HAL open science

A Broad NLP Training from Speech to Knowledge

Maxime Amblard, Miguel Couceiro

To cite this version:

Maxime Amblard, Miguel Couceiro. A Broad NLP Training from Speech to Knowledge. NAACL 2021 - 5th Workshop on Teaching NLP at the Annual Conference of the North American Chapter of the Association for Computational Linguistics, Jun 2021, Mexico / Virtual, Mexico. hal-03200480

HAL Id: hal-03200480

https://inria.hal.science/hal-03200480v1

Submitted on 16 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Broad NLP Training from Speech to Knowledge

Amblard Maxime
Miguel Couceiro
LORIA, UMR 7503, Université de Lorraine, Inria and CNRS
{maxime.amblard, miguel.couceiro}@univ-lorraine.fr

Abstract

In 2018, the Master of Science programe in NLP[1] opened at the IDMC[2]. Far from being a creation ex-nihilo, it is the product of a history and many reflections on the field and its teaching. This article proposes epistemological and critical elements on the opening and maintainance of this so far new master's programe in NLP.

This article discusses the epistemological background of the creation of the Master of Science program in natural language processing (NLP) at the University of Lorraine in 2018. It puts forward a critical analysis of the environment, of the methodology chosen to produce this program, and it highlights the salient elements for the teaching of NLP.

Currently, the master's degree is taught at the IDMC of the University de Lorraine. It is accessible to students trained within the undergraduate programe of the institute or to students that successfully apply to the programe. A jury from the pedagogical team evaluates the adequacy of the candidates' profile with the requirements and expectations of the training. For the detailed description of the master's degree in NLP, please visit pour dedicated website[3].

We propose a singular training at the French level, and probably at the international level. It is a Master's degree that gives the tools and methods to carry out language data processing. If NLP is at the heart of the training, we have opened up to speech and knowledge processing. The objective is to train tomorrow's professionals in both the economic and academic fields for language data processing.

In this article we present the Master's degree in NLP at the IDMC. After quickly reviewing the history, we present the approach explaining the constitution of the program, then we return to the program itself. Finally we put forward some elements of analysis.

1 History

The Department of Mathematics and Computer Science at the University of Nancy 2 pursued a policy of opening up both of these disciplines to human and social sciences. At the beginning of the 2000s, the department proposed a training program based on these aspects by integrating economics and business management. Following this interdisciplinary view, the pedagogical team created a bachelor's degree program in cognitive sciences. Several thematic openings were made, in particular, towards psychology, biology, linguistics and neurosciences. The dynamics was successful and paved the way to the opening of a complete master's programe in Cognitive Sciences (two years). The core of the training has clearly been the cognitive sciences.

1.1 Research environment

The French higher education and research implies that research and teaching are carried out in different components. Thus, the teachers carry out their research in a different research laboratory. In Nancy (France), two laboratories are particularly concerned: Loria[4] and ATILF[5]. These two laboratories are mixed research units, i.e., university laboratories co-accredited by a research center. This co-accreditation makes it possible to integrate staff who only have research duties. It appears as a quality label for the research carried out. The ATILF is co-accredited by the CNRS[6] and the Loria is co-accredited by the CNRS and Inria[7]. The two

[1] Natural Language Processing

[2] Institut des Sciences du Digital, du Management et de la Cognition https://idmc.univ-lorraine.fr/

[3] https://idmc.univ-lorraine.fr/idmc-master-degree-in-natural-language-processing/

[4] https://www.loria.fr/en/

[5] https://www.atilf.fr//

[6] Centre National pour la Recerhche Scientifique https://www.cnrs.fr/

[7] Institut National de la Recherche en Informatique et Automatique https://inria.fr/

laboratories ensure a sustained research activity in the field in Nancy. ATILF is organized with three themes: lexicons; corpora and knowledge; and dynamic aspects of language. Atilf is known for the *Trésor de la Langue Française Informatisé*, a large online dictionary of French.

Research in computer science is carried out jointly by loria and the Inria Nancy Grand-Est center. Beyond institutional issues, the researchers work jointly on all computer science domains. In particular, Loria is organised into research departments, one of which, the largest, is entirely devoted to NLP research. The department gathers about fifty permanent staff members in eight teams of different sizes (Cello, Team K, Multispeech, Orpailleur, Read, SMarT, Sémagramme, Synalp).

1.2 Creation of a training program in NLP

The dynamics of the teaching of mathematics and computer science open to other disciplines, and the strong research activity in Nancy have proven to be a favorable ground for the creation of a training program in NLP. In the early 2000s, the Master's degree in Cognitive Science integrated language-related issues into some of its courses. A DESS, an applied training program at the master's level, was also opened for a few years under the direction of Fiammetta Nammer and Yannick Toussaint. However, as the people in charge were not in the teaching component and the theme was not yet explicitly recognized by the French industry, this attempt had to be stopped quickly. It nevertheless established the possibility of developing a curriculum in NLP.

It was in 2005 that the main turning point took place. At that time, a double dynamic was set up, embodied by Patrick Blackburn and Guy Perrier. The latter was a university professor in computer science and set up a training program integrating students from computer science programs with linguistics programs within the Master's degree in Cognitive Science. During this time, Patrick Blackburn was working on the creation of the Erasmus Mundus Language and Communication Technologies consortium, of which Nancy would be the partner for France.

The Erasmus Mundus european program on Language and Communication Technologies[8] is a joint master's program between seven European Universities. The program is supported by Europe and propose grants for students. They are selected at the international level for the qualitiy of their trainning and motivation. Some students integrate the program as self-funded. Each student spends one year in one partner university of the consortium and the second year in another one. They share some activities and events all together during the year.

In 2009, Maxime Amblard took over the responsibility of the Master's program on Cognitive Sciences (SC), institutionalizing the existence of the NLP theme in the university. From that moment on, he has been in charge of the NLP theme, first as head of the SC Master's program from 2009 to 2012[9], and then as head of the M2 TAL[10] speciality from 2010 to 2012 and from 2015 to 2018. He carried out two years of sabatical leave[11] fully dedicated to research.He was then the project leader for the creation of the Master's program in NLP that was opened in 2018. He was the creator and organizor of the teaching from 2009 to today, both in Cognitive Sciences and in NLP. The responsibilities of 2nd year of the NLP program were assumed by Fabienne Venant in 2009-2010, Laure Buhry in 2013-14, then Miguel Couceiro from 2018. The coordination of the Erasmus Mundus LCT was carried by Miguel Couceiro since 2015.

1.3 Other contextual elements

In addition to the synergy between the cognitive science and NLP courses, it is worth noting the presence of other important training factors. On the one hand, there is a Master's degree in Language Sciences as well as a second Erasmus Mundu program specialized in Lexicology: EMLEX[12]. This program works very differently because the classes share semesters together on a given campus. The students are therefore not systematically present in Nancy. In all cases, they are also international students selected for their results and motivation.

In addition, the Université de Lorraine integrates engineering schools into its structure. In particular, the Ecole des Mines de Nancy. This engineering training is recognized as being of high quality in France. However, students from this type of training rarely go on to complete a PhD, despite their qualities. For those who are interested in this type of opening, it is traditional to offer double courses

[8] https://lct-master.org/

[9] The responsibility of the SC master has since been assumed by Manuel Rebuschi

[10] "Traitement automatique des langues" that then became NLP.

[11] *délégation*

[12] https://www.emlex.phil.fau.eu/

with Master's programs. Thus we have set up an exchange protocol that allows students from the Ecole des Mines to join the Master's program during the last year of the program.

All these elements made the Université de Lorraine a suitable environment for developing the master's program in NLP.

2 Definition of the NLP program

The first question is to define for whom the program is intended. To do this we went back to the ambition we wanted to give to the program. Once the objective was clarified, we were able to work on defining the content of the program.

2.1 Program's objectives

The first observation made in 2016 was that there was an effective increase in the need for NLP, both for industrial issues and for the development of research. At that time, a very strong dynamic was already in place, driven by AI and deep learning.

We realized that the training we had developed until then was attached to a more traditional definition of NLP, rooted in computer science and linguistics. Its objectives being mainly to accompany students towards research, this was not very surprising, but it proved to be out of step with scientific and societal evolutions.

The project took shape in the idea of considering that students should leave the Master's program for industry as much as for research. We already considered that a significant part of the companies concerned were also concerned with research issues. This shift towards more applicative issues also seemed necessary to us to be able to integrate more students into the training.

Moreover, our experience with the second year of the Master's program showed us that the students who successfully completed the program could have very heterogeneous profiles. The challenge for us was more to define the type of profile that we wanted to find at the end of the training and to propose paths to bring the students there. This being the case, in view of the size of the classes, it was not possible to multiply the paths and above all, it seemed important to us that the profiles be built in interaction with each other. To achieve this, we need to have a precise vision of the type of training we can accept.

We have therefore chosen to build a training program centered on computer science and mathematics for the NLP that is as open to research issues as it is to applications.

2.2 Methodology

To build the program, a project manager was appointed in 2015 by the teaching component in the person of Maxime Amblard with the task to design a project to be presented to the University in June 2016. Some others are used are underpinning as (Bender et al., 2008).

The project manager made the choice to start from scratch, considering that the contents present at that time needed to be renovated. He set up a working group gathering members of the Loria and ATILF teams working on language data who wanted to invest in the new training. The aim was to integrate new colleagues, new views and possibly new themes/issues.

A shared space was set up online allowing all parties to access the working material as it became available. The working group started with 13 people and ended up with 25 people. It was decided to meet the equivalent of once a month until July 2026. After each meeting, the leader wrote a report that was sent to all members, as well as a doodle to choose the date of the next meeting and the agenda of this meeting. The participants thus had the agenda well in advance, which allowed them to prepare the meeting as well as possible or to send their points of view and their work in case of absence. This methodology ensured that no information was lost and that the work dynamic was maintained throughout the year. In addition, meeting times were strictly adhered to. We return to the major orientations in the following section.

Higher education and research are undergoing numerous transformations. This is obviously the case in France. In particular, at that time, major programs were being launched and it was important to position training in this ecosystem. In the end, we quickly dismissed this issue, considering that the constitution of a scientifically solid program would always be easier to defend.

After several discussion sessions, we came to conclusions:

- the environment has many competences as well in data processing as in linguistics, which moreover with habits of work in common

- the Erasmus Mundus LCT obliged us to teach in English, which is not always easy in France,

but this should be a strength for the recruitment of international students

- we had a scientific positioning neither in computer science nor in linguistics, clearly based on mathematics and computer science
- it is now possible to award a French diploma in NLP
- the needs both in the industrial and academic world are important

We have also considered two hypotheses on the organization of the training: the non-opening of the training to distance learning because it requires another type of organization that we cannot propose for now, and the possibility of carrying out the training in alternation. This program is a French specificity which allows students to study while being employed by a company. Thus, the training alternates (in the literal sense) periods of training at the institute and periods of work in a company.

One important aspect is to have concluded that if we have many strengths in the field of NLP, we have the specificity of covering a wider field of language data processing. If we are not the only ones in France, we wanted to make a specificity of our mathematic and computer science positioning in the field of language data processing. We have therefore decided to offer a curriculum entitled "Computer Science, Speech, Text and Knowledge". Thus, the new program deals with speech processing, NLP and knowledge. It is obvious that the reconciliation of formal and statistics tools operated in the last ten years facilitates this closeness.

2.3 Broad Program Directions

We have therefore chosen to integrate students whose initial training is either in computer science or linguistics, with an appetite for formalization, programming and mathematics. A single pathway is proposed through the training that brings together these different profiles. The objective is to bring all students to the same exit point. Beyond this declaration of intent, it is obvious that we cannot transform in two years linguists who have never done programming into operational computer scientists, just as we cannot transform computer scientists into linguists in the same time. On the other hand, students must be comfortable enough to overcome the paralysis of working in a new field, to go beyond naive approaches and above all to be able to discuss the different aspects precisely with specialists in the other field. To achieve this, there is nothing better than to mix these profiles throughout the training.

This mixing adds entropy to the construction of individual paths. To compensate for this, we have proposed a very legible and regular architecture in the training. This apparently very rigid organization allows students to build their path together.

Moreover, as we have already mentioned, the aim is to offer a course with a strong research component, while at the same time offering numerous applications.

It was therefore decided that the first year would serve as a substrate to establish the background by opening up to the problematic of long time. The second year would focus on NLP topics, with many more applications and especially a significant amount of time devoted to an internship either in a laboratory or in a company. The training is given over two years, i.e. 4 semesters:

- two semesters of the first year of 260 effective teaching hours each - 30 credits each
- one long semester of the second year of 350 effective hours of teaching each - 30 credits
- one internship of at least 5 months (one semester) - 30 credits

The architecture of each of the three semesters is the same with 5 teaching units of equal importance in the validation of the training. The organization of studies in France implies that each semester validates 30 credits (ECTS).

2.4 Renewal of Teaching Practices

As we have already mentioned, we did not want to transform our teaching practices by switching to distance learning, especially because we thought it was important to have different profiles working together. It is difficult to measure this at a distance.

However, we questioned these practices to propose more applied teaching or with effective data manipulation, as well as the implementation of group work, in particular in the form of projects.

Concerning the discovery of research, we propose a project in groups of 2 to 4 students, called "supervised project". This is a project carried out during the first year, at the initiative of a researcher, and which leads to the writing of two reports. The first part of the year is a bibliographic work. It

consists of taking the time to read and understand scientific articles on the state of the art and to put them in perspective with the proposed project. The work is synthesized in a bibliographic report which opens to the second part of the work which is a more classical realization part. The students produce a report, which is a first experience of long writing before their final report of master, as well as a defense of 20 minutes in front of a jury and the presentation of a poster. The objective is to have them carry out a research project over a long period of time, with links to the literature on the one hand and actual achievements on the other, and also to master the codes of scientific presentation. This work often leads to publications in international conference workshops.

As an example, here are some titles of projects carried out in recent years on different themes: Speaker Adaptation Techniques for Automatic Speech Recognition, Testing Greenberg's Linguistic Universals on the Universal Dependencies Corpora using a Graph Rewriting tool, Does my question answer your answer, Anomaly detection with deep learning models, ...

The counterpart of this work is the realization of an applied project, in group, also on the long time. It is set up in the second year by groups of 3 to 4 students. The initiative is left to the students, who are nevertheless supervised by two teachers. Once the application is identified, the students implement the concepts they have encountered in their training to produce an application. The functional and deployable applications are put online to showcase the work done. The students also produce a report explaining their approach and their achievement, and they make a defense. Here some examples of realization subjects:

- GECko+ is built on top of two Artificial Intelligence models in order to correct spelling and grammar mistakes, and tackling discourse fallacies with BERT (Devlin et al., 2018).

- Askme is a Question Answering model built on the Stanford Question Answering Dataset (SQuAD) with a fine-tuned BERT. Askme is able to automatically answer factual questions without being aware of the context.

- IGrander Essay: Automated Essay Grading systems provide an easy and time-efficient solution to essay scoring.

- Multilingual multispeaker expressive Text-to-speech system: The main goal of this work is from text input to be able to generate speech with expressivity for multiple languages, which are currently French and English, with an end-to-end multilingual text-to-speech (TTS) system.

The first experiences have shown that both formats are very formative. It is common for groups of students to go from very enthusiastic to overwhelmed phases. In both cases, although very different from each other, the situation allows them to better understand where the interests of NLP are and especially where the difficulties are. Working over a long period of time shows them the importance of anticipating problems in both research and development. We make sure that the groups are mixed in terms of profiles to avoid the pitfall of groups stuck on IT developments, as well as groups missing out on linguistic issues. We note that an additional exercise is paper writing in the format of the main conferences in the field.

2.5 Support for internationalization

As part of the development of French universities, the gouvernement has set up the development of major projects, under the name Project Investment of the Future (PIA). The Université de Lorraine benefits from a major project of this type called Lorraine University of Excellence (LUE). This project covers several themes around systems engineering. The training program has benefited from financial support from this program to support internationalization. For this purpose, we are translating the presentation documents into several languages, in particular into Russian, Persian, Greek and Turkish. The objective is to accompany as effectively as possible the arrival of students in the training.

In addition, we have made a promotional film for international students to highlight the necessary complementarity between computer science and linguistics that is achieved within our training. The production was made in a professional way by relying on the students and their profile. Once again, the aim is to highlight the diversity and quality of the students' profiles.

3 Program Design

3.1 Education

It is not a question here of making an advertising brochure of the training, also we do not give explic-

itly the titles of each teaching unit. This being said, the training is thought to put forward continuum between the semesters, as much as possible on the whole training, at least between the semesters. It is this dynamic that we put forward. We invite the readers to refer to the descriptions of the training for more details. We have the three semesters with regular teaching. Units of the first semester are numbered 70X, ones of the second semester with 80X and those of the last semester with 90X, where X takes its value in $\{1, 2, 3, 4, 5\}$

The **first units** gathers the fundamental teachings in mathematics and computer science. For this unit, the continuum is natural: on one side probabilities and statistics - machine learning - neural network; on the other one programming - semantic web - data mining/recommendation

701 Probabilities, Statistics and Algorithms for AI: The course deals with fundamental mathematical tools, particularly statistical and algorithmic tools, which are necessary to define and resolve an artificial intelligence problem. The course unit stresses a case study approach in order to ensure the acquisition of theoretical aspects and practical application (Elementary mathematics tools, propabilities and statistics; and Python programming, both for beginners and advanced users)

801 Machine Learning and Semantic Web: This course takes on the fundamental principles of Machine Learning, of data mining and knowledge extraction. All the notions are illustrated through practical applications on real data (Machine learning theory and Web Semantic)

901 Deep Learning and Data Mining: The objective of this course unit is to acquire machine learning tools, mainly deep neural networks and factorisation matrices, to be able to manipulate these tools when considering practical applications (tweets, traces of e-learning), as well as to expand the students' knowledge in semantic web and the extensions of analysis of formal concepts for textual and relational data processing (Neural Networks, Deep Neural Networks; Data mining (structured data and text); and Collaborative filtering)

The **second units** gathers the teachings around corpora and formal toolsFor units two and three in the second year, we make sure to offer state of the art teaching applied to language data..

702 Design and Acquisition of corpus: This course aims to introduce techniques of construction, structuring, annotating and archival of textual oral or multimodal corpora, which play an essential role in the analysis of the structure of spoken and written language, and on the other hand in the training and evaluation of NLP algorithms. This subject is complex as it is necessary that (1) the corpora be restricted to a reasonable size in order to guarantee the proper collection of corresponding data and that (2) this data sufficiently represents the phenomena studied (Written Corpora; and Spoken Corpora).

802 Formal Tools: This course unit is dedicate to the introduction to theoretical frameworks and logic used in symbolic approaches to the modeling of language. It consists of mathematical logic and of formal languages. The objective is to familiarize the students with these formal models, their properties, the demonstration techniques associated with them, and the notions of calculability and complexity.

902 Text and Speech Processing: Automatic processing of texts and speeches involves different methods of machine learning. This class will introduce these methods and illustrate their use through examples and practical application using tools developed (Speech processing, Processing Textual Data, Terminology and ontology).

The **third units** gathers the teaching on software engineering and data sciences.

703 Software Engineering: Collection, analysis and formalization of customers' needs (Software design and; Functional analysis; specifications; and Project management)

803 Data Science: This course unit introduces fundamental techniques for the extraction, storage, cleaning, visualisation and analysis of data. We give a practical introduction to the tools and software libraries which allows the processing of data. We combine theoretical sessions with programming exercises which allows students to put into practice the software and concepts taught during the course.

903 Natural Language and Discourse: This course unit gathers courses which deal with discur-

sive and semantic processing of language (Application to texts; Computational semantics; and Discourse and Dialog modelling)

The **fourth units** is that of linguistics.

704 Linguistics for NLP-1: This course takes on one hand the fundamental elements of NLP and on the other hand the phonological and morphological elements, which are studied through a language sciences based approach (Methods for Natural Language Processing; Phonology; and Morphology)

804 Linguistics for NLP-2: This course takes up where the previous teachings of linguistics left off, wherein the content and methodologies of the courses are, however, independent from these prior teachings. The courses are concentrated on syntax and semantics, as well as lexicology. In this context, the focus is put on the question of formalisation of linguistic rules (Lexicology: lexical units and phraseology; Syntax; and Semantics)

904 Lexicon and Grammars for NLP: This course introduces advanced tools for the computational modeling of different types of linguistic information which describe lexical units (lexical resources) or the rules of organisation of larger units (grammar) (Diachronic and synchronic lexicology; Lexical resources; and Syntactic framework)

Finally, the **fifth units** is the project-based teaching that we have detailed above, to which we add language teaching (French for non-French speakers, English for the others). In the second year, we add opening lessons, in particular around ethical issues.

705 This course unit is composed of the first part of the year-long supervised project (which will be finalized in the second semester), as well as language classes. The project consists of group work (in pairs), which is supervised by researchers, in which the students will carry out the bibliographic part of the final report. Language classes will allow non-anglophone students to become more familiar with scientific english, the language in which all courses are conducted, whereas the french classes will facilitate non-francophone students' social and cultural integration. The course allows students to test their first skills acquired during the semester while synthesizing the different research issues concerned by a more open research topic. Students are evaluated on the acquisition of targeted skills, identified for the Sc. Master.

805 This course unit is made up of language classes as well as the second part of the year-long supervised project (UE 705) which is finalized during the second semester. The second part of the year-long project begins by following through with the procedure introduced in the first semester groupwork, and leads up to the presentation at the end of the year, explaining the implementation of the project, which puts to test all of the skills the student acquires during the first year of the program. The language classes allow students to become more comfortable with the knowledge of scientific english.

905 Projects and Foreign Language: This course unit gathers many teaching including the project and language classes (Software project; Law and ethics; Research methods; Professional integration; Foreign language courses (French or English))

To compensate for the differences in levels, the technical courses of the first semester are accompanied by a refresher course.

The teachers are free to choose the teaching methods they follow, within the general perspective of the program. We clearly share the objectives for the end of the program where students are autonomous in dealing with the scientific literature and in participating to produce new results.

3.2 Example of a course

It is obviously not possible to describe the all courses of the training and it is not the object of this article. We want to highlight one course in particular which allows us to put forward the principle of training by project by mixing profiles.

This is **Methods4NLP** given in unit 704. It is one of the very first courses introducing NLP to students. The teaching is divided into several sequences: a first one allowing to set up a common culture around NLP, a second academic one presenting the basic formalizations that they will develop throughout the two years of training, and the setting up of a project.

The first phase can be divided into three parts:

- a first phase of seminar to set the spectrum of NLP, from linguistic aspects to technical developments. This teaching allows to give a common culture to the students at the beginning of the program by positioning NLP in relation to the challenges of AI.

- a phase of apprehension of the concepts by experimentation with unplugged activities (Bell et al., 2009)(Romero et al., 2018): one simulating machine learning with a machine playing Nim's game, the other on Hoffman coding with groundhogs. The students meet the two paradigms of NLP in an intuitive way.

- first experiments to highlight the possibility of getting results with few developments with two labs: one on FastText (Bojanowski et al., 2016)(Joulin et al., 2016) and theone on Sckitlearn (Pedregosa et al., 2011), are proposed. The objective is to make all students aware of the ease of manipulating data without understanding what is behind it, and the difficulty of confronting how to improve the results.

The two other phases are carried out in parallel. For the project part, the students form groups of 4 in a balanced way between the profiles. It is requested that the students of the same nationality do not stay together and especially that all the groups have explicit different profiles (linguist and computer scientist). The groups choose the subject of their development. They are only required to explicitly include NLP aspects in their project (POS tagging, Named Entity recognition, Machine learning, etc.).

This project is shared with another course that deals with written corpora. For this course, the project must contain the constitution of a corpus, its normalization, as much as possible its annotation with study of the quality of the annotation, and thus implementation in a defined context. Thus, the themes of the projects are more naturally related to NPLP than to speech processing, or even to knowledge processing.

For information, here are a few topics chosen by the students: Gender bias in young targeted literature: 19th century vs. early 21st century, Author Identification for Philosophers, Song Lyrics Generator, Autonomous Vocabulary Assistant, ...

It is interesting to let the students choose a topic that interests them. Sometimes we see students with very specific subjects, letting them work on them allows us to build on their engagement. If the subject is not relevant, it allows us to put them back into a more appropriate training perspective. There is no such thing as a bad topic. The different topics make students aware of the difference between the desire to achieve a development and the possibility of achieving it. In general, we see that some groups prefer to stay close to a very classical topic and often regret not having explored more diversity of topics. Moreover, a recurrent element appears on the question of the evaluation of development. This practice makes them aware of the significance of anticipating the implementation of the evaluation in order to measure the quality of the final project.

Students are monitored weekly. They have to make a 1 minute presentation of the progress of their developments in front of the whole class. They may not have made any progress, what is important is that all students follow the progress of all groups. During the Covid-19 period, the follow-up was done by videoconference and it was difficult to share this experience with the whole class. During the semester, students submit 3 progress reports, the first ones being only a few pages long. They are used to document problems, issues and progress. The final report is due before the exam period during which a defense is organized.

Finally, the last part consists of more academic teaching that takes up the two main paradigms of NLP and defines them explicitly (Agarwal, 2013). The first part deals with out-of-context grammars and automata. These concepts are often known by students more or less well. This allows on the one hand to bring everyone up to speed and on the other hand to show how basic concepts in computer science find links with linguistics. This is done by looking at Turing's work on abstract machines, or through Chomsky's hierarchy. Then the course highlights the use of statistical techniques by explaining Bayes' models and automatic classification as done by Jurafsky (Jurafsky and Martin, 2018). Finally, the last part explains dynamic programming by focusing on the syntax with the algorithm of CKY (Kozen, 1977). This part of the course is carried out in a traditional way with plenary teaching and tutorials.

This teaching is the course which launches the training. Indeed, if it begins with a phase requiring little knowledge a priori, it continues with a phase that requires the mobilization of technical knowledge delivered in the other courses, as well as linguistic knowledge. The fact of leaving the choice

of the theme allows to motivate the students, while showing them the need to be autonomous and proactive in the training. In addition, a very traditional part of the course allows students to establish common ground by building epistemological bridges between computer science and linguistics.

4 Analysis of student profiles

After having presented the structure, the organization and the stakes of the training, we propose to come back on quantitative elements concerning the disciplinary profile of the students, the diversity of their origin and their success. In order to give a little more perspective to these elements, we rely on the data of the Master since 2018, that is to say 3 academic years, as well as on the M2 TAL speciality over the 3 previous years.

Before the definition of the program, we had made a study on the 2008-2017 classes. We noted that 36% of the students were French, 45% from another European country and 19% from a non-European country, which is proof of a good diversity of origin. Of those who continued their studies after the course, 40% joined a company, 33% were doing a thesis and 27% had an academic post.

The table 1 presents the evolution of the number of candidates and students in the training before the creation of the Master (3 years) and for the first 3 years of the training. It also contains the distribution of students according to the profiles (Fosler-Lussier, 2008): computer science, mixed or linguistics. We observe that the distribution is homogeneous over the different years of training, with an average of 41.6% computer scientists over the first three years of the Master's program for 33.5% linguists. This good distribution makes it possible to build balanced work groups. Moreover, we observe that the opening as a Master has significantly increased the average number of this profile from 27 to 33.5. This implies reinforcing the course for this type of profile and increasing the means implemented on their programming skills.

Another interesting element that appears in this table is the increase in the number of candidates, with 117 candidates for the M1 year in 2020-21. It should be noted that this number of applicants does not include students trained within the IDMC or Erasmus Mundus LCT students who are selected through another process. For M2 students, only a few students are selected, the majority coming from the first year of training. Our experience leads us to limit even more the number of students entering the second year directly, because on the one hand these students have gaps that they are unable to fill, and on the other hand they have difficulties integrating the class. These elements must of course be put into perspective because the effects of the health crisis must also be taken into account.

Finally, the other axis of analysis proposed is to look at the nationalities of the students. It can be seen that the average number of French students has decreased significantly, from 44% to 34.5%. This decrease did not affect the rate of oversea students, keeping their number slightly above 50.

Finally, an important element is to analyze the profiles at entry with the success at the end of the Master's degree. The first element is that the students who follow the two years of training have better results and rankings. This supports the idea that the training has singular characteristics that are difficult to catch up on. This is understandable because it is necessary to master concepts in computer science and linguistics, and moreover it is necessary to master the tools and methods of AI, which are characteristic and rather abstract.

One should not believe that a profile at the entrance would be privileged in this type of training. The ringleader or the first ones of the promotions do not have *a priori* determined profiles. They may come from traditional computer science or linguistics. What seems to make the difference is, in addition to their aptitude at entry, their investment in the training throughout the two years. This reassures us that the aim is to bring together the profiles, while respecting their original specificities.

5 Conclusion and perspectives

We highlighted the process of creating the Master's degree in NLP at the IDMC. This training is unique in that it delivers a French diploma in NLP. The students who enter the program have a background either in computer science or in linguistics. The training aims to give them strong skills to train future professionals in language data processing.

We have integrated collaboration with several components in our environment (Erasmus Mundus LCT, EMLEX, Ecole des Mines, etc.). The training is attractive and has good results. The students trained in this way are currently making choices for their careers rather than being affected by the economic situation, despite the health crisis.

Over the different years observed, we note that

	# applic.	# students	Comp.	Mixed	ling	Fr	Eu	World
2015-16 M2	35	17	41	24	35	24	12	65
2016-17 M2	20	6	67	17	17	67	0	33
2017-18 M2	33	17	47	18	29	41	0	59
Mean	29,33	13,33	51,66	19,66	27	44	4	52,33
2018-19 M1	47	16	25	38	38	50	13	38
M2	21	7	57	29	14	29	14	57
2019-20 M1	65	20	30	25	45	35	30	35
M2	32	20	45	20	35	30	0	70
2020-21 M1	117	33	45	18	36	36	12	52
M2	33	27	48	19	33	26	15	59
Mean M1	76,33	23	33,33	27	39,66	40,33	18,33	41,66
Mean M2	28,66	18	50	22,66	27,33	28,33	9,66	62
Mean	52,5	20,5	41,66	24,83	33,5	34,33	14	51,83

Table 1: Distribution of students in the training over the last 6 years: number of candidates, number of students, then distribution of students in percentage (%) between the 3 profiles, then according to their nationality.

the number of internship proposals on the one hand and job offers on the other hand are constantly increasing. We see the search for profiles explicitly in NLP. Moreover, our opening towards more applied profiles has not been at the expense of the relationship with the academic world. Out of the first class that followed the 2 years of the training, 7 are pursuing a thesis which is very positive. These two dynamics reinforce our commitment to the development of the program.

For the future, we want to consolidate the training by slightly increasing the flow of students until we reach 40 students per year, especially with students from the European area. In addition, we are working on developping an alternation for research, which would allow us to offer training courses of very good quality towards PhD.

References

Apoorv Agarwal. 2013. Teaching the basics of NLP and ML in an introductory course to information science. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 77–84, Sofia, Bulgaria. Association for Computational Linguistics.

Tim Bell, Jason Alexander, Isaac Freeman, and Mick Grimley. 2009. Computer science unplugged: School students doing real computing without computers. *The New Zealand Journal of Applied Computing and Information Technology*, 13(1):20–29.

Emily M Bender, Fei Xia, and Erik Bansleben. 2008. Building a flexible, collaborative, intensive master's program in computational linguistics. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, pages 10–18.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eric Fosler-Lussier. 2008. Strategies for teaching "mixed" computational linguistics classes. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, pages 36–44.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Daniel Jurafsky and James H Martin. 2018. Speech and language processing (draft). *Chapter A: Hidden Markov Models (Draft of September 11, 2018). Retrieved March*, 19:2019.

Dexter C Kozen. 1977. The cocke—kasami—younger algorithm. In *Automata and Computability*, pages 191–197. Springer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Margarida Romero, Benjamin Lille, Thierry Viéville, Marie Duflot-Kremer, Cindy de Smet, and David Belhassein. 2018. Analyse comparative d'une activité d'apprentissage de la programmation en mode branché et débranché. In *Educode-Conférence internationale sur l'enseignement au numérique et par le numérique*.