financiers et ayant prouvé son efficacité pour la collecte d'annotations en TAL est celle de la myriadisation (*crowdsourcing*) : « activité qui consiste à faire produire (des idées, des annotations, un dessin, un vote, à une masse de gens, aujourd'hui principalement *via* Internet » (Fort, 2016). On peut notamment citer dans le domaine de la production collaborative de ressources langagières le succès de la plate-forme JEUXDEMOTS (réseau lexical de 835 000 termes connectés par plus de 42 millions de relations lexicales et sémantiques (Lafourcade *et al.*, 2015) verifier la ref), ou de ZOMBILINGO (application d'annotation collaborative grâce à laquelle 646 internautes ont permi de produire 107 000 annotations en syntaxe en dépendance pour le français à ce jour).

Le but de ce projet est donc de proposer une plate-forme pour l'annotation linguistique en parties du discours par myriadisation pour l'alsacien. L'application BISAME permet en effet

de mettre à contribution tout locuteur désireux de participer à la création de ressources annotées pour sa langue. Nous forçons certaines bonnes pratiques et proposons une méthodologie saine d'évaluation des données recueillies permettant de garantir la qualité du corpus annoté obtenu.

Grâce à ce projet, nous montrons que la mise à contribution des locuteurs peut pallier l'absence de ressources financières ou humaines dédiées à l'élaboration de corpus annotés de qualité.

CHAPITRE 1

PRÉSENTATION DU SUJET - PROBLÉMATIQUES

1.1 LANGUES PEU DOTÉES ET TAL

1.1.1 TECHNOLOGIE ET SURVIE DES LANGUES

La problématique de l'extinction des langues n'a rien de nouveau et inquiète entre autres la communauté linguistique. En 2000, Claude Hagège conclut *Halte à la mort des langues*, sur les opportunitées offertes par Internet où « se multiplient les échanges au moyen [...] de ces langues » (yiddish, langues régionales de France etc.) « Il s'agit d'un support qui donne une nouvelle voix, même dans le mirage du virtuel, à des idiomes qu'on risquait de ne plus entendre ».
La conservation de la diversité linguistique dans les zones connectées passe par la possibilité pour chacun de pouvoir utiliser les moyens de communication et d'information actuels dans sa langue maternelle. Faciliter l'utilisation d'une graphie au sein des technologies modernes permet en outre d'asseoir une tradition écrite tout en familiarisant les anciens comme les nouveaux locuteurs à celle-ci.

Les domaines d'application du TAL sont alors multiples lorsque l'existence et le maintien d'une écriture sont assurés. Par exemple :

— le domaine de la santé, des média (diffusion d'une presse écrite, accès à l'information),
— le domaine de l'éducation - enseignement (élaboration de manuels scolaires, standardisation de la langue),
— le tourisme,
— la gestion des cas de crise (en cas de catastrophe naturelle et d'intervention externe par exemple).

Par ailleurs l'essor des technologies de TAP (Traitement automatique de la parole) ouvre la possibilité d'inclure aux technologies du langage les langues à tradition purement orale.

Nous nous plaçons ainsi dans la continuité des travaux menés par Vincent Berment en 2004 (Berment, 2004), proposant des méthodes pour informatiser les langues, en particulier les « peu dotées » : « En dépit du caractère manifestement politique de ce mouvement d'affirmation des langues — si l'on s'accorde, avec Hannah Arendt ([Arendt 1995]), pour dire que « la politique repose sur un fait : la pluralité humaine » — l'idée s'impose alors qu'aux moyens traditionnels doivent s'ajouter les outils informatiques appropriés sans lesquels les buts visés ne peuvent plus être atteints. L'informatisation occupe ainsi une place essentielle dans cette vaste mobilisation culturelle et linguistique. »

Il apparaît ainsi primordial de mettre à disposition des utilisateurs les outils leur permettant d'employer leur langue : claviers adaptés, outils de traitement de texte, correcteurs orthographiques, outils de traduction automatique etc.

1.1.2 Pourquoi construire des corpus pour le TAL ?

Les corpus sont indispensables dans la discipline du TAL, du fait notamment de l'omniprésence des algorithmes d'apprentissage. Une grande majorité des outils issus du TAL (moteurs de recherche, traducteurs automatiques, outils de reconnaissance de la parole, de désambiguïsation lexicale, de résolution d'anaphore, d'étiquetage morpho-syntaxique) sont en effet au moins partiellement développés grâce à des techniques d'apprentissage supervisé sur corpus de référence. Par ailleurs, l'évaluation de tout outil requiert l'utilisation de corpus annotés de référence sur lesquels sont testées les performances de l'outil.

Deux axes de recherche évoluent parallèlement : d'une part, il s'agit trouver de nouveaux moyens de collecter des données adaptées au traitement automatique pour les langues peu dotées, et d'autre part d'améliorer les algorithmes nécessitant peu de ressources annotées.

En ce qui concerne la tâche d'annotation en TAL pour les langues peu dotées, on distingue trois classes principales de méthodes, reposant chacune sur un type d'algorithme d'apprentissage :

- **L'apprentissage non supervisé** : l'absence de ressources pour les langues peu dotées a été un moteur pour le développement de l'apprentissage non supervisé.

Un algorithme non supervisé fonctionne par extraction de connaissances sans données d'apprentissage, donc sans nécessité de ressource langagière spécifique. Ces technologies sont aujourd'hui peu utilisées en TAL car l'exactitude atteinte est encore trop faible. (Voir (Christodoulopoulos *et al.*, 2010), pour l'état de l'art des technologies non supervisées pour l'annotation en parties du discours.)

- **L'apprentissage semi-supervisé** : cette approche repose sur l'existence de ressources et d'outils pour les langues bien dotées et sur l'adaptation de ces derniers permettant un « transfert » vers une langue peu dotée. Deux méthodes principales ont été identifiées : (Täckström *et al.*, 2013) tire parti conjointement de l'existence de corpus bilingues alignés et de lexiques frustes (en l'occurrence, le Wiktionnaire [1]) pour développer de nouveaux outils d'annotation pour les langues peu dotées. L' exactitude des outils d'annotation obtenus n'a pu être testée que sur des langues bien dotées pour lesquelles existent des corpus de référence annotés.

  C'est l'approche empruntée par (Scherrer et Sagot, 2013) sur les couples allemand–palatin et allemand–néerlandais. L'exactitude des outils ainsi crées est respectivement de 67.2 % pour le palatin et 60.7 % pour le néerlandais. Une autre méthode consiste à tirer parti de la proximité étymologique et morphologique pouvant exister entre une langue bien dotée et une qui l'est moins. Il est ainsi possible d'identifier des couples de cognats entre ces langues proches et de traduire aisément les mots dits « outils » dans la langue bien dotée ce qui améliore la performance des outils existant lorsqu'ils sont appliqués sur la langue peu dotée. Cette approche est également celle de (Bernhard et Ligozat, 2013) pour la création d'outils pour l'alsacien à partir de l'allemand, obtenant une exactitude de l'ordre de 75 %.

  Une troisième approche a également été rencontrée, dans le cas du bengali : (Dandapat *et al.*, 2007) tirent parti de la richesse morphologique spécifique à cette langue pour contraindre les étiquettes possibles et compenser le manque de ressources.

- **L'apprentissage supervisé** : les ressources suffisantes pour entraîner de tels outils existent aujourd'hui pour la vingtaine de langues « bien dotées » mentionnée en partie 2.4.1.

Outre ces méthodes statistiques, les méthodes symboliques fonctionnant sans apprentissage à l'instar de l'analyseur syntaxique Leopar (Perrier et Guillaume, 2013), construit à partir de grammaires développées par des linguistes, nécessitent également des corpus annotés de référence afin d'évaluer leurs performances.

1. Voir `https://fr.wiktionary.org/wiki/Wiktionnaire:Page_d%E2%80%99accueil`.

1.2 OBJECTIFS

1.2.1 OBJECTIF GÉNÉRAL

L'objectif général est de démontrer qu'il est possible de recueillir des ressources annotées de qualité auprès des locuteurs d'une langue peu dotée au moyen d'une application de myriadisation. Dans le cadre de ce mémoire, nous produisons un corpus annoté collaborativement en parties du discours pour l'alsacien.

1.2.2 OBJECTIFS SPÉCIFIQUES

Les objectifs spécifiques définis pour la collecte d'annotations collaboratives pour l'alsacien sont les suivants :

— définition des besoins de l'alsacien,

— état de l'art de l'existant pour l'alsacien,

— développement de la plate-forme,

— collecte des données,

— évaluation des ressources.

CHAPITRE 2

ÉTAT DE L'ART

2.1 LANGUES RÉGIONALES DE FRANCE ET PROJETS EN COURS

La figure 2.1.1 est une carte établie par l'UNESCO place les différents parlers régionaux en danger plus ou moins élevé d'extinction.

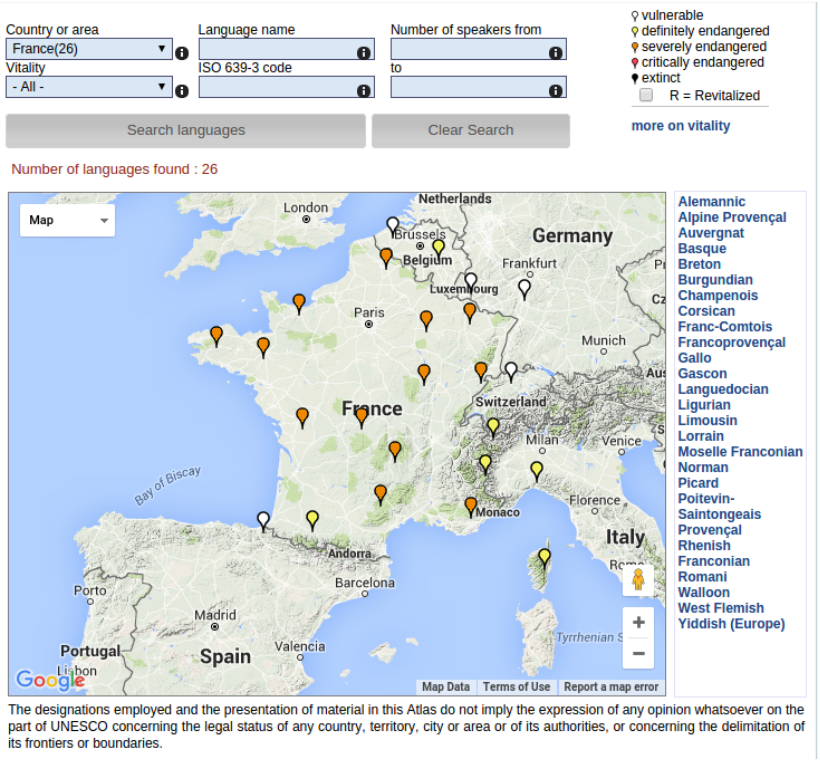

FIGURE 2.1.1 – Atlas UNESCO des langues en danger : les langues de France.

Depuis janvier 2015, le projet ANR RESTAURE (RESsources informatisées et Traitement AUtomatique pour les langues REgionales) s'emploie à développer ressources et outils pour

le TAL pour l'alsacien, l'occitan et le picard. Le projet *Lofloc* a vu le jour pour la création d'un lexique des formes fléchies de l'occitan, qui dispose d'ores et déjà d'un inventaire des ressources disponibles et d'un calendrier d'objectifs à atteindre quant à l'informatisation de l'occitan : « Diagnostic et feuille de route pour le développement numérique de la langue occitane : 2015-2019 » [1]. L'occitan dispose en outre d'une base de textes de 3 millions de mots, contenant des écrits dans neuf genres et cinq dialectes : BaTelÒc (Bras et Vergez-Couret, 2014), et un outil pour l'analyse morpho-syntaxique est développé en tirant parti de la parenté existant entre occitan et catalan. En ce qui concerne le picard, il existe également une base de corpus textuel de cinq millions de mots, apparentée à Frantext : Picartex (Eloy *et al.*, 2015) [2].

Pour la langue bretonne, (Foret *et al.*, 2015) dresse un état de l'art à jour des différentes ressources existant : plusieurs dictionnaires, un correcteur orthographique, plusieurs sites d'information en langue bretonne, mais aucun outil spécifique au traitement automatique du breton. Néanmoins le projet *Akenou-Breizh* mis en avant prévoit également d'utiliser la méthodologie de la myriadisation sous la forme d'un jeu ayant un but pour créer un corpus bilingue breton-français.

Les langues régionales ont en commun des caractéristiques devant être prises en compte dans la conception d'outils et d'applications visant à leur informatisation :

— absence de norme : existence d'un continuum dialectal, pas d'orthographe définie, tradition majoritairement orale,

— faible quantité de corpus disponible qu'ils soient monolingues, bilingues, écrits ou oraux,

— faible présence sur Internet comparativement aux langues « principales »,

— moyenne d'âge élevée des locuteurs.

2.2 L'ALSACIEN

L'alsacien appartient au groupe hétérogène (plus que celui des langues latines ou slaves (Malherbe, 1983)) des langues germaniques.

1. Voir `http://locongres.org/images/docs/feuille_route_numerique_occitan_fr.pdf`.
2. Voir `https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/corpusCherchePub.php`.

On compte par ailleurs cinq aires linguistiques principales en Alsace décrites comme suit par Steible (2014) :
« *- Le francique rhénan, parlé en Alsace Bossue (région de Sarre-Union, La Petite Pierre) et dans une partie de la Moselle contigüe. Proche voisin des autres parlers dialectaux de Moselle ou du Luxembourg,*
- Le francique rhénan méridional, parlé dans l'extrême nord-est de l'Alsace (région de Wissembourg, Lauterbourg). Proche voisin des dialectes du Palatinat ou de Hesse,
- Le bas-alémanique du nord, parlé dans les régions de Saverne, Haguenau, Strasbourg et Sélestat. Proche voisin des dialectes du Bade-Wurtemberg,
- Le bas-alémanique du sud, parlé dans les régions de Colmar et de Mulhouse. Proche voisin des dialectes parlés en Brisgau (Bade-Wurtemberg),
- Le haut-alémanique, parlé au sud de la région d'Altkirch, c'est-à-dire dans le Sundgau (extrême sud de l'Alsace). Proche voisin du suisse allemand. »

Ces aires linguistiques forment le continuum dialectal, c'est à dire l'altération, phonétique et lexicale (Bernhard et Ligozat, 2013) s'opérant de proche en proche dans la région. Ce continuum est illustré par la figure 2.2.1.

L'INSEE estime en 2002 à environ 500 000 le nombre de locuteurs de l'alsacien résidant en Alsace [3]. En 2012, EdInstitut mène pour l'OLCA (Office pour la Langue et la Culture d'Alsace) une enquête [4] selon laquelle 43 % des alsaciens se déclareraient dialectophones (33 % des sondés déclarent savoir parler un peu ou comprendre un peu). 74 % d'entre eux auraient 60 ans et plus [5].

Si 62 % des locuteurs en 2012 affirment avoir transmis l'alsacien, il est clair que la transmission par le biais familial a chuté au cours du dernier siècle, alors même que d'après l'enquête de 2012, il s'agit du vecteur principal de son apprentissage (95 % des locuteurs déclarant alors avoir appris l'alsacien par le biais de leur famille).

3. INSEE, Chiffres pour l'Alsace, revue n° 12, décembre 2002.

4. Enquête menée sur la base de 801 personnes résidant en Alsace interrogées par téléphone selon la méthode des quotas entre le 1er et le 9 mars 2012.

5. Cette enquête semble surévaluer le nombre de locuteurs par rapport à l'enquête de l'INSEE menée en 1999 qui estimait alors que 39 % de la population était dialectophone

FIGURE 2.2.1 – Atlas historique d'Alsace, CRESAT, Université de Haute-Alsace.

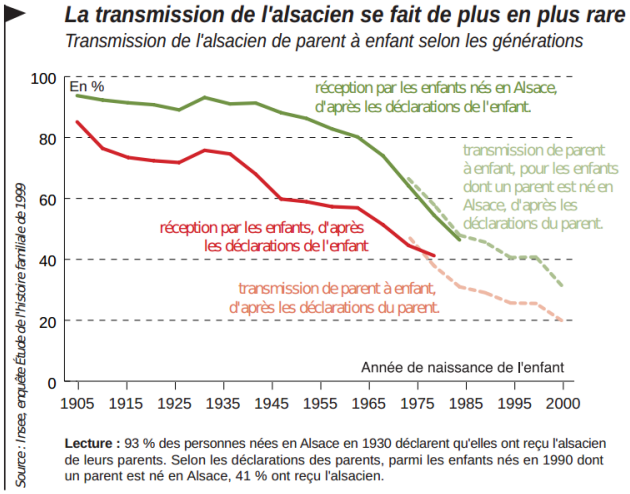

FIGURE 2.2.2 – Évolution de la transmission de l'alsacien (source : INSEE, 1999).

Néanmoins, des initiatives associatives proposant des cours d'alsacien sont nées, et la possibilité d'apprendre l'alsacien est donnée à l'université de Strasbourg. Par ailleurs, toujours d'après l'enquête de 2012, parler l'alsacien serait un facteur facilitant l'apprentissage d'autres langues et serait un atout professionnel pour la majorité des jeunes interrogés entre 18 et 29 ans.

Pour 38 % des personnes interrogées, l'action principale à mener est de donner une place à l'alsacien dans le milieu scolaire, afin de pallier la chute de la transmission familiale.

En 2008, la méthode ORTHAL est conçue par Danielle Crévenat-Werner et Edgar Zeidler (Crévenat-Werner et Zeidler, 2008) afin de fixer une graphie et de faciliter l'écriture de l'alsacien dans le continuum dialectal qu'il représente. L'orthographe respecte toutes les variantes tout en fixant des règles d'inter-compréhension qui permettent à chacun d'écrire à sa manière tout en étant lisible du nord au sud de l'Alsace. Il n'a pas été possible de mesurer l'impact de l'introduction de cette méthode.

2.3 Langues et présence sur internet

Les graphiques de la figure 2.3.1, illustrant respectivement la langue maternelle estimée des internautes, et la langue effective des contenus du Web [6], montrent qu'il existe une marge importante de progression pour qu'Internet soit représentatif de la réalité linguistique connectée, l'anglais dominant aujourd'hui largement la toile.

Par ailleurs, la couverture du réseau Internet croît et des zones sont nouvellement connectées. Cette extension donne l'opportunité d'enrichir la pluralité linguistique du Web, si celle-ci est rendue possible.

D'après l'étude de 2014 effectuée par Maaya, réseau mondial pour la diversité linguistique [7] l'alsacien est néanmoins très bien représenté sur Internet, notamment comparativement aux autres langues de France. La population alsacienne est effectivement bien connectée à Internet (80% des locuteurs ayant accès à Internet), et très active : on trouve un grand nombre de pages personnelles et groupes Facebook dont les contenus sont au moins partiellement en alsacien (25 000 usagers de Facebook déclarent parler cette langue [7]).

6. Source : Wikipedia - Langues utilisées sur Internet : `https://en.wikipedia.org/wiki/Languages_used_on_the_Internet`.

7. Voir *Étude sur la place des langues de France dans l'Internet*, `http://maaya.org`.

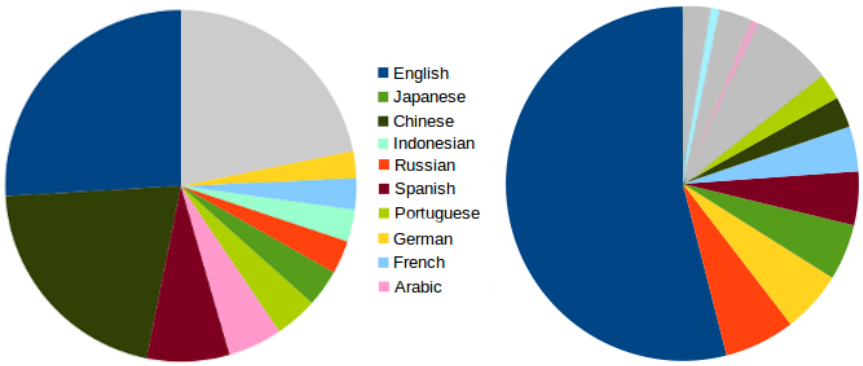

FIGURE 2.3.1 – Comparaison des langues des internautes (gauche) avec les langues des contenus sur Internet (droite).

Par ailleurs, l'OLCA [8] (Office pour la Langue et Culture d'Alsace) recense un nombre important de contenus écrits et oraux libres de droits en alsacien : recettes, poèmes issus du concours de poésie annuel *Friehjohr fer unseri Sproch*. Enfin, une Wikipédia [9] partagée entre alsacien, suisse-allemand, badois et souabe est un réservoir imparfait de contenus écrits en parlers régionaux (environ 18 000 articles).

2.4 LANGUES PEU DOTÉES

2.4.1 LE « CASSE-TÊTE » DES LANGUES PEU DOTÉES POUR LE TAL

Le terme de « langue peu dotée » se définit actuellement en négatif comme l'immense majorité des langues n'appartenant pas au groupe restreint des langues « bien dotées ». Une multitude d'appellations différentes a été trouvées dans la littérature, traduisant à la fois la diversité des réalités représentées et l'absence de définition faisant référence et de consensus sur le sujet.

La formulation anglaise la plus répandue semble être « low resource languages » (729 résultats obtenus lors d'une recherche sur Google Scholar) On rencontre également les appellations suivantes (est donné entre parenthèses le nombre d'articles rencontrés sur Google Scholar la

8. Voir `http://www.olcalsace.org/`.
9. Voir `https://als.wikipedia.org/wiki/Wikipedia:Houptsyte`.

mentionnant) : « low-resourced languages » (146) (Cucu et al., 2012), « under resourced languages » (Bauman and Pierrehumbert 2014), « resource free language » (Klementiev et Roth, 2006), « resource disadvantaged languages » (Sing et al., 2006), « non-resourced language » (Meftouh et al., 2012), « low-density language » (430) (Resnik et Smith, 2003), « lower-density languages » (Maxwell et Hugues, 2006), « resource-poor language » (Das et Petrov, 2011) (Nakov et Ng, 2012), « resource-scarse languages » (Hasan et Ng, 2009), « less-resourced languages » (Walter and Sagot, 2010 (Sorani Kurdish)), « little equipped » (Berment, 2004), « less-represented » (44) (Pellegrini, 2008), « Limited Resources Scenario » (Khanam et Murthy, 2014), « scarce resource » (Khanam et al., 2013) , « Poor Resource Scenario » (Dandapat et al., 2007), « langue non-dotée » (Scherrer et Sagot, 2013).

Chaque appellation moins répandue semble être utilisée par un microcosme de chercheurs travaillant ensemble et traitant d'une langue particulière ou d'un groupe de langues restreint. La multiplicité de ces appellations paraît en outre être due à l'inexistence d'une plate-forme ou d'une méthodologie unique fédérant les travaux menés.

Il n'existe pas de seuil permettant de fixer une langue dans l'une ou l'autre de ces catégories, les avancées technologiques pouvant aussi bien accroître la complexité des ressources langagières annotées nécessaires, qu'en limiter la dépendance (dans le cas des technologies non supervisées par exemple).

Néanmoins, si on prend comme critère les langues pour lesquelles il existe des corpus arborés (*treebanks*) de qualité ainsi que des outils performants pour l'annotation en parties du discours, on peut considérer que la majorité des langues officielles de l'union européenne, ainsi que l'anglais américain, l'arabe moderne, le chinois mandarin, le coréen, le japonais, le russe sont effectivement « bien dotées » (voir (Petrov *et al.*, 2012), décrivant la création d'un « Universal POS Tagset », jeu d'étiquettes pour l'annotation en partie du discours universel, pour 22 langues).

L'appellation « langue peu dotées » recouvre donc un ensemble de réalités linguistiques très varié :

- Statut :
  - — les langues ayant le statut de langue officielle parlées par un nombre réduit de locuteurs (exemple : l'islandais, 310 000 locuteurs),
  - — les langues ayant le statut de langue officielle parlées par un nombre important de locuteurs (exemple : le lao, 4 à 5 millions de locuteurs au Laos),

  - les langues n'ayant pas le statut de langue officielle coexistant avec une langue bien dotée étant parlées par un nombre important de locuteurs (exemple : l'igbo, parlé par plus de 20 millions de personnes au Nigéria dont la langue officielle est l'anglais).

- Parenté avec d'autres langues :
  - les langues apparentées à une langue bien dotée (exemple : l'occitan lanquedocien, étymologiquement proche du catalan (Vergez-Couret et Urieli, 2015).),
  - les langues appartenant à une famille de langues dont certaines sont bien dotées (exemple : l'alsacien au sein de langues germaniques, pouvant profiter des travaux existant sur l'anglais et l'allemand) ,
  - les langues « isolées » au sein d'un groupe (exemples :l'arménien, le grec appartenant au groupe indo-européen).
- Géographie :
  - les langues de peu de locuteurs géographiquement isolées (exemple : l'inuktitut (34 000 locuteurs), qui n'est pas soumis à la pression d'une autre langue),
  - les langues dont les locuteurs appartiennent à plusieurs pays (exemple : le berbère (Maroc, Algérie, Mali, Niger)),
  - les langues dont les locuteurs sont peu groupés (exemples : le breton, le berbère, dont les locuteurs subsistent dans des zones rurales où elles ont été confinés par les politiques linguistiques),
  - les langues cohabitant avec une langue « nationale » ou langue-toit (*Dachsprache*) comme les langues indonésiennes au regard du bahasa indonesia (seule enseignée et bénéficiant d'une littérature écrite) ou les parlers régionaux en France.
- Graphie :
  - les langues à pure tradition orale,
  - les langues aux graphies multiples (exemple : alsacien),
  - les langues à graphies partagées (exemple : les langues latines), non partagées (exemples : le chinois, l'hébreu).
- Intérêt :
  - les langues présentant ou ne présentant pas (encore ?) d'intérêt économique global.

On note par ailleurs qu'une langue à tradition orale sans orthographe ni grammaire consensuelle requiert un effort bien plus élevé pour mettre en place des outils d'acquisition de ressources linguistiques, qu'une langue disposant d'une graphie installée et étant enseignée *via* des manuels de référence. La distinction entre « langue peu dotée » et « langue bien dotée » peut et doit donc être affinée, à la manière de Vincent Berment dans sa thèse (Berment, 2004). Il classifie ces langues sous trois appellations : langues-π (π pour peu ou pas dotées), par opposition aux langues-τ (τ pour très bien dotées) et aux langues-μ (μ pour moyennement dotées) et définit un tableau évaluant le niveau d'informatisation d'une langue : pour un certain nombre de services ou ressources sont attribuées une note de criticité et une note de qualité de l'existant.

Ainsi, sont établies par exemple les notes de 5,46/10 pour le birman et de 6,14/10 pour le khmer. Le terme « langues-μ » pour « moyennement dotée » a été repris par Maxwell en 2006 (Maxwell et Hughes, 2006) sous la dénomination de *medium-density languages*.

L'absence de définition claire de critères fins pour qualifier la richesse en ressource des langues, ainsi que d'une plate-forme unique fédérant les ressources pour le TAL rend ardue la tâche de classification des langues selon ces degrés.

Plusieurs initiatives, à l'instar de l'OLAC (Open Archives Language Community [10], qui propose un moteur de recherche pour les resources linguistiques englobant 58 catalogues tels que ELRA (European Language Resources Association) Catalogue of Language Resources [11], ou le projet An Crúbadán [12] (qui exploite la grande quantité de ressources textuelles en accès libre sur le Web pour constituer des corpus, notamment pour les langues peu dotées) tentent de remédier à ce manque mais nous n'avons pu trouver à ce jour aucun document faisant un état complet et à jour de l'existant en ressources spécifiques au TAL pour chaque langue.

En 2003, ELSNET (European Network of Excellence in Language and Speech) et ELRA créent le Basic Language Resource Kit (BLARK), dont le but est de définir pour chaque langue un tableau recensant les ressources, outils et compétences nécessaires au développement d'outils de TAL (Maegaard *et al.*, 2006). L'élaboration du BLARK consiste à remplir pour chaque langue deux matrices évaluant respectivement :

10. Voir `http://search.language-archives.org/index.html`.
11. Voir `http://catalog.elra.info/index.php`.
12. Voir `http://crubadan.org/`.

— l'importance de chaque type de ressource langagière pour la construction de composants logiciels spécifiques (exemple d'entrée : lexique monolingue pour le développement d'un analyseur sémantique),

— l'importance de chaque brique logicielle dans le développement d'applications (exemple d'entrée : importance d'avoir un système de reconnaissance des entités nommées pour le développement d'un système de dialogue).

Les ressources orales et écrites sont traitées dans des matrices séparées.

Les tables BLARK de l'arabe ont été établies au cours du projet NEMLAR [Maegaard, 2004] ainsi que celles du néerlandais (Strik *et al.*, 2002).

Néanmoins, cette méthodologie ne peut s'appliquer en l'état aux langues peu dotées, pour lesquelles aucune ressource ou presque n'est disponible et une critique en a été faite par Prys (2006) qui demande un RELARK (Preliminary Language Resources Kit) pour les langues « critiquement peu dotées » . Il s'agirait d'une couche antérieure au BLARK permettant d'identifier les tout premiers composants nécessaires pour permettre l'entrée d'une langue au sein des technologies du langage.

2.4.2 Méthodologie : myriadisation et jeux ayant un but

L'efficacité de la myriadisation dans le domaine de la collecte de données pour le TAL n'est plus à prouver. On peut notamment citer, pour les jeux ayant un but, la plate-forme JeuxDeMots (réseau lexical de 835 000 termes connectés par plus de 42 millions de relations lexicales et sémantiques à ce jour [13]), ou l'application Phrase Detective (annotation de corpus en référence de plus de 2,5 milions d'annotations produites entre 2008 et 2012 (Poesio *et al.*, 2012)).

2.4.2.1 Typologie de la myriadisation

La myrisadisation peut prendre trois forme principales : la participation rémunérée et directe, bénévole et directe, bénévole et indirecte.

13. Voir `http://www.jeuxdemots.org/jdm-about.php` pour les statistiques actualisées.

Dans le premier cas, on peut citer l'exemple d'Amazon Mechanical Turk. Amazon met en effet en place en 2005 une plate-forme offrant la possibilité de mettre en relation toute personne ayant une tâche à faire réaliser par l'intelligence humaine (« Human Intelligence Tasks », dont des tâches de traduction, de reconnaissance de contenu d'image, par exemple), avec une communauté de travailleurs (les « turkers »). Cette plate-forme pose des problèmes éthiques du fait du système de micro-rémunération mis en place : notamment, il n'existe pas de salaire minimum ni de garantie pour les « turkers » d'être effectivement rémunérés pour le travail fourni (Fort *et al.*, 2011). Par ailleurs, la plate-forme AMT ne ne permet pas de former le contributeur à la tâche spécifique proposée et paraît inadaptée aux tâches d'annotation dans le cadre des langues peu dotées car la communauté concernée par la tâche proposée peut être réduite et difficile à identifier.

Les projets Wikipedia et le Projet Gutenberg illustrent quant à eux le principe de participation bénévole et directe, c'est-à-dire pour laquelle le but à atteindre est connu de l'utilisateur.

Une troisième alternative est la participation bénévole et indirecte. C'est celle des jeux ayant un but, pour lesquels l'aspect ludique de la plate-forme proposée masque à l'utilisateur le but réel à atteindre et compense la dimension pouvant être fastidieuse ou complexe de la tâche à réaliser. Cette méthodologie ouvre des champs d'investigation multiples pour la communauté du traitement automatique des langues qu'elle soient bien ou peu dotées. Ainsi, (Jurgens et Navigli, 2014) explorent la possibilité d'aller plus loin dans la démarche du jeu ayant un but concernant la tâche de désambiguïsation lexicale pour l'anglais et l'italien en développant de véritables jeux-vidéo autour de la tâche plutôt qu'une interface ludifiée très proche de la tâche d'annotation traditionnelle.

Cette dernière alternative est celle vers laquelle nous souhaitons tendre, bien que très peu d'éléments ludiques aient à ce jour été intégrés à l'application. Dans la classification de Good et Su (2013) pour les jeux ayant un but, nous nous situons donc aujourd'hui dans la dynamique du « *crowdsourcing* a des fins scientifiques pour une micro-tâche » (sans résolution d'un problème complexe).

2.4.2.2 Myriadisation et enjeux de la ludification (*gamification*)

Il existe des faiblesses inhérentes aux jeux ayant un but telle que la tentation de certains joueurs de contourner les système ou de tricher pour obtenir plus de points, par conséquent « trouver le degré maximal de « ludification » est crucial au regard de la motivation suscitée chez le

joueur. »[14] (Lee *et al.*, 2013). Néanmoins, dans le cas qui nous occupe, ce phénomène paraît négligeable, d'autant plus que le degré de ludicité mis en place est quasi nul dans ce projet.

Le projet réalisé se situe dans le cadre de la participation bénévole et directe : en effet, la recherche du divertissement à été négligée dans ce projet, au profit de la volonté de participer à un projet de recherche (science citoyenne) visant à promouvoir les parlers alsacien. Les éléments de ludicité classiques (score, possibilité de collectioner des trophées, grades, évolutivité etc.) ont été très peu développés.

Dans notre cas, la difficulté majeure a donc consisté à définir et à atteindre la communauté de locuteurs disposant des connaissances nécessaires à l'annotation – ou pouvant facilement les acquérir grâce à la formation proposée – et motivée pour promouvoir sa langue en contribuant *via* l'application proposée.

14. « *Finding the maximum safe level of gamification is crucial for motivational design* » (Lee *et al.*, 2013).

CHAPITRE 3

BISAME, UNE APPLICATION POUR RECUEILLIR DES ANNOTATIONS

3.1 ACCÈS À L'APPLICATION ET CHOIX TECHNIQUES

L'application développée est accessible à l'adresse `http://bisame.herokuapp.com`. Le code source peut être téléchargé *via* le dépôt Git mis en place à l'adresse `https://github.com/milloura/bisame`

Il s'est agi dans le cadre de ce projet de développer intégralement un site Internet répondant aux spécifications définies.

Compte tenu du peu de compétences préalables spécifiques aux technologies du Web, nous avons choisi des technologies courantes, pour lesquelles il existe une documentation riche et une communauté assurant le support des développeurs.

La plate-forme BISAME a ainsi été développé en PHP avec le framework Laravel. Ces choix nous ont permis de développer en un temps court une application robuste, sécurisée, facile à maintenir et portable. Nous avons pu ainsi développer toutes les fonctionnalités nécessaires et proposer en parallèle un design qui semble avoir été apprécié par la communauté des utilisateurs.

La base de données a été construite grâce à la technologie MySQL, et le schéma de la base est le suivant :

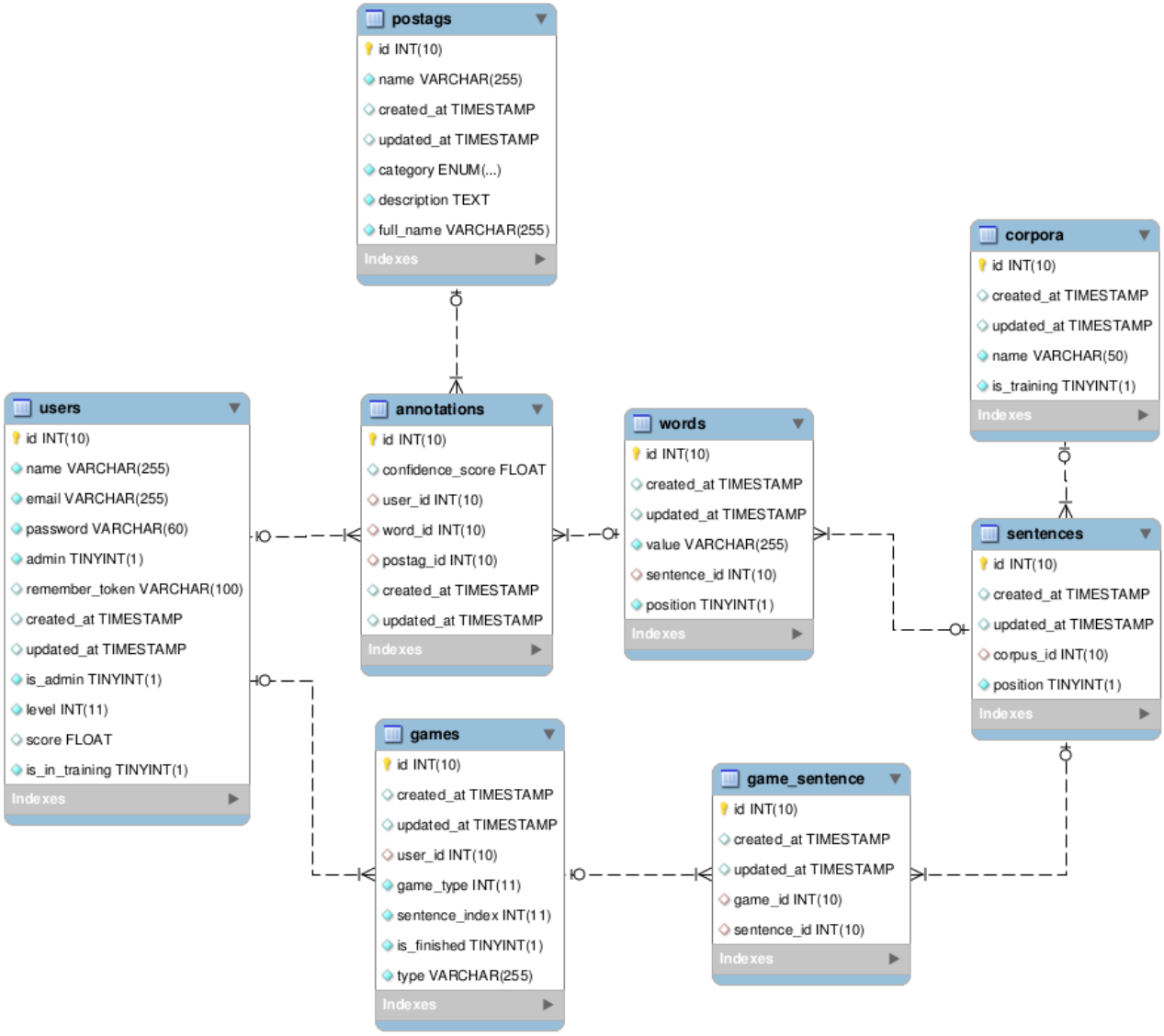

FIGURE 3.1.1 – Schéma de la base de données développée pour l'application BISAME.

Un utilisateur produit ainsi des `annotations`, soit un lien entre un `mot` et une `étiquette` au cours de `jeux`. Ces `jeux` peuvent être de type `entrainement` ou `production`. Les `jeux` sont alimentés par des `phrases` contenant des `mots`, issues de `corpora` pouvant être de type `entraînement` ou non.

3.2 Choix méthodologiques

3.2.1 Description générale

La méthodologie générale adoptée dans le cadre de cette expérience pour l'obtention de corpus annotés et le développement d'outils d'annotation de qualité est illustrée par le schéma suivant :

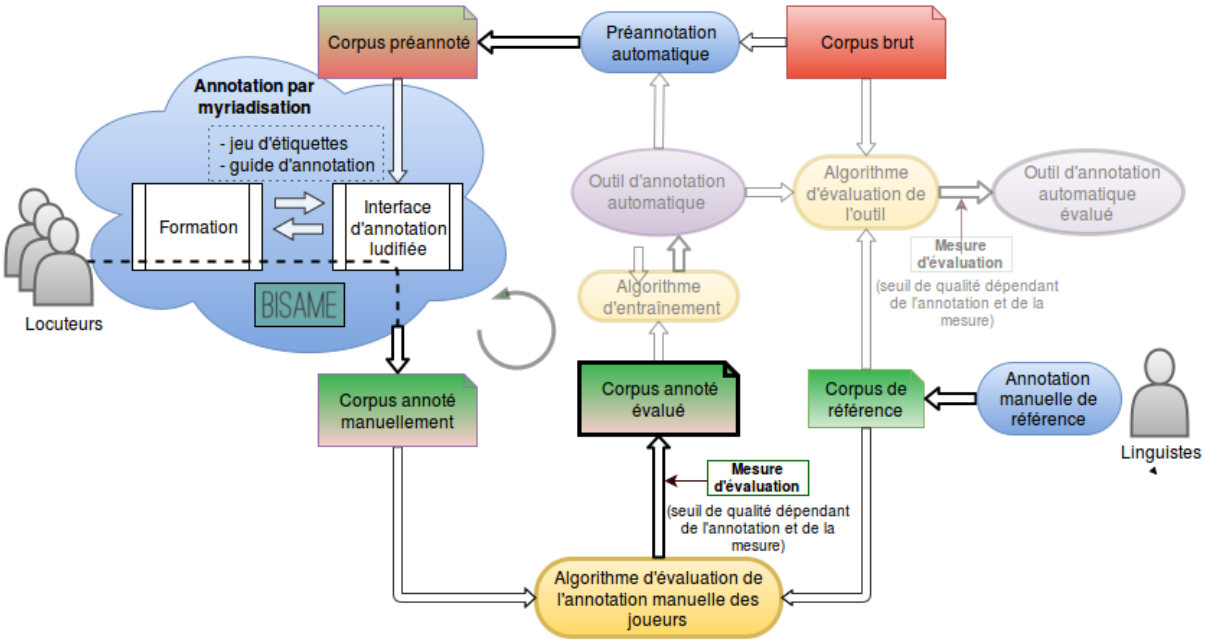

FIGURE 3.2.1 – Méthodologie de collecte des données pour la linguistique outillée par myriadisation. Le corpus annoté obtenu apparaît entouré en gras.

Nous identifions plusieurs étapes :

1. Préparation des données :
   1.1. constitution d'un corpus brut représentant au mieux l'état de la langue,
   1.2. constitution d'un corpus de référence annoté par des linguistes,
   1.3. pré-annotation de ces corpus grâce aux outils imparfaits d'annotation existants,
   1.4. identification des points difficiles de la tâche d'annotation à réaliser et rédaction du guide d'annotation,

2. Mise en place de la plate-forme :
   2.1. alimentation de l'application en corpus,
   2.2. conception de la phase de formation relativement au points durs identifiés,
3. Collecte des annotations :
   3.1. mise en ligne de la plate-forme,
   3.2. communication publique sur la disponibilité de la plate-forme et l'intérêt des contributions individuelles,
   3.3. assistance aux utilisateurs,
4. Analyse des résultats obtenus :
   4.1. normalisation des annotations produites par les joueurs,
   4.2. comparaison des résultats obtenus par les joueurs sur certaines phrases de référence et évaluation de leurs performances,
   4.3. construction d'une annotation unique issue de la mise en commun des annotations fournies par les différents utilisateurs
5. Évaluation des résultats.

Les parties grisées du schéma représentent les étapes qu'il faudrait mettre en place pour pouvoir entraîner et évaluer des outils d'annotation grâce aux corpus annotés obtenus.

3.2.2 Tâche spécifique : l'annotation en parties du discours

L'annotation en parties du discours telle qu'elle est réalisée par les utilisateurs est relativement peu complexe, du fait de la faible taille du jeu d'étiquettes considéré. Les dimensions de complexité donnés sur la figure suivante ont été déterminés *a priori* selon la grille d'analyse proposée par Fort *et al.* (2012) :

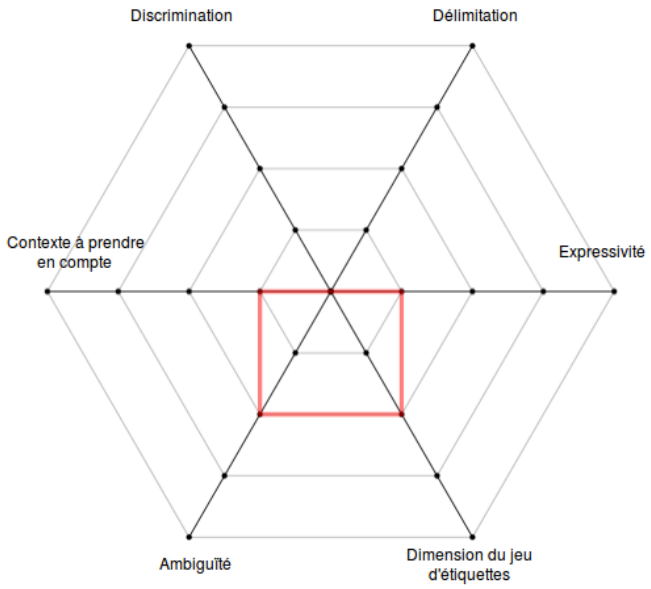

FIGURE 3.2.2 – Dimensions de complexité de la tâche d'annotation en parties du discours.

Ainsi, la délimitation et la discrimination sont nulles, les phrase étant préalablement séparées en tokens et tous les mots pouvant être annotés. L'expressivité du langage d'annotation est faible car il s'agit d'un langage de types, la dimension du jeu d'étiquettes est relativement faible (le travail effectué grâce à la pré-annotation a permis de faire baisser ce degré de complexité en ne proposant qu'un sous ensemble des étiquettes à chaque annotation).

Nous avons défini que l'ambiguïté était également faible mais non nulle, bien que l'annotation soit *a priori* non ambiguë. Cela traduit la méconnaissance potentielle des subtilités des étiquettes des utilisateurs non grammairiens (différence entre `CONJ` et `SCONJ` par exemple). Enfin, le contexte à prendre en compte est également faible car restreint à la phrase contenant le token à annoter. La seule source externe que l'utilisateur doit consulter est le guide d'annotation, intégré sous la forme des aides-mémoire.

3.2.3 CHOIX DU JEU D'ÉTIQUETTES

Nous avons choisi d'utiliser dans le cadre de ce projet les « *Universal POS tags* » (Petrov *et al.*, 2012) [1] afin de pouvoir tirer parti facilement des données fournies par l'équipe du LiLPa

1. Voir `http://universaldependencies.org/u/pos/all.html`

travaillant déjà sur l'alsacien avec ce jeu d'étiquettes. L'« *Universal POS tagset* », synthétisant les des jeux d'étiquettes spécifiques de 22 langues, est facilement extensible aux spécificités de chaque langue considérée tout en étant standard, reconnu et soutenu par la communauté du TAL.

Ce jeu d'étiquettes grossier comprenant 17 catégories n'a finalement pas été adapté à l'alsacien faute de temps. Seule la catégorie `X` a été remplacée par la catégorie `Mots en langue étrangère` afin de limiter le risque d'obtenir des données difficiles à analyser, l'étiquette `autre` étant trop vague. Il est apparu que l'annotation des `Mots en langue étrangère` a posé des difficultés. En effet, il est courant (notamment dans la pièce de théâtre appartenant au corpus) que des termes en allemand et en français soient intégrés dans une phrase en alsacien. Les utilisateurs ont étés gênés par le fait de qualifier ces mots comme appartenant à une langue étrangère, ce qui a été une source de désaccords lors de la comparaison avec la référence dont nous disposions.

Les modifications suivantes ont également été considérées :

— introduction de la catégorie `VERB:AUX_MOD` désignant les auxiliaires modaux,

— introduction d'une catégorie `je ne sais pas` permettant éventuellement de donner une signification à l'absence d'annotation parfois observée pour un mot de la phrase.

Par ailleurs, la tokénisation imparfaite a conduit à des problèmes de choix d'étiquette : Par exemple dans le cas de la contraction de la préposition et de l'article défini « vum », celui-ci n'est ni préposition ni déterminant, puisqu'il est l'amalgame de l'un et de l'autre.

3.2.4 Utilisation de la pré-annotation

Un des enjeux propres aux tâches linguistiques à accomplir est la réduction de la complexité de la tâche d'annotation (Fort *et al.*, 2012). Cette réduction passe notamment par la qualité de la pré-annotation, et l'intégration de celle-ci. La pré-annotation dont nous disposions pour le corpus utilisé dans l'application permet par exemple de proposer un nombre réduit d'étiquettes lors de l'annotation diminuant ainsi la complexité du jeu d'étiquettes.

Fort et Sagot (2010) ont en effet montré qu'un corpus de 50 phrases peut suffire à entraîner un outil de pré-annotation permettant de réduire notablement (de l'ordre le la division par trois) le temps d'annotation humain tout en conservant la qualité (notamment l'exactitude des annotations produites et les accords inter-annotateurs) de cette annotation manuelle. Nous avons

ainsi utilisé les travaux de Bernhard et Ligozat (2013), ayant développé deux outils d'annotation automatique présentant une exactitude de l'ordre de 70%. La pré-annotation peut dans certains cas introduire un biais si l'utilisateur n'est pas disposé à remettre en cause l'étiquette qui lui est suggérée. C'est pourquoi dans le cadre de cette application, nous proposons systématiquement deux étiquettes, même s'il existe une pré-annotation consensuelle. Cela force l'utilisateur à devoir rejeter systématiquement une des propositions qui lui est faite.

3.2.5 Identification de la communauté

Un des enjeux majeurs du succès d'une application fonctionnant grâce à la myriadisation (*crowdsourcing*) est de parvenir à réunir une communauté acceptant d'accorder du temps et de participer à créer des données. Dans notre cas, l'OLCA a fait une communication le 2 mai, mais cela n'a pas suffi à mobiliser les utilisateurs ayant en effet contribué. En effet, il semblerait que les utilisateurs soient répartis dans des micro-communautés qu'il est difficile d'atteindre grâce à une seule communication.

Aussi, le relais effectué par des membres de diverses associations comme Le FILAL [2] (Fonds international pour la lange alsacienne), ou d'entreprises telles que la Marque Alsace [3], ainsi que le contact individuel *via* Facebook grâce à l'identification de groupes de dialectophones a permis de réunir plus d'utilisateurs. Il est également possible que le cumul des diffusions ait contribué à convaincre des utilisateurs de se rendre sur la plate-forme.

3.2.6 Évaluation des annotations

3.2.6.1 Scores de confiance

Les utilisateurs et les annotations présentent un score de confiance. Celui de l'utilisateur est mis à jour au cours de son utilisation de la plate-forme, celui de l'annotation est égal au score de confiance de l'utilisateur au moment de son annotation.

L'évolution du score de confiance des utilisateurs n'a pas été évaluée dans ce projet, nous n'avons donc pas pu mesurer leur progression.

2. Voir `https://filalsace.net/about/`.
3. Voir `http://www.marque-alsace.fr/`.

3.2.6.2 Accord inter-annotateur et coefficient Kappa

En raison de la faible quantité de données annotée parallèlement par plusieurs annotateurs différents, nous n'avons pas pu calculer d'accord inter-annotateurs entre les utilisateurs. Celui-ci aurait pu nous permettre de mesurer la clarté des consignes données et l'intelligibilité de la tâche. Néanmoins, nous avons calculé l'accord entre les annotations produites par les utilisateurs et la référence qui nous a été fournie par l'équipe du LiLPa. En effet, cette annotation dite de « référence » comporte des erreurs au regard du guide d'annotation, celui-ci ayant été rédigé *a posteriori* par les deux expertes linguistes. La « référence » doit donc encore être corrigée afin d'être cohérente avec le guide. Ce scénario s'apparente à une campagne de pré-annotation, telle que décrite par Fort (2012).

L'accord entre utilisateur et référence est calculé grâce au coefficient Kappa de Cohen définit par Cohen (1960). Au-delà de l'accord observé, ce coefficient permet de prendre en compte qu'une partie de l'accord est due au hasard, et que l'utilisateur peut être biaisé. En effet, les utilisateurs peuvent avoir différentes interprétations des consignes et annoter selon une tendance biaisée qui leur est propre. On considère donc la proportion d'accord atteinte se trouvant au-dessus du hasard.

Sont ainsi calculés :

— l'accord observé A_0, soit la proportion de « bonnes réponses » de l'utilisateur,

— l'accord attendu A_e^k, soit la valeur attendue de l'accord observé :
$A_e^k \sum_q \frac{n_{A1q}}{i} . \frac{n_{A2q}}{i}$ pour $\begin{cases} i = \text{nombre total d'items : 16} \\ n_{Axqa}, \text{la probabilité que l'utilisateur } A_x \text{choisisse la catégorie } q_a \end{cases}$

3.2.6.3 Définition d'un utilisateur de confiance

Afin de pouvoir comparer l'annotation fournie par les utilisateurs et la référence, une annotation unique est déduite des données recueillies. En cas de désaccord, nous décidons quelle étiquette est la plus fiable en calculant quelle étiquette parmi celles proposées présente la probabilité maximale d'être correcte, au regard des scores de confiance qui ont été attribués aux annotations.

Supposons que pour un mot, les deux étiquettes E1 et E2 aient été choisies par trois utilisateurs U1, U2 et U3, dont les scores de confiance sont ceux donnés dans le tableau 3.2.1.

Étiquette	Utilisateur	Score de confiance
E1	U1	0.8
E2	U2	0.6
E2	U3	0.4

TABLEAU 3.2.1 – Calcul de l'étiquette la plus probable (étape 1).

En calculant les probabilités jointes d'erreur pour les événements indépendants que constituent les annotations, on obtient le score de confiance final de chaque étiquette (tableau 3.2.2).

Étiquette	Probabilité jointe d'erreur	Score de confiance de l'étiquette
E1	0.2	0.8
E2	0.24	0.76

TABLEAU 3.2.2 – Calcul de l'étiquette la plus probable (étape 2).

Pour le cas de figure décrit, on choisira donc l'étiquette E1.

3.3 RESSOURCES DE DÉPART

3.3.1 LE CORPUS

Nous avons choisi de travailler uniquement sur des corpus pour lesquels nous disposions d'une référence afin de pouvoir évaluer effectivement les résultats obtenus.

L'équipe du LiLPa de Strasbourg nous a ainsi fourni un corpus constitué de quatre textes dont la description est donnée dans le tableau 3.3.1.

Ce corpus traduit une certaine réalité de l'écriture de l'alsacien mais demeure très imparfait. Il manque notamment de textes issus des réseaux sociaux comme Twitter ou Facebook qui

4. Disponible à l'adresse `https://archive.org/details/drhoflieferantel00stos`.
5. Disponible à l'adresse `https://www.olcalsace.org/fr/autres-publications`.
6. Disponible à l'adresse `https://als.wikipedia.org/wiki/Els%C3%A4ssisches_Museum_%28Stra%C3%9Fburg%29`.
7. Disponible à l'adresse `https://als.wikipedia.org/wiki/Johannes_Mentelin`.

Nom	Description	Nombre de phrases	Nombre de mots
Hoflieferant_p53 [4]	pièce de théâtre D'r Hoflieferant de Gustave Stoskopf	26	181
recettes [5]	texte de trois recettes	29	311
wikipedia1 [6]	page Wikipedia sur le musée alsacien de Strasbourg	20	345
wikipedia2 [7]	page Wikipedia sur Johannes Mentelin	27	425
TOTAL		102	1 262

TABLEAU 3.3.1 – Description du corpus

est apparue comme une source importante de contenu écrit en alsacien. Par ailleurs, aucun document journalistique ni littéraire (en particulier récit ou poésie) ne fait partie de ce corpus. C'est en partie dû au fait qu'il est difficile de trouver des contenus libre de droits exploitables pour des travaux de recherche, et cela constitue également un enjeu auquel nous nous sommes intéréssés.

3.3.2 ANNOTATIONS FOURNIES PAR L'ÉQUIPE DU LILPA

Deux expertes linguistes, Delphine Bernhard, maître de conférence à l'univesité de Strasbourg et Lucie Steiblé en post-doctorat à l'université de Strasbourg nous ont fourni les annotations suivantes pour chacun des textes du corpus :

(a) deux annotations manuelles indépendantes réalisées par chacune des expertes linguistes,

(b) une annotation unique de référence issue de l'arbitrage entre les deux annotations (a),

(c) les annotations produites par deux outils d'annotations automatiques : le German Tree-Tagger (Schmid, 1995) et le Stanford POS Tagger (Toutanova *et al.*, 2003), deux taggers de l'allemand ayant été entraînés pour l'alsacien grâce à la méthodologie décrite par Bernhard et Ligozat (2013).

3.3.2.1 Analyse des annotations manuelles concurrentes

Nous avons souhaité disposer de deux annotations manuelles effectuées indépendamment afin de pouvoir identifier les éventuelles catégories spécifiques portant à désaccord. L'accord inter-annotateur entre les deux expertes linguistes a ainsi été évalué pour chacun des documents du corpus. Les résultats des calculs des coefficients κ de Cohen (Cohen *et al.*, 2005) et π (Scott, 1955) sont présentés dans le tableau 3.3.2.

Corpus	Hoflieferant_p53	recettes	wikipedia1	wikipedia2
Accord observé	0.92	0.87	0.91	0.89
Coefficient κ de Cohen	0.91	0.85	0.90	0.88
Coefficient π	0.91	0.85	0.90	0.88

Tableau 3.3.2 – Accords inter-annotateur calculés pour les annotations manuelles fournies

Les coefficients calculés étant tous supérieurs à 0.8, on peut déduire que l'annotation manuelle fournie est de bonne qualité selon le critère défini par Artstein et Poesio (2008). On constate également que les coefficients κ de Cohen et π sont quasiment égaux, ce qui signifie que le biais des deux annotatrices est très faible.

Les matrices de confusion des deux annotations manuelles fournies sont disponibles en annexe 3.6.2.

Les désaccords portent en majorité sur les catégories suivantes :

— `ADJ` et `ADV` : par exemple *licht* (*légèrement*)

— `ADP` et `ADV` : par exemple *zamme* (*ensemble*)

— `ADV` et `CONJ` : par exemple *àwwer* (*mais*)

— `AUX` et `VERB` : par exemple *ìsch* (*est*)

Ces divergences, pouvant être dûes à une mauvaise compréhension de la langue ou à un contexte ambigu, ont été résolues dans l'adjudication fournie par les deux expertes. Cette adjudication, réalisée sur le corpus complet, nous sert de référence.

Nous avons par ailleurs fortement poussé à la rédaction d'un guide d'annotation contenant des exemples pour chaque étiquette et détaillant les cas pouvant porter à confusion. Celui-ci [8] nous

8. *Guide d'annotation morphosyntaxique pour les dialectes alsaciens*, Delphine Bernhard, Pascale Erhart, Dominique Huck, Lucie Steiblé, LiLPa, Université de Strasbourg

a ainsi également été fourni, les aides-mémoire pour chaque étiquette ayant été intégrés à la plate-forme s'en inspirent très largement.

3.3.2.2 Analyse des annotations produites par les outils automatiques

La comparaison des annotations fournies par les deux outils automatiques (German TreeTagger (Schmid, 1995) et Stanford POS Tagger (Toutanova *et al.*, 2003)) a permi de définir que la performance conjointe des deux outils sur le corpus de référence complet est de 90% : dans 10% des cas, aucun des deux taggers ne propose l'étiquette de référence donnée par l'adjudication des deux annotations manuelles.

3.3.3 Script pour la tokénisation

L'équipe du LiLPa nous à également fourni un outil de tokénisation simple définissant comme token :

— Les nombres,

— les abréviations,

— les URL et courriels,

— les « mots » compris entre deux séparateurs,

— les séparateurs.

Outre les signes de ponctuation classiques permettant de délimiter les tokens, (« ? », « « », « & » etc.), les séparateurs spécifiques de l'alsacien sont pris en compte. Par exemple, dans le contexte « *ich mach's* » (« *je le fais* »), « *mach* » (« *fais* ») est considéré comme un token car « *'s* » (pronom « *le* ») est considéré comme un séparateur. Les séparateurs sont par ailleurs considérés comme des tokens à part entière. Nous appellerons « mot » dans la suite de ce mémoire tout token n'étant pas un élément de ponctuation.

3.4 Description de l'application Bisame

Le but de l'application Bisame est donc de permettre la construction collaborative d'un corpus annoté en parties du discours pour l'alsacien. Il s'adresse à tout internaute dialectophone,

sans connaissance spécifique requise ; ce n'est en aucun cas une application réservée aux enseignants ou grammairiens.

L'application comporte deux phases : la phase de formation et la phase de production des annotations. Cette dernière n'est rendue accessible qu'une fois la première terminée.

3.4.1 Éléments communs aux deux phases

L'utilisateur a accès *via* la barre de navigation au nombre total d'annotations produites par les utilisateurs, à son `niveau`, son `score` et au nombre d'annotations qu'il a produites.
Le `niveau` varie entre 0 et 1 selon que la phase de formation a été complétée, et le `score` correspond au degré de confiance accordé à l'utilisateur multiplié par le nombre d'annotations qu'il a produites.

FIGURE 3.4.1 – Page d'accueil de la plate-forme.

Quelle que soit la phase du jeu, l'annotation se déroule de la manière suivante : l'utilisateur clique sur un mot, ce qui déclenche l'affichage d'un tableau de catégories grammaticales à droite.

Lorsqu'une étiquette est sélectionnée, elle apparaît sous le mot sélectionné et les mots restants peuvent alors être annotés dans n'importe quel ordre avant validation de la proposition. L'utilisateur peut modifier la catégorie associée à un mot à tout moment.

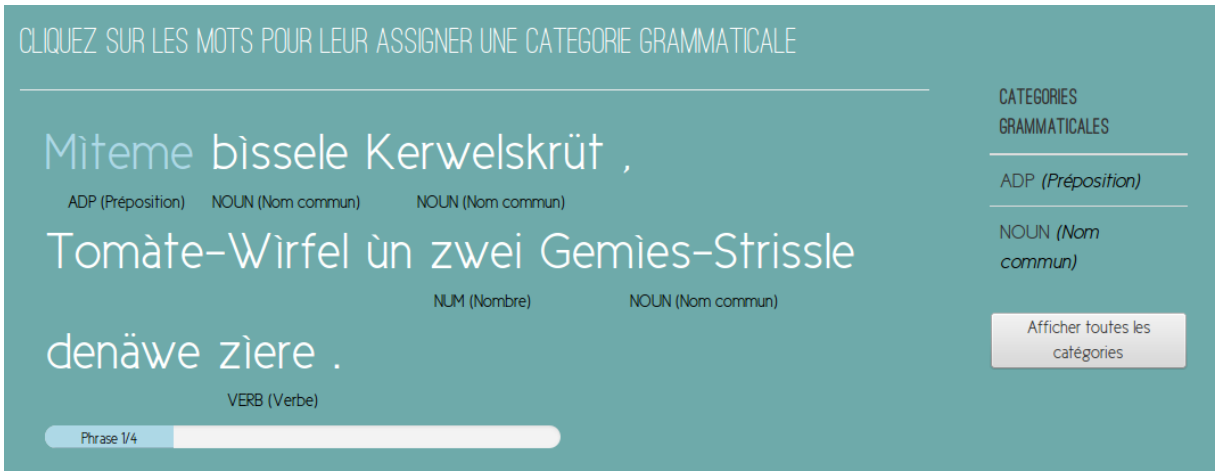

FIGURE 3.4.2 – Annotation de plusieurs mots : dans une phrase donnée, plusieurs mots peuvent être annotés avant de procéder à la validation de la proposition.

Lors de l'annotation d'un mot, deux propositions d'étiquettes sont faites à l'utilisateur : ce sont les deux étiquettes issues des pré-annotations effectuées par le TreeTagger et le StanfordTagger. Lorsqu'aucune des deux étiquettes proposées par les taggers n'est correcte, l'utilisateur doit alors faire afficher la liste complète des catégories.

Lorsque l'utilisateur passe sa souris sur une catégorie, un aide-mémoire s'affiche, constitué d'exemples et des cas problématiques ayant été identifiés. Dans la figure 3.4.3, on a fait apparaître l'aide mémoire de la catégorie `ADV` : adverbes. Des exemples de phrases contenant des tokens à annoter comme adverbes sont donnés, ainsi que des exemples spécifiques de tokens devant être annotés comme adverbes dans certains contextes : ici, on donne l'exemple d'une phrase où « wo » et « wenn » doivent effectivement être étiquetés comme adverbes. Une section « ATTENTION » donne les contre-exemples classiques pouvant générer des confusions : on donne ainsi le cas de « wenn », devant être annoté `SCONJ` lorsqu'il est employé dans le sens de « si », et de « wo » dans un contexte où il doit être annoté comme `PRON`.

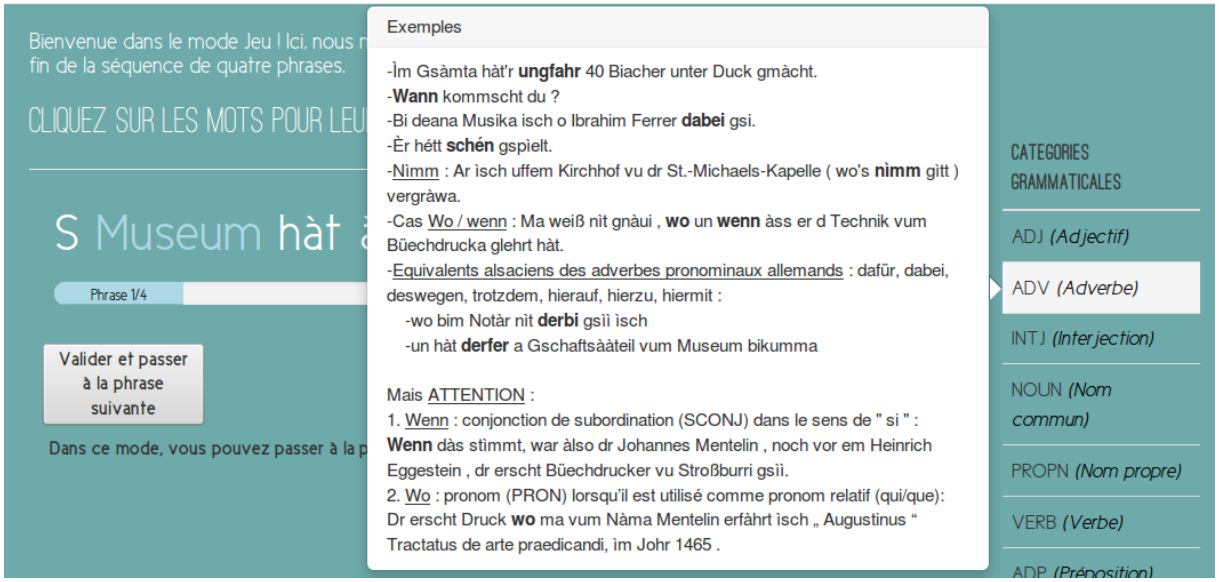

FIGURE 3.4.3 – Aide-mémoire pour la catégorie ADV.

Chaque séquence est constituée de quatre phrases.

3.4.2 PHASE DE FORMATION

Le corpus utilisé dans la phase de formation est le corpus `recettes`. Toutes les étiquettes y apparaissent au moins une fois.

Nous avons voulu tenir compte des points durs identifiés grâce à l'analyse des désaccords entre les annotations manuelles des deux chercheuses du LiLPa (en annexe 3.6.2). C'est pourquoi les propositions d'étiquettes sont conçues dans cette phase de manière à confronter l'utilisateur à ces cas difficiles. Notamment,

— lorsque l'étiquette à identifier est `VERB`, on proposera également `AUX`,

— lorsque l'étiquette à identifier est `ADV`, on proposera également `ADJ`,

— lorsque l'étiquette à identifier est `CONJ`, on proposera également `SCONJ`.

Par ailleurs, la phase de formation doit être la plus proche possible des conditions réelles de production des annotations. En particulier, en situation d'annotation réelle, l'étiquette correcte peut ne pas faire partie des étiquettes suggérées à l'utilisateur. Ce phénomène a donc également

été reproduit artificiellement : dans un cas sur dix, deux étiquettes incorrectes sont choisies au hasard et proposées à l'utilisateur.

Lors de cette phase, l'utilisateur ne peut pas passer à la phrase suivante tant que tous les mots n'ont pas été étiquetés correctement. Il peut vérifier ses réponses au fur et à mesure et doit les corriger pour pouvoir avancer. La figure 3.4.4 illustre une situation où l'utilisateur a annoté tous les mots de la phrase : les mots en vert sont ceux dont l'annotation proposée à été reconnue comme correcte par rapport à la référence dont nous disposons. Les mots en rouge sont ceux dont l'étiquette proposée reste à être corrigée afin de pouvoir passer à la phrase suivante.

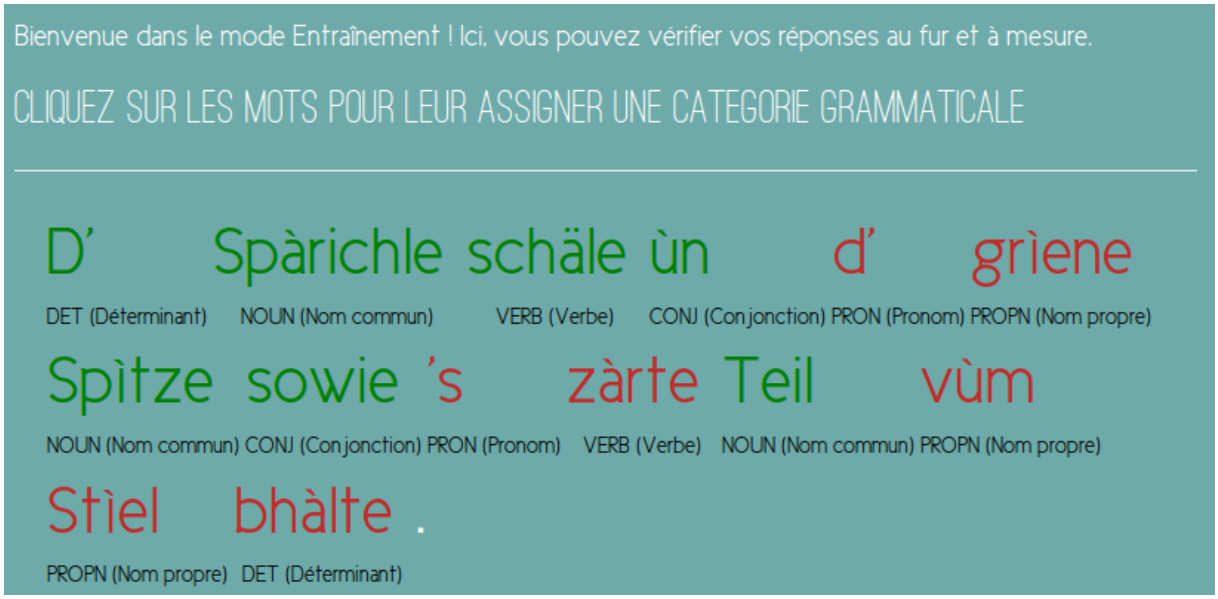

FIGURE 3.4.4 – Phase de formation.

La coloration des termes en fonction de l'étiquette choisie s'accompagne d'un rappel des exemples pour les annotations fausses ayant été proposées. Une fois la formation terminée, l'utilisateur peut commencer à produire des annotations.

Il peut par ailleurs revenir à tout moment à la phase de formation.

3.4.3 PHASE DE PRODUCTION DES ANNOTATIONS

Dans cette phase, l'utilisateur ne peut pas vérifier ses réponses, le passage à la phrase suivante signifie la production des annotations sélectionnées. Contrairement à la phase de formation,

l'utilisateur n'est pas obligé d'annoter tous les mots pour passer à la phrase suivante. Dans chaque série de quatre phrases est placée une phrase du corpus de référence. L'utilisateur est ainsi évalué régulièrement et son niveau de confiance est mis à jour en même temps que ses points à la fin de chaque séquence.

L'accord inter-annotateur entre l'utilisateur et la référence est ainsi recalculé à la fin de chaque séquence : la métrique choisie pour cela est le coefficient Kappa de Cohen (Cohen, 1960).

Les annotations sont quant à elles créées avec un niveau de confiance égal à celui de l'utilisateur.

3.5 RÉSULTATS

3.5.1 ANALYSE DE LA PLATE-FORME MISE EN PLACE

Nous avons pu, notamment grâce aux retours des utilisateurs, évaluer notre plate-forme au regard des critères définis par Lafourcade *et al.* (2015). Les auteurs fournissent un état de l'art riche des jeux ayant un but dans différents domaines. L'analyse du fonctionnement de chacun de ces jeux comparé à son succès leur permet de lister les caractéristiques d'un bon jeu ayant un but en distinguant les caractéristiques orientées joueur et les avantages tirés par le concepteur. Bien que l'application BISAME ne soit pas en l'état un jeu ayant un but, peu de fonctionnalités ludiques ayant été développées, nous avons évalué notre application selon les critères définis grâce à une note allant de 1 à 5 :

Caractéristiques orientées joueur		Avantages tirés par le concepteur	
Amusant	2	Ressources produites	4
Facile à appréhender	5	Constitution d'une base de connaissances	4
Relance par reclic	3	Prix de revient faible	x
Valorisation des joueurs	2	Acquisition rapide	5
Évolutive	1	Public large	2
Parties courtes	5	Éthique	5
Portabilité	1		

TABLEAU 3.5.1 – Évaluation de l'application BISAME selon les critères définis par Lafourcade *et al.* (2015).

Le coût de la réalisation de ce projet n'a pas pu être évalué. Il apparaît néanmoins qu'outre la conception et le développement de la plate-forme, la partie communication et publicité à réaliser ainsi que le support aux utilisateurs se sont avérés coûteux en temps. En outre, nous n'avons pas pu obtenir les temps cumulés d'annotation des utilisateurs. Les temps effectifs passés sur la tâche d'annotation ne peuvent être calculés à partir des dates d'interaction avec la base de donnée, l'utilisateur pouvant s'interrompre au milieu d'une séquence de phrases et reprendre plus tard : les durées passées sur une séquence terminée vont ainsi de quelques minutes à plusieurs jours. La réalisation du projet n'a par ailleurs engagé aucun coût financier spécifique, les utilisateurs de la plate-forme n'étant pas rémunérés pour leurs contributions.

3.5.2 PARTICIPATION

Les annotations ont été recueillies entre le 2 et le 15 mai 2016 : 3 189 annotations ont ainsi été produites en quinze jours par 22 utilisateurs.

104 personnes ont créé un compte, 35 ont complété la phase de formation et 22 ont effectivement produit des annotations.

Le graphique 3.5.1 montre la participation enregistrée en nombre de parties terminées par jour : après l'annonce effectuée par l'OLCA, la participation s'est rapidement essoufflée. C'est pourquoi nous avons entrepris de contacter directement *via* Facebook les utilisateurs s'exprimant dans des groupes tels que le « Centre Culturel Alsacien / Elsässisches Kulturzentrum » ou « Alsace Bilingue ». Cela nous a permis d'atteindre une nouvelle communauté à partir de

laquelle l'information s'est propagée de proche en proche de manière plus efficace que grâce à une communication officielle.

L'absence d'évolutivité dans le jeu genère néanmoins une lassitude chez les utilisateurs qui explique l'abandon de l'application après quelques séquences annotées. L'ultime mail de relance a permis d'obtenir des annotations supplémentaires auprès des utilisateurs ayant déjà participé.

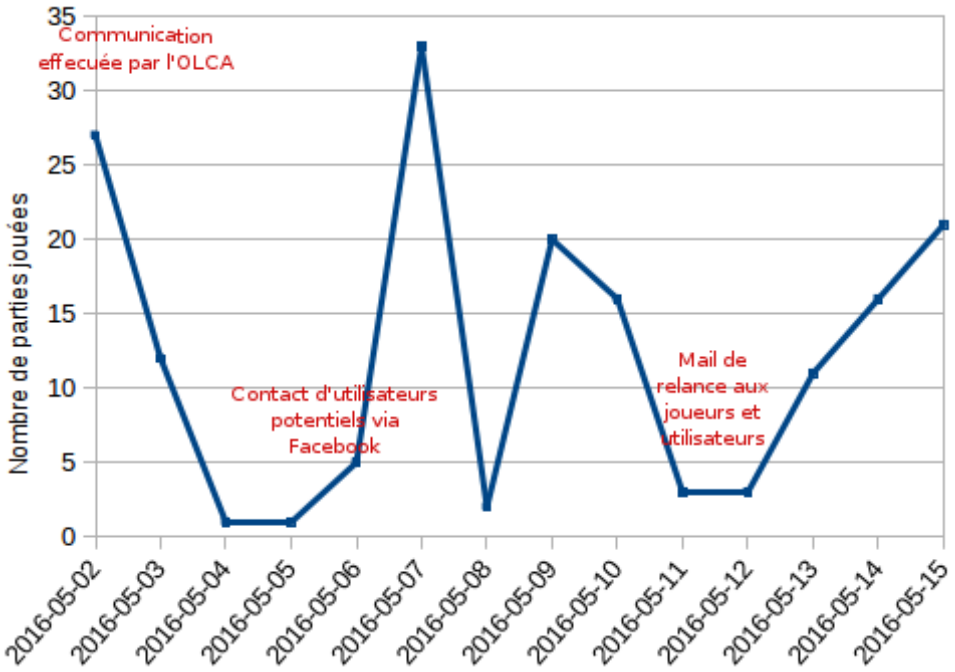

FIGURE 3.5.1 – Nombre de parties effectuées par jour.

3.5.3 UTILISATEURS

Les 22 utilisateurs se répartissent comme suit par intervalle de nombres d'annotations produites par chacun :

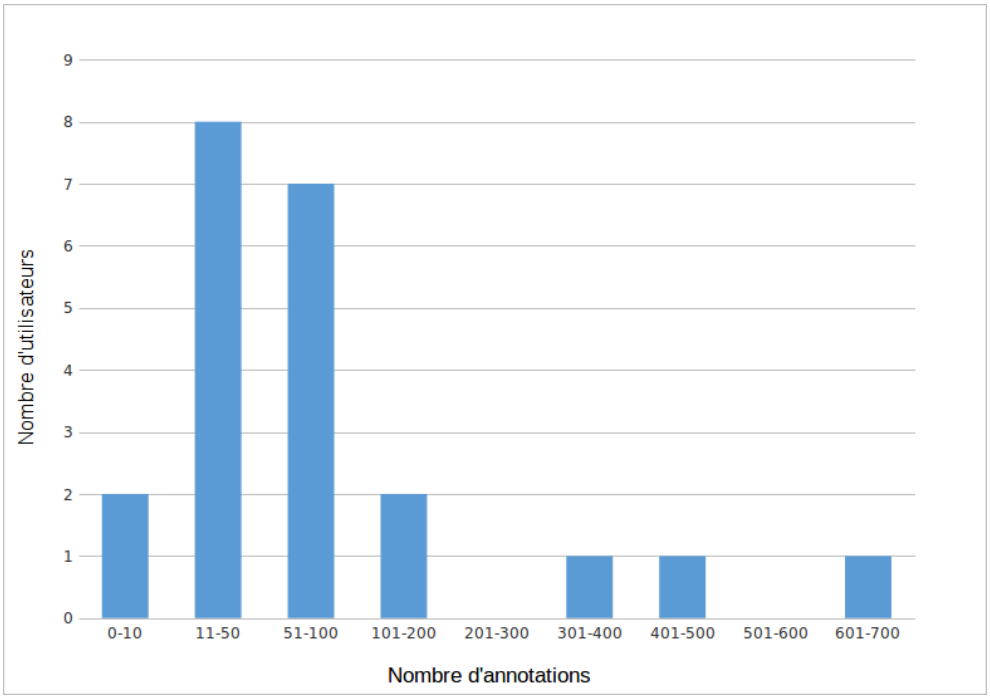

FIGURE 3.5.2 – Répartition des utilisateurs selon leur participation (en nombre d'annotations produites).

On observe ainsi une tendance très proche de celles observées au cours d'autres projets de myriadisation pour l'annotation en TAL (et en général dans les projets de *crowdsourcing*) : il ne s'agit pas de mobiliser une « foule » de joueurs, la grande majorité des annotations étant produites par un petit nombre d'utilisateurs motivés et se spécialisant dans la tâche (Fort, 2015). Ci-dessous les graphiques décrivant respectivement les nombres de points par joueurs pour les plate-formes Phrase Detectives (Chamberlain *et al.*, 2013) et Jeux de Mots [9], le nombre d'annotations pour la plate-forme Zombilingo, et le score des joueurs pour BISAME :

9. Voir `http://www.jeuxdemots.org/generateRanking-4.php`.

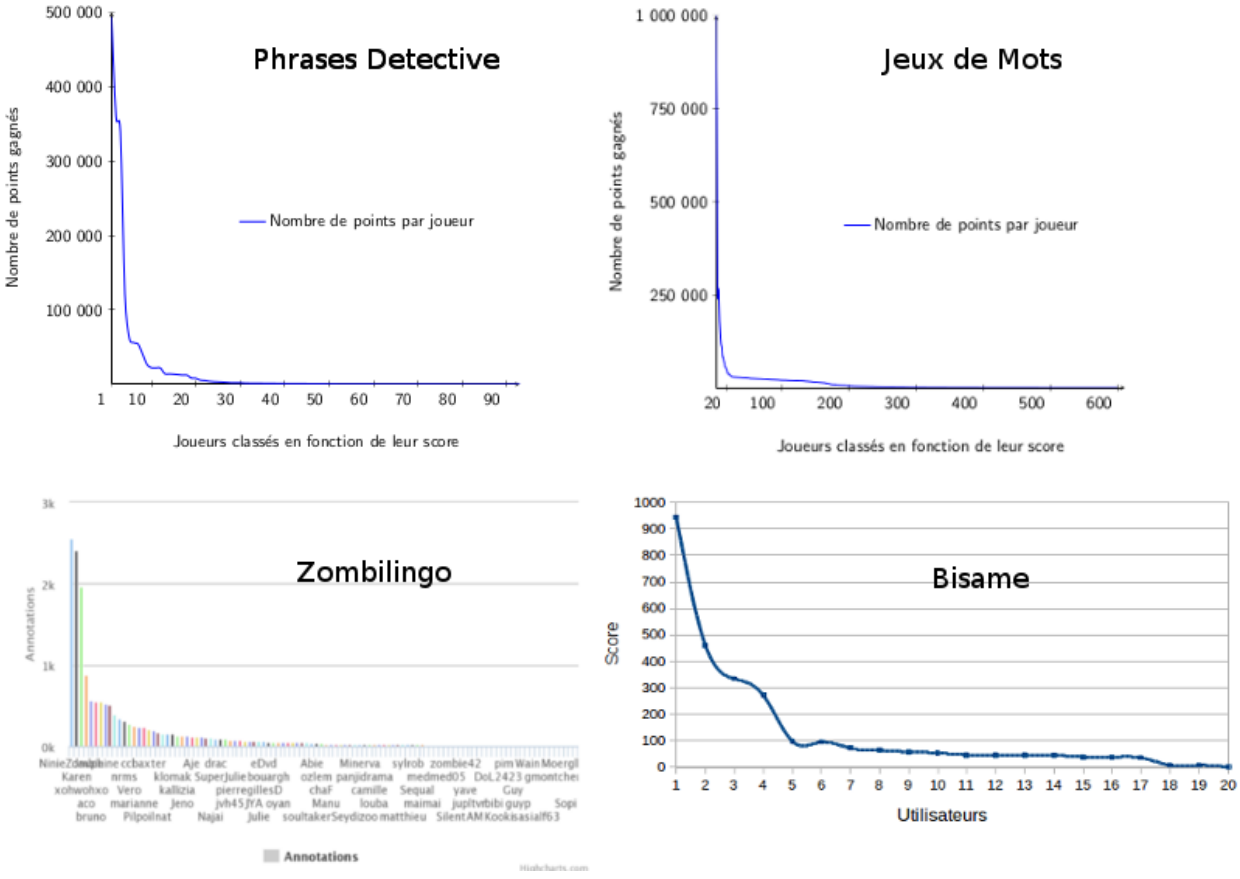

FIGURE 3.5.3 – Répartition des utilisateurs par score ou équivalent pour quatre applications d'annotation par le jeu.

Par ailleurs, le contact avec les utilisateurs a permis de définir que les communautés intéressées pour participer à un projet de ce type sont multiples. Parmi les utilisateurs de la plate-forme on trouve effectivement aussi bien :

— des étudiants,

— des jeunes actifs engagés dans des associations de défense de la langue,

— des personnes retraitées, anciens enseignants notamment.

Une dizaine d'utilisateurs a demandé des précisions sur le projet. Les communications sur la plate-forme ainsi que le bref descriptif du projet donnés sur le site se sont avérés insuffisants. Cela pose la question du degré de détail qui doit être communiqué de prime abord. Il est

prévu d'intégrer un onglet « Information » à l'application, mais il est apparu que le caractère individuel d'une explication détaillée du projet et des enjeux est un facteur très motivant pour les utilisateurs.

La question de l'écriture et de l'orthographe choisie s'est beaucoup posée lors des dialogues qui ont été menés avec les utilisateurs. Si certains ont été déconcertés par l'écriture de l'alsacien qui ne leur est pas familière, l'ayant pratiqué à l'oral exclusivement, quelques utilisateurs ont suggéré des corrections orthographiques. Il a fallu rappeler à plusieurs reprises que le projet mis en place ne prétend nullement imposer une vision ni de l'orthographe, ni de la grammaire.

Plusieurs utilisateurs ont par ailleurs fait part de leur souhait de pouvoir utiliser l'application dans une variante de l'alsacien leur étant davantage familière. Si un utilisateur se dit « déconcentré par l'alsacien haut-rhinois », une utilisatrice souhaiterait pouvoir rajouter des phrases en « Mulhousien ».

3.6 Évaluation

3.6.1 Une annotation produite de qualité

L'évaluation a été effectuée selon la méthodologie décrite en partie 3.2.6. Nous définissons ainsi un « utilisateur global » en synthétisant les annotations recueillies, et nous calculons l'accord inter-annotateur avec les valeurs de référence dont nous disposons.

Nous atteignons une couverture du corpus de plus de 90 %.
L'annotation produite par les joueurs présente une exactitude de 89,21 %, avec 123 erreurs pour 1 141 mots différents annotés. Le coefficient kappa calculé est de $k = 0.85$, l'annotation obtenue est donc satisfaisante (en accord avec Artstein et Poesio (2008), qui définit $k = 0.8$ comme seuil au-delà duquel on peut considérer que l'annotation est satisfaisante [10]).

10. « We therefore feel that if a threshold needs to be set, 0.8 is a good value. »

3.6.2 Analyse de l'annotation globale obtenue

La précision, le rappel et la F-mesure des annotations obtenues par catégorie ont été calculées et comparées aux deux annotations automatiques fournies (voir figures 3.6.1 à 3.6.3).

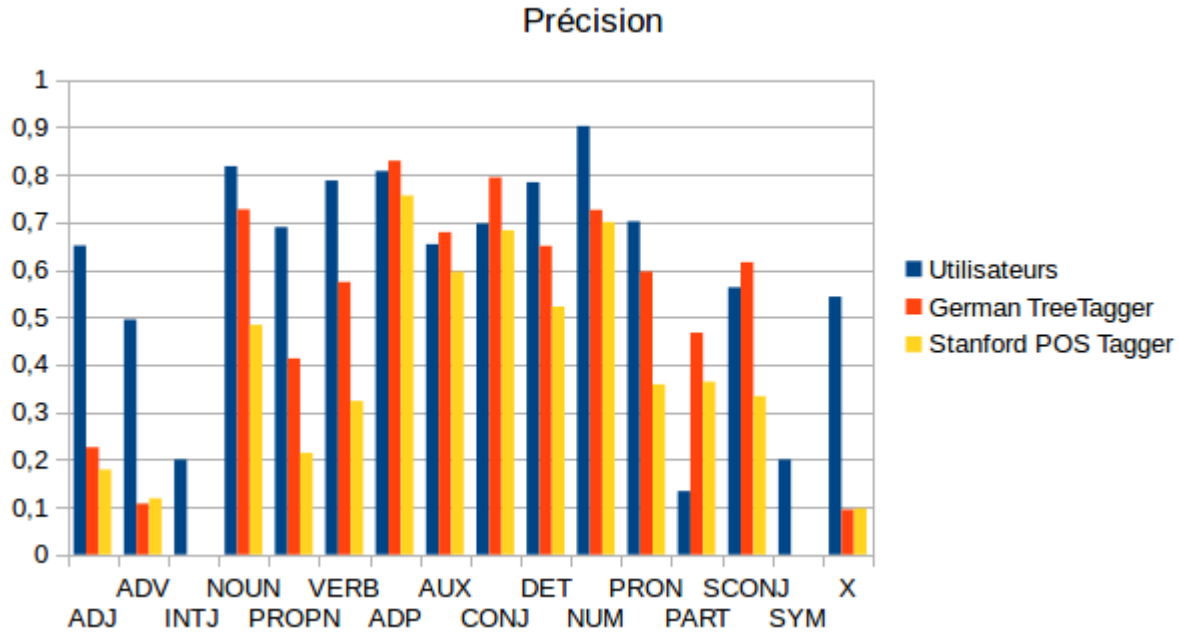

FIGURE 3.6.1 – Précision.

Les erreurs commises par l'« utilisateur global » ont été comparées aux mots difficiles qui avaient été identifiés du fait du désaccord entre les deux annotations manuelles qui avaient été fournies par les membres du LiLPa (en annexe) : 40 % des erreurs commises par les utilisateurs concernaient des mots ayant posé problème lors de l'annotation manuelle par les expertes linguistes. Le reste des erreurs commises par les annotateurs seuls se trouve en majorité dans la catégorie `X`, catégorie des mots étrangers, avec seulement un tiers des occurrences identifiées, et les catégories `CONJ` et `AUX`.

Par ailleurs, d'après les résultats répertoriés figure 3.6.3 on observe que si l'annotation manuelle des utilisateurs est globalement meilleure (et en particulier pour les catégories `ADJ`, `ADV`, `NOUN`, `PROPN`, `VERB`, `SCONJ`, `DET`, `NUM`, et `X`) le German TreeTagger obtient de meilleurs résultats pour les catégories `AUX`, `CONJ` et `PART`.

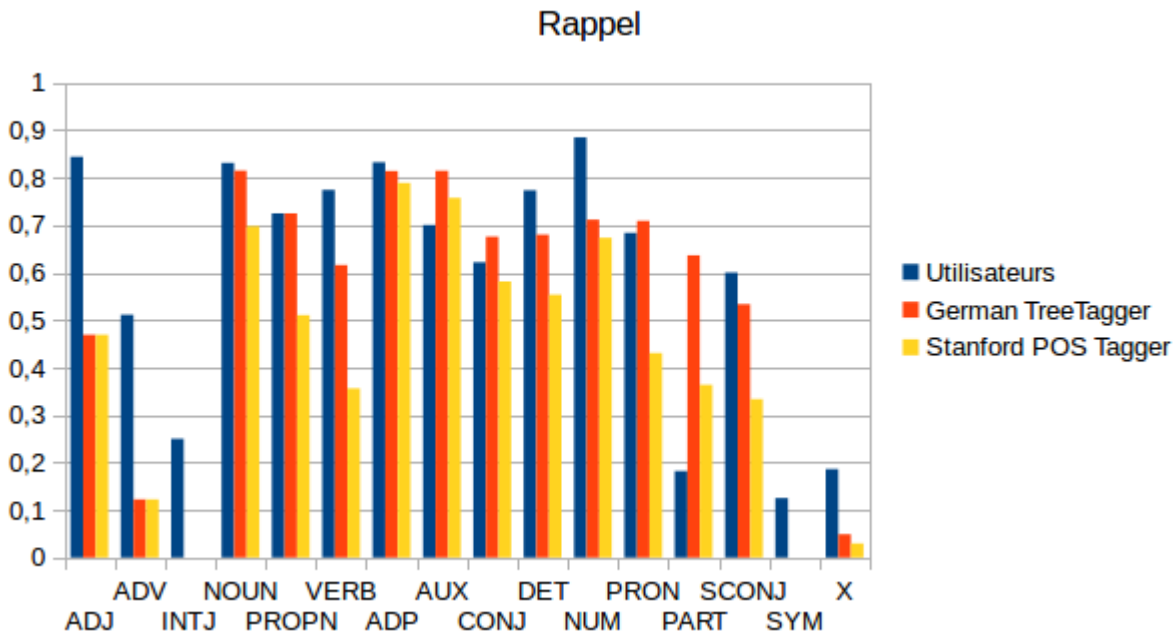

FIGURE 3.6.2 – Rappel.

Les difficultés propres aux utilisateurs humains et les catégories pour lesquelles les outils automatiques sont satisfaisant ont été identifiées. Ces observations doivent être intégrées à la conception de l'application. Par exemple un entraînement renforcé pourra être proposé pour les catégories les plus problématiques. Par ailleurs en connaissant les catégories pour lesquelles la compétence humaine dépasse celle des outils automatiques, on pourra restreindre l'action de l'utilisateur humain à la correction de certaines étiquettes spécifiques.

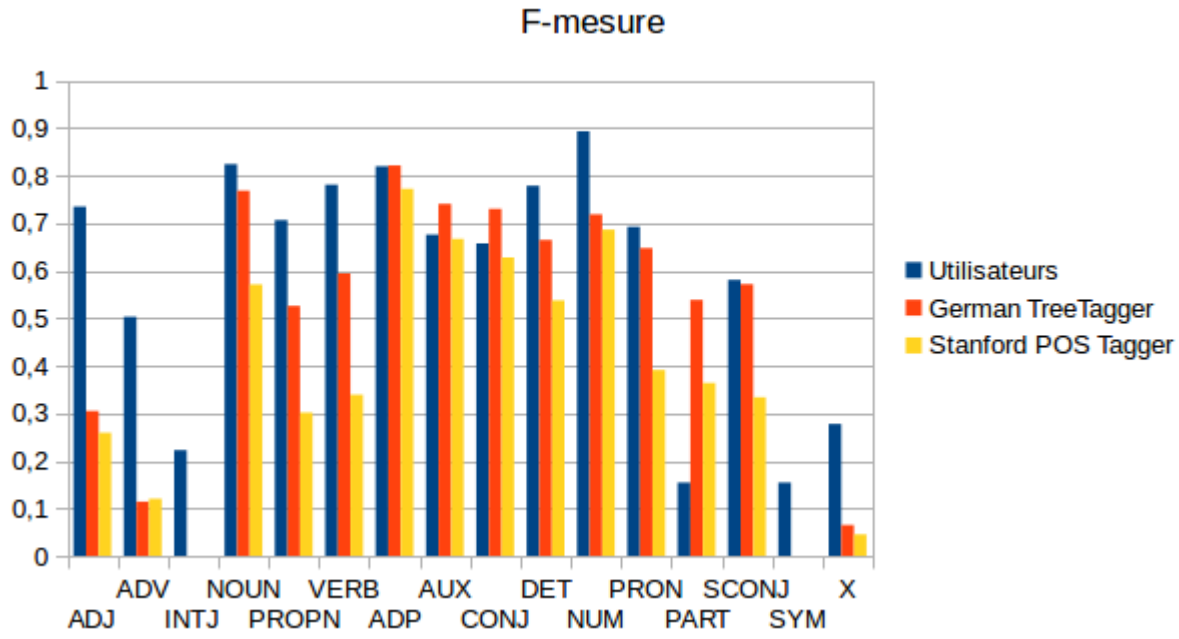

FIGURE 3.6.3 – F-mesure.

CONCLUSION

Ce projet a permis de montrer qu'il est possible de mobiliser rapidement une communauté de locuteurs afin de constituer un corpus annoté. La spécificité de l'annotation en parties du discours ne nous permet pas de généraliser sur la validité de l'extension de cette méthode pour tous types d'annotation.

Néanmoins, pour une tâche qui présente une certaine difficulté, comme le montrent les résultats des annotations manuelles de deux linguistes, nous avons obtenu des résultats satisfaisants et très encourageants compte tenu de la qualité du corpus obtenu et de la brièveté de l'expérience menée.

Ces résultats nous poussent à envisager l'annotation de corpus bruts pour lesquels nous n'avons pas d'annotation de référence, l'application BISAME devenant alors une réelle plate-forme d'annotation collaborative.

La constitution d'un corpus de l'alsacien annoté en parties du discours de taille supérieure à l'existant (la référence dont nous disposions) permettra à terme d'entraîner à nouveau les outils d'annotation automatique afin d'améliorer leurs performances de pré-annotation. En suivant ce cercle vertueux, l'annotation manuelle à effectuer serait de plus en plus simplifiée ce qui accélérerait l'annotation de corpus variés et il deviendrait possible d'obtenir des corpus de l'alsacien annotés automatiquement de bonne qualité.

Le dialogue avec les différents utilisateurs de la plate-forme a été extrêmement enrichissant et nous a permis d'acquérir une vision d'une certaine réalité linguistique de l'alsacien, ainsi que de comprendre les attentes des locuteurs face à un projet de ce type.

Les pistes d'amélioration pour l'application mise en place sont nombreuses et sont en partie le fruit du retours des utilisateurs. Notamment, nous souhaitons développer l'aspect ludique de l'application, en introduisant entre autres un classement des meilleurs joueurs, une évolutivité dans le jeu, la possibilité d'annoter « en duel ». D'un point de vue des fonctionnalités, la possibilité de pouvoir choisir plusieurs étiquettes doit être envisagée, ainsi que le fait de pouvoir proposer des orthographes alternatives.

Il apparaît en outre que la non-utilisabilité de l'application sur smart-phone a été un élément bloquant pour plusieurs utilisateurs *a priori* motivés, et doit être résolue.

Nombre d'utilisateurs se sont montrés très enthousiastes quant au projet BISAME, suggérant d'eux-mêmes des nouvelles fonctionnalités, telles que la possibilité de proposer une traduction ou la volonté de pouvoir travailler dans certaines variantes de l'alsacien. Cette motivation laisse supposer qu'une interface plus ludique et enrichie pourrait réunir une communauté solide que nous sommes déjà parvenus à intéresser et à mobiliser dans le cadre de ce court projet.

Il est par ailleurs envisagé de combiner l'annotation comme jeu ayant un but avec un enseignement de la méthode ORTHAL sous la forme d'un jeu sérieux. Cela permettrait de renforcer l'utilisation de cette méthode facilitant l'emploi de l'écrit pour l'alsacien, tout en consolidant une communauté de locuteurs autour de l'application.

Enfin, compte tenu des similarités typologiques existant entre les différentes langues de France (présence d'un continuum dialectal, absence de norme pour l'écriture, existence d'une communauté de locuteurs désireux de défendre leur langue), il est également envisagé d'étendre cette méthodologie à d'autres langues de France, ce qui nous permettrait également de valider l'adaptabilité de la plate-forme.

ANNEXES

1. IDENTIFICATION DES MOTS PROBLÉMATIQUES PAR COMPARAISON DES ANNOTATIONS MANUELLES FOURNIES

Étiquette1	Étiquette2	Formes (traduction | Étiquette retenue lors de l'adjudication)
ADJ	ADP	bis (jusqu'à) wittersch (en continuant) erscht (!premier ADJ)
ADJ	CONJ	unn
ADJ	DET	jedes
ADJ	ADV	gùssàrtig bìssele (NOUN) lang licht (ADV) nawadràà (ADV) erscht (ADV) friajher (plus tôt ADV) nawadràà (à côté ADV) wittersch (en continuant)
ADJ	NUM	drìtta (ADJ)
ADJ	NOUN	Làtiinisch (NOUN) Hochditsch (NOUN) {structure Uf + nom / ìn + nom}
ADJ	PRON	àlles (PRON)
ADJ	PROPN	elsassischa ditscha
ADJ	VERB	bekànnta (connu ADJ) gstudiarta (cultivé ADJ)
ADP	ADV	denäwe zàmme debi (ADV) zuere (vers) sogàr (ADV) wìdder (ADV)
ADP	CONJ	sowie ùn'em soboel (CONJ) Denn (CONJ) ebbena (ADV)
ADP	DET	üs
ADP	NOUN	Ànna (en ADP)
ADP	NUM	Zwìscha (ADP)
ADP	PART	ùm (à propos de) züe (trop PART)
ADP	PROPN	von (Wolfram von Eschenbach ADP)
ADP	SCONJ	wo (que ->PRON) Àls (comme ->CONJ)
ADP	VERB	lon (laisser)
ADV	CONJ	dànn sovìel àwwer (mais CONJ) dànn (ADV) Wahrend'm (pendant le ->ADP) àlso (donc ADV) sogàr (voire ADV) àlso (donc ADV) wia (comme CONJ)
ADV	NOUN	àlso (donc ADV)

ADV	PART	nìt (PART) genue
ADV	PRON	derfer (pour cela ADV)
ADV	SCONJ	wenn (si SCONJ) wo (ADV) wenn (quand ADV)
ADV	VERB	wìdder (de nouveau)
AUX	VERB	wäre hab han wird worra (!AUX) sìn ìsch war (AUX) känna (VERB) worra (!VERB) hann (AUX) gsii (AUX)
AUX	PRON	ìsch
CONJ	PRON	ich
CONJ	SCONJ	dàss (SCONJ) wie wial (parce que) àss (que SCONJ)
DET	NOUN	’s (le/la) de (le) d’ (la)
DET	PRON	se (les) dàs (le DET) dia (les/cette DET) da (les DET)
INTJ	X	Eh
NOUN	NUM	Ànderthàlb
NOUN	PART	nit (PART)
NOUN	PRON	’s (ich mach’s PRON) ’s (wo’s nìmm gìtt PRON)
NOUN	PROPN	Israelita (NOUN) Baron (...de Rose PROPN)
NOUN	VERB	igstellt (VERB) Biacher (livre NOUN) Duck (impression)
NUM	PART	2008
NUM	PROPN	1478
NUM	X	Th7
PRON	SCONJ	wo (qui PRON) wo-n-ìhm (que lui PRON)
PRON	VERB	’s (il)
SYM	X	B49 (PROPN)
VERB	X	garantier gsaat fànga-n-à

TABLEAU 3.6.1 – Comparaison des annotations manuelles fournies par les chercheuses du LiLPa.

2. Matrices de confusion et accords inter-annotateurs calculés pour les deux annotations manuelles fournies par les chercheuses du LiLPa

Hoflieferant_P53

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADP	0	1 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADV	0	0	1 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AUX	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	2	0
CONJ	0	0	1	0	7	0	0	0	0	0	0	0	0	0	0	0	0
DET	0	1	0	0	0	1 0	0	1	0	0	0	0	0	0	0	0	0
INTJ	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
NOUN	0	0	0	0	0	0	0	2 2	0	0	0	0	0	0	0	0	0
NUM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PART	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
PRON	0	0	0	0	1	0	0	1	0	0	3 0	0	0	0	0	0	0
PROPN	0	0	0	0	0	0	0	1	0	0	0	8	0	0	0	0	0
PUNCT	0	0	0	0	0	0	0	0	0	0	0	0	5 0	0	0	0	0
SCONJ	0	0	2	0	2	0	0	0	0	0	0	0	0	3	0	0	0
SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VERB	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2 5	2
X	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	4

recettes

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	2 1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADP	2	3 1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ADV	6	4	7	0	1	0	0	0	0	0	0	0	0	0	0	0	0
AUX	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CONJ	0	3	3	0	2 2	0	0	0	0	0	0	0	0	0	0	0	0
DET	1	0	0	0	0	3 4	0	3	0	0	0	0	0	0	0	0	0
INTJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOUN	0	0	0	0	0	0	0	7 7	0	0	0	0	0	0	0	0	0
NUM	0	0	0	0	0	0	0	1	9	0	0	0	0	0	0	0	0
PART	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRON	1	0	0	0	0	4	0	0	0	0	4	0	0	0	0	0	0
PROPN	0	0	0	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	5 1	0	0	0	0
SCONJ	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VERB	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	5 8	1
X	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

wikipedia1

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	9	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0
ADP	0	3 9	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0
ADV	2	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	1
AUX	0	0	0	2 3	0	0	0	0	0	0	0	0	0	0	0	3	0
CONJ	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	0	0
DET	0	0	0	0	0	4 6	0	0	0	0	3	0	0	0	0	0	0
INTJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOUN	0	0	0	0	0	0	0	6 6	0	0	0	1	0	0	0	0	1
NUM	0	0	0	0	0	0	0	0	1 6	1	0	0	0	0	0	0	1
PART	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
PRON	0	0	1	0	0	0	0	0	0	0	5	0	0	4	0	1	1
PROPN	0	0	0	0	0	0	0	0	0	0	0	3 7	0	0	0	0	0
PUNCT	0	0	0	0	0	0	0	0	0	0	0	0	4 8	0	0	0	1
SCONJ	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
VERB	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	3 4	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1 9

wikipedia2

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	1 7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0
ADP	0	5 2	0	0	0	0	0	5	0	0	0	1	1	0	0	0	0
ADV	2	0	2 1	0	0	0	0	1	0	0	0	0	0	0	0	1	0
AUX	0	0	0	2 7	0	0	0	0	0	0	0	0	0	0	0	5	0
CONJ	0	1	4	0	1 1	0	0	0	0	0	0	0	0	0	0	0	0
DET	0	1	0	0	0	4 5	0	0	0	0	2	0	0	0	0	0	0
INTJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOUN	3	0	0	0	0	0	0	7 1	0	0	0	0	0	0	0	1	0
NUM	0	0	0	0	0	0	0	0	1 7	0	0	2	0	0	0	0	0
PART	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRON	0	0	0	0	0	0	0	2	0	0	2 0	0	0	0	0	0	0
PROPN	0	0	0	0	0	0	0	0	0	0	0	4 5	2	0	0	0	0
PUNCT	0	0	0	0	0	0	0	0	0	0	0	0	7 8	0	0	0	1
SCONJ	0	3	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0
SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
VERB	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	3 9	1
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	5

BIBLIOGRAPHIE

Ron ARTSTEIN et Massimo POESIO : Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Vincent BERMENT : *Méthodes pour informatiser les langues et les groupes de langues «peu dotées»*. Thèse de doctorat, Université Joseph-Fourier-Grenoble I, 2004.

Delphine BERNHARD et Anne-Laure LIGOZAT : Hassle-free pos-tagging for the alsatian dialects. *Non-Standard Data Sources in Corpus Based-Research*, pages 85–92, 2013.

Alena BÖHMOVÁ, Jan HAJIČ, Eva HAJIČOVÁ et Barbora HLADKÁ : The prague dependency treebank. *In Treebanks*, pages 103–127. Springer, 2003.

Myriam BRAS et Marianne VERGEZ-COURET : Batelòc : a text base for the occitan language. *In Language Documentation & Conservation*, pages 133–149, 2014.

Jon CHAMBERLAIN, Karën FORT, Udo KRUSCHWITZ, Mathieu LAFOURCADE et Massimo POESIO : Using games to create language resources : Successes and limitations of the approach. *In The People's Web Meets NLP*, pages 3–44. Springer, 2013.

Christos CHRISTODOULOPOULOS, Sharon GOLDWATER et Mark STEEDMAN : Two decades of unsupervised pos induction : How far have we come ? *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584, Massachusetts, États-Unis, octobre 2010. Association for Computational Linguistics.

Jacob COHEN : A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Kevin Bretonnel COHEN, Lynne FOX, Philip V. OGREN et Lawrence HUNTER : Corpus design for biomedical natural language processing. *In Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases : mining biological semantics*, pages 38–45, 2005.

Danielle CRÉVENAT-WERNER et Edgar ZEIDLER : *Orthographe alsacienne - Bien écrire l'alsacien de Wissembourg à Ferrette*. Jérôme Do Bentzinger, 2008.

Sandipan DANDAPAT, Sudeshna SARKAR et Anupam BASU : Automatic part-of-speech tagging for bengali : An approach for morphologically rich languages in a poor resource scenario. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 221–224, Prague, République tchèque, juin 2007. Association for Computational Linguistics.

Jean-Michel ELOY, Fanny MARTIN et Christophe REY : Picartext : Une ressource informatisée pour la langue picarde. *In Actes de la Traitement Automatique des Langues Régionales de France et d'Europe*, pages 19–25, Caen, France, 2015.

Annie FORET, Valérie BELLYNCK et Christian BOITET : Akenou-breizh, un projet de plateforme valorisant des ressources et outils informatiques et linguistiques pour le breton. *In Actes de la Traitement Automatique des Langues Régionales de France et d'Europe*, pages 26–37, Caen, France, juin 2015. Association pour le Traitement Automatique des Langues.

Karën FORT : *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Thèse de doctorat, Université Paris-Nord-Paris XIII, 2012.

Karën FORT : Experts ou (foule de) non-experts ? la question de l'expertise des annotateurs vue de la myriadisation (crowdsourcing). *In 8es Journées Internationales de Linguistique de Corpus*, Orléans, France, septembre 2015.

Karën FORT : *Collaborative Annotation for Reliable Natural Language Processing*. Focus series. ISTE Wiley, 2016.

Karën FORT, Gilles ADDA et K Bretonnel COHEN : Amazon mechanical turk : Gold mine or coal mine ? *Computational Linguistics*, 37(2):413–420, 2011.

Karën FORT, Adeline NAZARENKO et Sophie ROSSET : Modeling the complexity of manual annotation tasks : a grid of analysis. *In International Conference on Computational Linguistics*, pages 895–910, Mumbai, Inde, 2012.

Karën FORT et Benoît SAGOT : Influence of pre-annotation on pos-tagged corpus development. *In Proceedings of the fourth linguistic annotation workshop*, pages 56–63, Uppsala, Suède, 2010. Association for Computational Linguistics.

Benjamin M GOOD et Andrew I SU : Crowdsourcing for bioinformatics. *Bioinformatics*, page btt333, 2013.

David JURGENS et Roberto NAVIGLI : It's all fun and games until someone annotates : Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464, 2014.

Mathieu LAFOURCADE, Nathalie LE BRUN et Alain JOUBERT : *Jeux et intelligence collective : résolution de problèmes et acquisition de données sur le web*. ISTE éd., 2015.

Tak Yeon LEE, Casey DUGAN, Werner GEYER, Tristan RATCHFORD, Jamie C RASMUSSEN, N Sadat SHAMI et Stela LUPUSHOR : Experiments on motivational feedback for crowdsourced workers. *In ICWSM*, Ann Arbor, Michigan, États-Unis, 2013.

Bente MAEGAARD, Steven KRAUWER, Khalid CHOUKRI et L JØRGENSEN : The blark concept and blark for arabic. *In Fifth International Conference on Language Resources and Evaluation, LREC'06*, Gênes, Italie, 2006.

Michel MALHERBE : Les langages de l'humanité (une encyclopédie des 3000 langues parlées dans le monde). *Bouquins*, 1983.

Mitchell MARCUS, Beatrice SANTORINI et Mary Ann MARCINKIEWICZ : Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Mike MAXWELL et Baden HUGHES : Frontiers in linguistic annotation for lower-density languages. *In Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, LAC '06, pages 29–37, Stroudsburg, PA, États-Unis, 2006. Association for Computational Linguistics.

Guy PERRIER et Bruno GUILLAUME : Leopar : an interaction grammar parser. *In ESSLLI 2013-Workshop on High-level Methodologies for Grammar Engineering*, pages 121–122, Düsseldorf, Allemagne, 2013.

Slav PETROV, Dipanjan DAS et Ryan MCDONALD : A universal part-of-speech tagset. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie, mai 2012. European Language Resources Association (ELRA).

Massimo POESIO, Jon CHAMBERLAIN, Udo KRUSCHWITZ, Livio ROBALDO et Luca DUCCESCHI : The phrase detective multilingual corpus, release 0.1. *In Collaborative Resource Development and Delivery Workshop Programme (LREC'12)*, page 34, Istanbul, Turquie, mai 2012.

Delyth PRYS : The blark matrix and its relation to the language resources situation for the celtic languages. *In Fifth International Conference on Language Resources and Evaluation, LREC'06, Strategies for developing machine translation for minority languages*, page 31, Gênes, Italie, 2006.

Yves SCHERRER et Benoît SAGOT : Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche. *In Atelier TALARE, TALN 2013*, Les Sables d'Olonne, France, juin 2013. ATALA.

Helmut SCHMID : Improvements in part-of-speech tagging with an application to german. *In Natural language processing using very large corpora*, pages 13–25. Springer, 1995.

William A SCOTT : Reliability of content analysis : The case of nominal scale coding. *In Public opinion quarterly*, pages 321–325. JSTOR, 1955.

Lucie STEIBLE : *Timing of plosives in Alsatian and French spoken in Alsace*. Thèse de doctorat, Université de Strasbourg, décembre 2014.

Helmer STRIK, Walter DAELEMANS, Diana BINNENPOORTE, Janienke STURM, Folkert de VRIEND et Catia CUCCHIARINI : Dutch hlt resources : from blark to priority lists. *In INTERSPEECH*, Denver, Colorado, États-Unis, 2002.

Oscar TÄCKSTRÖM, Dipanjan DAS, Slav PETROV, Ryan MCDONALD et Joakim NIVRE : Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12, 2013.

Kristina TOUTANOVA, Dan KLEIN, Christopher D MANNING et Yoram SINGER : Feature-rich part-of-speech tagging with a cyclic dependency network. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

Marianne VERGEZ-COURET et Assaf URIELI : Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan. *In TALARE 2015*, pages 61–71, Caen, France, juin 2015.