Unsupervised word discovery for computational language documentation

Pierre Godard

université PARIS-SACLAY

NNT : 2019SACLS062

Thèse de doctorat

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Unsupervised Word Discovery for Computational Language Documentation

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°580 Sciences et technologies de l'information et de la communication (ED STIC)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Orsay, le 16 avril 2019, par

PIERRE GODARD

Composition du Jury :

Pierre ZWEIGENBAUM Directeur de Recherche, CNRS (LIMSI)	Président
Christophe CERISARA Chargé de recherche, CNRS (LORIA)	Rapporteur
Adam LOPEZ Professeur associé, University of Edinburgh (ILCC)	Rapporteur
Emmanuel DUPOUX Directeur d'Études, EHESS (LSCP)	Examinateur (absent)
Laurent BESACIER Professeur, Université Grenoble-Alpes (LIG)	Co-encadrant
François YVON Professeur, Université Paris-Sud, CNRS (LIMSI)	Directeur de thèse

Abstract

Language diversity is under considerable pressure: half of the world's languages could disappear by the end of this century. This realization has sparked many initiatives in documentary linguistics in the past two decades, and 2019 has been proclaimed the International Year of Indigenous Languages by the United Nations, to raise public awareness of the issue and foster initiatives for language documentation and preservation. Yet documentation and preservation are time-consuming processes, and the supply of field linguists is limited.

Consequently, the emerging field of computational language documentation (CLD) seeks to assist linguists in providing them with automatic processing tools. The *Breaking the Unwritten Language Barrier* (BULB) project, for instance, constitutes one of the efforts defining this new field, bringing together linguists and computer scientists. This thesis examines the particular problem of discovering words in an unsegmented stream of characters, or phonemes, transcribed from speech in a very-low-resource setting. This primarily involves a segmentation procedure, which can also be paired with an alignment procedure when a translation is available.

Using two realistic Bantu corpora for language documentation, one in Mboshi (Republic of the Congo) and the other in Myene (Gabon), we benchmark various monolingual and bilingual unsupervised word discovery methods. We then show that using expert knowledge in the Adaptor Grammar framework can vastly improve segmentation results, and we indicate ways to use this framework as a decision tool for the linguist. We also propose a tonal variant for a strong nonparametric Bayesian segmentation algorithm, making use of a modified backoff scheme designed to capture tonal structure. To leverage the weak supervision given by a translation, we finally propose and extend an attention-based neural segmentation method, improving significantly the segmentation performance of an existing bilingual method.

To those opening pathways,
in grateful memory of Isabelle Tellier.

Acknowledgments

I am particularly indebted to Professor François Yvon for giving me the opportunity, time and space, to make this research. His trust despite my struggles, his deep and broad scientific knowledge as much as his appetite for new ideas, his high standard of expectations and relentless commitment for improvement (he is the kind of person sending you annotated pdfs mere hours after having been hit by a car), gave me the inspiration to insist when I was in doubt and swallowed by the requirements of my other activities. I greatly admire his scientific brilliance and integrity.

I am equally and deeply grateful for Professor Laurent Besacier's supervision; without his passion for computational language documentation, insatiable curiosity and enthusiasm, and seemingly effortless sunny disposition, I probably would not have been able to complete this work. I will miss our weekly videoconference meetings and the pleasure of learning under his guidance, but I believe our conversations will continue in other ways. Our time together at CMU in Pittsburgh will also stay with me as a manifold experience and a very happy memory.

I want to warmly thank the members of the jury for their careful examination of my work, and their comments and questions on the day of my defense: Professor Pierre Zweigenbaum who presided the jury, Professor Emmanuel Dupoux who inspired and led a project that I was lucky to join during the 2017 Jelinek Summer Workshop on Speech and Language Technology (JSALT), and referees Professor Christophe Cerisara and Professor Adam Lopez who generously committed time in their very busy schedules to write sharp and enlightening in-depth reports on this work. I am genuinely honored, humbled, and grateful.

Alexandre Allauzen deserves a special mention for answering many of my questions, especially during our trips in the infamous RER back from Orsay. I am proud and fortunate to now call him a friend. I also have to give a special thank you to Hélène Maynard who accompanied me at the beginning of this work and was the most welcoming and open force in the laboratory. She often cleared my doubts about my capacity to conduct this research while still being involved in making art. I additionally enjoyed very much my time sharing an office with Guillaume Wisniewski in what was then the 'Bâtiment S', and our conversations about life or dance, as much as those more closely related to the topics discussed here.

The BULB project, funded by French ANR and German DFG under grant ANR-14-CE35-0002, centrally supported and motivated this work. It fostered many collaborations and I want to address my warmest appreciation firstly to Gilles Adda, Annie Rialland, and Martine Adda-Decker, from whom I learnt a lot and received much benevolence and support, as well as to Lori Lamel, Jamison Cooper-Leavitt, Joseph Mariani, and Sebastian Stüker. I also want to specifically acknowledge the work of Guy-Noël Kouarata and Odette Ambouroue in the collection of the Mboshi and Myene data, as well as the stimulating conversations I had with Mark Van de Velde, Dmitry Idiatov, Fatima Hamlaoui, and Élodie Gauthier. The JSALT workshop mentioned above was also very formative and inspiring in many ways, and I would like to wholeheartedly thank Graham Neubig, Florian Metze, and Alan Black, who embody kindness and the spirit of research at its best. I also enjoyed very much getting to know, and working with, Rachid Riad, Lucas Ondel, and Markus Müller during that period.

Even though I cannot acknowledge properly and in detail all the people who provided insightful comments, suggestions, or help, during this PhD, I very much want to thank Kevin Löser and Marcely Zanon Boito for our fruitful collaborations, and Alexandre Bérard who offered me the most precious help. I also received generous help from Philip Arthur and Matthias Sperber. To some of the past and present researchers and PhD students at LIMSI in the 'TLP group' led by Jean-Luc Gauvain who shared some of their knowledge or their time with me — Marianna Apidianaki, Éric Bilinski (who saved the day so many times), Ophélie Lacroix, Hervé Bredin, Philippe Boula De Mareüil, Nicolas Pécheux, Yong Xu, Gong Li, Benjamin Marie, with a special mention to the kernel of students I got to share most of the ups and downs of this journey with, Matthieu Labeau, Lauriane Auffrant, Elena Knyazeva, Rachel Bawden, Julia Ive —, and also to Aurélien Max, Thomas Lavergne, and Anne Vilnat: merci ! I hope that our paths will cross again soon. And to a whole new generation of PhD students, Yuming Zhai, Syrielle Montariol, Léo Galmant, Margot Lacour, Aina Gari Soler, Benjamin Maurice, Anh Khoa Ngo Ho, Minh Quang Pham, José Rosales, Pooyan Safari, Ruiqing Yin: my best wishes for your future endeavors.

This is an already very long list of names, and I apologize for those I forgot despite my best efforts, but I would like to extend my gratitude to some of the many wonderful people at LIMSI who always helped me with extreme kindness: Laurence Rostaing, Jean-Claude Barbet, Pascal Desroches, and Sophie Pageau-Maurice.

Lastly, and most importantly, I address a loving thank you to Liz, Anna, Mélanie, Cynthia, Anne, Yves, Marie, Jeanne, and to many friends, for your unconditional support and patience when cumulated work made me more absent than I had wished.

Contents

List of Figures

List of Tables

List of Acronyms

AER	Alignment error rate
AG	Adaptor Grammar
ASR	Automatic speech recognition
BF	F-measure on word boundaries
BP	Precision on word boundaries
BR	Recall on word boundaries
BULB	French-German *Breaking the Unwritten Language Barrier* project
CLD	Computational language documentation
CRP	Chinese restaurant process
DP	Dirichlet process
GRU	Gated recurrent unit
HDP	Hierarchical Dirichlet process
ITG	Inversion transduction grammar
LF	F-measure on word types
LP	Precision on word types
LR	Recall on word types
LSTM	Long short-term memory
MDL	Minimum description length principle
MT	Machine translation
NMT	Neural machine translation
NSE	Normalized segmentation entropy
PCFG	Probabilistic context-free grammar
PYP	Pitman-Yor process
RNN	Recurrent neural network
SMT	Statistical machine translation
UL	Unwritten language
WF	F-measure on word tokens
WL	Well-resourced language
WP	Precision on word tokens
WR	Recall on word tokens

Prologue

We usually think of science as an activity pursuing an apolitical and objective effort to uncover the truth. Art, on the other hand, often finds its impetus in the turmoil of the social world, filled with subjective preferences and partisan conflicts. Of course this is oversimplifying, and in the case of natural language or speech processing, I believe the line to be particularly blurry: it is safe to say that technology, and particularly language technology, now shapes — or significantly impacts — the life of billions of human beings.

This dissertation was written to be defended in computer science, a "hard science" discipline. Consequently, I tried my very best to follow the standards of rigor and objectivity required by this framework. An important perspective, as far as I am concerned, would be lost however if I did not make the liminary mention that I became involved in this research as an artist questioning aesthetic and political implications of language technologies, in order to gain the capacity to make better informed choices, be able to voice subjective concerns or hopes, and ultimately try to create new forms on stage throughout this process.

Computational language documentation represents, to my eyes, the best of what such technologies could provide to our societies: protection, diversity, inclusion, knowledge and open-mindedness. However, clever models and ideas often end up being used for things like efficient targeted advertising or the breaking into citizens' privacy. I form the wish that the research community will continue to support undertakings such as computational language documentation in the future, and will stand strong to safeguard ourselves from the misuse of many of the powerful techniques currently being developed to analyse and process language data.

Le rêve : connaître une langue étrangère (étrange) et cependant ne pas la comprendre : percevoir en elle la différence, sans que cette différence soit jamais récupérée par la socialité superficielle du langage, communication ou vulgarité ; connaître, réfractées positivement dans une langue nouvelle, les impossibilités de la nôtre ; apprendre la systématique de l'inconcevable ; défaire notre « réel » sous l'effet d'autres découpages, d'autres syntaxes ; découvrir des positions inouïes du sujet dans l'énonciation, déplacer sa topologie ; en un mot, descendre dans l'intraduisible, en éprouver la secousse sans jamais l'amortir, jusqu'à ce qu'en nous tout l'Occident s'ébranle et que vacillent les droits de la langue paternelle, celle qui nous vient de nos pères et qui nous fait à notre tour, pères et propriétaires d'une culture que précisément l'histoire transforme en « nature ».

Roland Barthes
L'empire des signes, 1970

Chapter 1

Introduction

Contents

Contributing to a response to language endangerment, a large-scale issue with dreadful consequences, motivates the work presented in this thesis. We give some elements to measure the magnitude of the problem, and why it matters. We then describe the emergence of a new field of research we call *computational language documentation*, and a French-German initiative contributing to define this field. Lastly, *unsupervised word discovery* – the scope of our work – is informally defined, and our main contributions are given, as well as an outline of the thesis.

1.1 Motivation: language endangerment

In 1992, Michael Krauss concluded a contribution (Krauss, 1992) to a special issue of the journal *Language* (Hale et al., 1992) with an alarming call:

> *Obviously we must do some serious rethinking of our priorities, lest linguistics go down in history as the only science that presided obliviously over the disappearance of 90% of the very field to which it is dedicated.*

As Krauss himself acknowledged then, it is difficult to estimate robustly the magnitude of a phenomenon some authors have called *language death* (Crystal, 2000; Harrison, 2007). Crystal (2000) and Nettle and Romaine (2000) suggest instead that "only" 50% of the world's language could disappear by the end of this century.

This is a threat to the world's complex and diverse linguistic landscape. Lewis (2015, cited by Romaine (2017)) counts 7,102 languages (among which 137 sign languages) in use throughout the world. The top 20 languages are spoken by 50% of the world's population (Austin and Sallabank, 2011), but less than 1% of the population accounts for 55% of the world's spoken languages (Romaine, 2017). Geographical disparity is also strong: Africa and Asia both bear 30% of the world's languages, America, 15%, and Europe, 4% (Romaine, 2017). The data in the 21st edition of *Ethnologue* (Lewis, 2018) indicate that 45% of known languages are unwritten.[1]

1.1.1 Magnitude of the issue

Discussing the causes for endangerment is beyond the scope of this thesis. Natural disasters, climate change, famine, disease, war and genocide, repression or assimilation, and various factors of dominance (culture, politics, economy) seem to be the most important identified causes (Austin and Sallabank, 2011). The role of globalization is also debated amongst linguists. But when should an expert declare that a language is officially *endangered*? This requires a somewhat empirical weighing of various criteria, for instance the number and age of native speakers, the state of intergenerational transmission,[2] the domains of use, or the presence of an ongoing language shift.[3] This has led Lewis and Simons (2010) to elaborate the Extended Graded Intergenerational Disruption Scale (EGIDS), which aligns several scales of endangerment – Fishman's 8-level scale (Fishman, 1991), a 6-level scale developed by UNESCO, and a 5-level scale previously used by *Ethnologue* (e.g. Lewis, 2009, 16th edition) – in a 13-level scale.[4] According to Romaine (2017), with data from (Lewis, 2015), 66% of the world's languages are vital (level between 0 and 6a), but 34% are endangered or dying (EGIDS scores between 6b and 9). A simpler endangerment threshold, set at 100,000 speakers for any given language, would however lead to a much higher estimate of the world's languages being at risk (80%). It also seems undoubtful that the rate of language extinction is on the rise. In a resolution adopted in 2016 by the United Nations General Assembly, 2019 was proclaimed as the International Year of Indigenous Languages, to increase public awareness of the issue and foster collaborations to "promote and protect indigenous languages and improve the lives of those who speak them".[5]

[1] It is unclear, though, how often the existing writing systems are used in the remaining 55%.

[2] For UNESCO's *Atlas of the World's Languages in Danger of Disappearing*, when less than 30% of young people learn a language, it should be considered endangered (Romaine, 2017).

[3] When a community of speakers shifts to a different language, e.g. the shift to French from most of the Alsatian-speaking community after World War II.

[4] Succinctly, from 0 to 4: institutional; 5: written; 6a: vigorous; 6b-7: in trouble; 8a-9: dying; 10: extinct (Lewis and Simons, 2010; Romaine, 2017).

[5] `https://en.iyil2019.org/`.

1.1.2 Consequences of language loss

Many aspects can be invoked as to why we should care for the consequences of language extinction, and why linguistic diversity matters. One central reason is that, when a language stops to be spoken, a vast network of cultural knowledge, creativity, and relationship to the environment becomes lost, especially for undocumented and/or unwritten languages. Oral literature – songs, legends, historical accounts – vanish. Harrison (2007) makes a case for the broad loss that comes with the extinction of small languages: specific knowledge to interact with animals or plants, singular counting systems and relationships to time and space, as well as an access to the creativity that allowed speakers to encode these knowledges into linguistic structures. The author recalls linguist Ken Hale once telling a reporter: "When you lose a language, you lose a culture, intellectual wealth, a work of art. It's like dropping a bomb on a museum, the Louvre."

Another important reason to document endangered languages is to prevent a loss for scientific knowledge. Diversity is of utmost importance for linguistic theory and cognitive sciences, as it provides a ground to challenge and improve existing models of the way human language emerged and how it functions. Many "language universals" have indeed been shaken when confronted to a larger number of languages (Evans and Levinson, 2009), and many areas are under-represented in linguistic research. Diversity is also a condition to extend our knowledge of linguistic structures and forms. In Pirahã for instance, a language spoken in Brazil by about 100 speakers, unknown sounds were found in the phone inventory[6] (Palosaari and Campbell, 2011).

Other authors also relate language diversity to human rights, educational achievement, or the need to protect regional identities in the advent of globalization (Austin and Sallabank, 2011). Some authors even suggest that we should treat linguistic diversity as we treat biodiversity (for instance Hale (1992): "[...] just as the extinction of any animal species diminishes our world, so does the extinction of any language."), although this parallel might overlook important differences between species and languages (see discussion in Crystal, 2000).

1.1.3 Response of the linguistic community

Some of the causes (or consequences) of language endangerment briefly mentioned above cannot be addressed easily. In fact, one may question in certain cases if documentary linguistics' goals are scientific, or defined by activism (Dobrin and Good, 2009).[7] Nevertheless, and to a certain extent, some criteria can inspire action at the academic level. Among these, the amount and quality of documentation certainly can. This is in fact one of the nine criteria elaborated by a group of UNESCO experts to assess the degree of vitality of a language. Other such criteria include "material for language education and literacy", as well as the "community members' attitudes towards their own language" (Brenzinger et al., 2003; Brenzinger, 2008).

[6] *"A voiced bilabial trill (rare in other languages), and a lateral-apical double-flap (unique to Pirahã)"* according to Palosaari and Campbell (2011).

[7] See also (Haspelmath, 2012).

Hence, the "wake-up call" from Krauss (1992) sparked a surge of interest for fieldwork and language documentation during the following two decades (Woodbury, 2011). In Woodbury's words, *"Language documentation is the creation, annotation, preservation and dissemination of transparent records of a language."* Fieldwork, on the other hand and as defined by Sakel and Everett (2012), *"describes the activity of a researcher systematically analysing parts of a language, usually other than one's native language and usually within a community of speakers of that language"*. It usually involves speech data elicitation and collection, transcription, translation, and analysis (phoneme or lexicon inventory, interlinear glosses, grammatical hypotheses, etc.).

1.2 Computational language documentation

Unfortunately, fieldwork is costly time-wise, and there are not enough field linguists to document all of the world's endangered languages. Therefore, a need for automation has arisen, and a new avenue of research has started to emerge in order to develop computational method for language documentation. We will refer to this new field as computational language documentation (CLD).

Note that if, in this thesis, we solely focus on CLD, a related endeavor is *language revitalization*, which aims at reviving languages, or reversing language shifts. This involves research at the crossing of social linguistics and indigenous studies, with an overarching goal to encourage child and adult language learning; various learning methods are compared, and the modernization of the language of interest or its writing system are typically explored (Tsunoda, 2006; Hinton, 2018).

1.2.1 Recent work

One of the earliest attempts to provide unwritten languages with computational tools is found in the work of Besacier et al. (2006), proposing a method for speech translation which was subsequently extended by Stüker et al. (2009). Amongst the pioneering works, Bird (2010, 2011), Hanke and Bird (2013), and Bird et al. (2014) addressed the demand for automatic processing with new methodologies to collect speech data, while Bird and Chiang (2012) proposed an early investigation into machine translation for language preservation. Kempton and Moore (2014) also participated to this effort with a machine-assisted method for the phonemic analysis of unwritten languages.

In recent years, research for low-resource languages has attracted a growing interest,[8] and automatic processing efforts to support language documentation have become more numerous. It is important to note that low-resource languages are not necessarily endangered (nor unwritten), but techniques developed for each condition are likely to benefit both. The Zero Resource Speech Challenge[9] (Versteegh et al., 2015; Dunbar et al., 2017) has been a powerful force to bring together researchers around questions related to language learning with limited resources. The ComputEL workshop[10] is another initiative, launched in 2014 and fostering the use of computational methods to

[8] For instance a query for "low-resource" on the ACL anthology (`https://aclanthology.info`) returns 7 papers with this expression in their title for 2015, 17 for 2016, 22 for 2017, and 53 for 2018.

[9] `https://zerospeech.com/`.

[10] `https://computel-workshop.org/`.

study endangered languages. A workshop for Computational Methods for Endangered Language Documentation and Description (CMLD) has also recently taken place in Paris.[11]

The work of Kamper (2016), addressing zero-resource conditions for speech processing, or the work of Adams (2017) and Anastasopoulos (2019) who specifically target the documentation of endangered languages, are exemplary of the sophisticated methods and models being currently developed in CLD. Most recent work includes speech-to-text translation (Bansal et al., 2018a,b), speech transcription using bilingual supervision (Anastasopoulos and Chiang, 2018b), both speech transcription and translation (Anastasopoulos and Chiang, 2018a), or automatic phonemic transcription of tonal languages (Adams et al., 2018).

1.2.2 The BULB project

The French-German project *Breaking the Unwritten Language Barrier* (BULB) (Adda et al., 2016) corresponds to one of the efforts defining the new field of CLD. Bringing together linguists and computer scientists inside a collaborative framework, its goal is to support the documentation of unwritten languages, using three very low-resource and mostly unwritten languages from the Bantu family as a test bed: Mboshi (Republic of the Congo), Myene (Gabon), and Basaa (Cameroon). The methodology of the project was conceived around three main steps:

1. Collection of a large[12] speech corpus for each Bantu languages, and oral translation into French;
2. Automatic transcription of Bantu speech (phoneme level), and of French speech (word level); automatic alignment between Bantu phonemes and French words;
3. Tool development to support linguists in their documentation and description work.

Speech collection used a mobile device application, Lig-Aikuma[13] (Blachon et al., 2016), extending the original Aikuma developed by Hanke and Bird (2013). The core functionalities (speech recording, oral translation and re-speaking[14]) were adapted to better suit the linguists' needs, and an elicitation mode was added.

The rationale of this methodology, following (Hanke and Bird, 2013), is that oral translations are easier to produce than speech transcriptions, which require time-consuming manual labor.[15] For endangered languages, where bilingualism is common due to language shift processes, it is (in theory, although proving more challenging in practice)

[11] http://lattice.cnrs.fr/cmld/.

[12] The original goal was to collect about 100 hours of speech per language, but at the end of the project, about 50 hours of speech were collected for each language. In comparison, a Griko-Italian corpus (Lekakou et al., 2013; Boito et al., 2018) for endangered language documentation contains less than half an hour of speech (Griko is an endangered dialect from South Italy).

[13] https://lig-aikuma.imag.fr/.

[14] A procedure in which potentially hard-to-hear collected speech is re-recorded slowly so as to facilitate transcription (Woodbury, 2003).

[15] Up to a 35:1 transcription-time to data-time ratio for less-known languages, see (Auer et al., 2010; Dingemanse et al., 2012).

easy to find a bilingual speaker capable of orally translating the recorded speech into a well-resourced language. Automatic transcription of the well-resourced language at the word level can then be performed robustly. Our own work, relevant to the second step of the BULB project's methodology, is concerned with word discovery and alignment between Bantu phonemes and French words (see Section 1.3). The complementary task of automatically transcribing Bantu speech has been examined by Müller et al. (2016), and the particular problem of discovering a phoneme inventory in an unwritten language has been studied in (Müller et al., 2017).

1.2.3 Challenges

Many difficulties for the researcher in CLD are specific to the kind of data available:

- Endangered languages corpora are often not easy to track down. An exception is the extremely rich Pangloss collection,[16] giving access to speech recordings and transcriptions for many endangered languages. Smaller in scale, the EOPAS project[17] follows the same philosophy. The Endangered Languages Archive (ELAR)[18] should also be mentioned.
- In many cases, transcriptions, annotations, or metadata are lacking.[19] If manual transcriptions are costly to produce, as already noted, morphological and syntactic annotations are even more time-consuming, yet a large number of automatic processing techniques require such annotations. This could lead to what Himmelmann (2006, cited by Adams (2017)) has described as "data graveyards", or *"large heaps of data with little or no use to anyone"*;
- Corpora are also typically very small in size (often under an hour of speech, sometimes only a few dozens of sentences). This is a central challenge, as natural language processing (NLP) relies on machine learning techniques, which require large quantities of data for training;
- As most endangered languages are also unwritten, or lack a stable writing system, speech is the primary data available. During language documentation, recordings are often produced in less-than-ideal conditions (e.g. background noise or imprecise articulation from elderly speakers), which adds to the usual challenges of processing speech data (speaker variability, overlapping speech, etc.).

1.3 Scope and contributions

In this section, we delineate the scope of the present work: unsupervised word discovery in the context of CLD. We then summarize our contributions, and present an outline of this thesis.

[16] http://lacito.vjf.cnrs.fr/pangloss/index_en.html.
[17] http://eopas.org/.
[18] https://www.soas.ac.uk/elar/.
[19] Anastasopoulos et al. (2017) for instance reports that half of the collections in the Archive of the Indigenous Languages of Latin America contain no transcriptions, and only a small fraction of the other half are fully transcribed.

1.3.1 Unsupervised word discovery

Documenting the lexicon is an important task in endangered language documentation (see Haig et al., 2011, *inter alia*). Assuming speech can be automatically and reliably transcribed into a sequence of phone-like units (following the methodology of the BULB project, see Section 1.2.2), the *word discovery* task consists to *segment* this sequence into words. In this document, we refer to this task indifferently as word discovery, or *word segmentation*. In our context, and for reasons invoked in Section 1.2.3, we attempt to perform word segmentation without supervision, or minimal supervision.

This task can be carried out in a monolingual setting. But with bilingual data, word discovery can be tightly coupled with word alignment: segmentation can be guided by the alignment between phone-like units on one side of the bilingual corpus, and words on the other side; conversely, automatically segmented words can be aligned to their well-resourced counterpart. A formal description of both tasks is in Section 2.1. For reasons explained in Section 3.1.1, and to make results presented in different chapters of this thesis comparable, we consider graphemic transcriptions made by linguists instead of automatically transcribed phone-like units in our experiments. We investigated the latter, more realistic, condition in (Godard et al., 2018a,d), as well as in (Ondel et al., 2018), and we refer to this work in several places.

Our work embraces the goals of the BULB project, and more largely CLD, to support the work of linguists in the documentation of endangered languages via automatic processing. More specifically, we first aim at benchmarking existing word discovery methods, in order to assess their usefulness for CLD, and their applicability within a language documentation scenario. Our goal is then to propose improvements to the most promising techniques, and make progress towards providing linguists with useful tools in their workflow. In our approach, we examine how, in a low-resource context, word segmentation can be improved with auxiliary information, such as expert knowledge, tonal patterns, or bilingual supervision. At the end of Chapter 2 (Section 2.7), and after introducing key concepts and problems more technically, explicit research questions motivating our work will be presented and discussed.

1.3.2 Outline of the thesis

Chapter 2 formally introduces the problems of unsupervised word segmentation and alignment. In this chapter, we then review early models for segmentation, paradigmatic approaches, and nonparametric Bayesian language models. A brief review of the automatic word alignment literature allows us to further examine the word discovery literature with joint models for segmentation and alignment. We conclude with the lessons learnt, and with open questions motivating our own work.

Chapter 3 describes two novel realistic corpora for endangered language documentation used throughout the whole thesis. We experiment with several word segmentation systems, either with a monolingual or a bilingual setting, and contrast our results with the results of experiments conducted on an additional French-English corpus. We vary data sizes, representations, and establish strong baselines for unsupervised word discovery, while showing the difficulty to take advantage of bilingual data in our context.

Chapter 4 strongly improves on previously established baseline results for word segmentation, by incorporating expert knowledge from linguists via the Adaptor Grammar framework. We devise a methodology allowing us to experiment with a large hierarchical grammar landscape, and we show how to use this grammatical landscape to explore and support linguistic hypotheses, extending the linguist's toolkit.

Chapter 5 questions whether tonal information can be put to good use in order to improve unsupervised word discovery. We show that tones can help in a supervised setting, and we propose a new Bayesian model designed to capture tonal information in the unsupervised setting, with unconvincing results.

Chapter 6 proposes and refines a neural segmentation method making use of translations. We describe the RNN encoder-decoder architecture with attention that our method leverages, and introduce two extensions to this method, one of which achieves significantly higher precision in the segmentations than the baseline.

Chapter 7 concludes the thesis and lays the ground for future work, in relationship to the current evolution of CLD.

1.3.3 Contributions

The main contributions of this thesis are the following:

- We propose a comprehensive survey of the literature and tools relevant to the unsupervised word discovery task, and its connection with the automatic word alignment task;
- We introduce two new corpora (one of which is now publicly released), and conduct systematic word segmentation experiments with various methods on very low-resource oral languages;
- We show that word segmentation can be vastly improved using expert knowledge and through a close collaboration with linguists; we also propose a methodology to integrate the Adaptor Grammar framework in the linguist's toolkit to explore new languages;
- We study the usefulness of tonal information for the word discovery problem, and introduce a new Bayesian model aiming at taking advantage of such information;
- We propose and improve a neural segmentation method using attention matrices; while unable to beat the best monolingual methods on the word discovery task, our method outperforms the bilingual method previously at our disposal, and is promising in a realistic (noisy) setup, using automatically transcribed phone-like units.

1.3.4 Author's publications

Main publications

- Pierre Godard, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Laurent Besacier, Hélène Bonneau-Maynard, Guy-Noël Kouarata, Kevin Löser, Annie Rialland, and François Yvon. Preliminary Experiments on Unsupervised Word Discovery in Mboshi. In *Proceedings of Interspeech*, San-Francisco, California, USA, 2016
- Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélène Maynard, Markus Müller, Annie Rialland, Sebastian Stüker, François Yvon, and Marcely Zanon Boito. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. In *Proceedings of LREC*, Miyazaki, Japan, 2018a
- Pierre Godard, Laurent Besacier, François Yvon, Martine Adda-Decker, Gilles Adda, Hélène Maynard, and Annie Rialland. Adaptor Grammars for the Linguist: Word Segmentation Experiments for Very Low-Resource Languages. In *Proceedings of the 15th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON)*, Brussels, Belgium, 2018b
- Pierre Godard, Kevin Loser, Alexandre Allauzen, Laurent Besacier, and Francois Yvon. Unsupervised Learning of Word Segmentation: Does Tone Matter? In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, Hanoi, Vietnam, 2018c
- Pierre Godard, Marcely Zanon Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, and Laurent Besacier. Unsupervised Word Segmentation from Speech with Attention. In *Proceedings of Interspeech*, Hyderabad, India, 2018d

Collaborations

- Lucas Ondel, Pierre Godard, Laurent Besacier, Elin Larsen, Mark Hasegawa-Johnson, Odette Scharenborg, Emmanuel Dupoux, Lukas Burget, François Yvon, and Sanjeev Khudanpur. Bayesian Models for Unit Discovery on a Very Low Resource Language. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018
- Odette Scharenborg, Laurent Besacier, Alan W. Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux. Linguistic Unit Discovery from Multi-Modal Inputs in Unwritten Languages: Summary of the "Speaking Rosetta" JSALT 2017 Workshop. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018a
- Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitri Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. Breaking the Unwritten Language Barrier: The Bulb Project. In *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia, 2016
- Annie Rialland, Martine Adda-Decker, Guy-Noël Kouarata, Gilles Adda, Laurent Besacier, Lori Lamel, Élodie Gauthier, Pierre Godard, and Jamison Cooper-Leavitt. Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville). In *Proceedings of LREC*, Miyazaki, Japan, 2018

- Sebastian Stüker, Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitri Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. Innovative Technologies for Under-Resourced Language Documentation: The Bulb Project. In *Proceedings of CCURL (Collaboration and Computing for Under-Resourced Languages : Toward an Alliance for Digital Language Diversity)*, Portorõz, Slovenia, 2016
- Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang. XNMT: The eXtensible Neural Machine Translation Toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts, USA, 2018

Chapter 2

Background

Contents

In Chapter 1, we introduced computational language documentation (CLD) and gave a broad picture of the challenges facing field linguists and computer scientists with respect to the preservation and documentation of endangered languages. The present chapter narrows down the scope of CLD, and introduces the particular problem examined in this thesis: the unsupervised segmentation of a stream of symbols into words, as well as its connection with the automatic word alignment task. In the BULB project's methodology, this corresponds to a subpart of the automatic processing step after speech data collection and translation (Section 1.2.2). We review the literature and tools necessary to understand our work, and introduce notations and metrics. Parts of this chapter have appeared in (Godard and Yvon, 2016), and Section 2.5 borrows from (Godard, 2014).

2.1 Word segmentation and alignment

As discussed in Chapter 1, collecting annotated data is costly, and non practical to meet the challenges of documenting a large number of endangered languages. Consequently, the work presented in this thesis is concerned with unsupervised, or minimally supervised, automatic processing of the "raw" bilingual data at our disposal after collection (see Section 1.2.2).

Such data, in the BULB project's methodology, consist in pairs of mutually translated sentences in the *unwritten language*[1] (henceforth UL) and in the *well-resourced language*[2] (WL). A sentence $\boldsymbol{\pi}$ in the UL is a sequence of L units, $\boldsymbol{\pi} = \pi_1, \ldots, \pi_l, \ldots, \pi_L$, and a sentence $\mathbf{w}$ in the WL is a sequence of I units, $\mathbf{w} = w_1, \ldots, w_i, \ldots, w_I$. In practice, units π_l in the UL can correspond to transcribed characters, phones, pseudo-phones,[3] phonemes, or even speech frames. Units w_i in the WL, on the other hand, correspond to transcribed words.

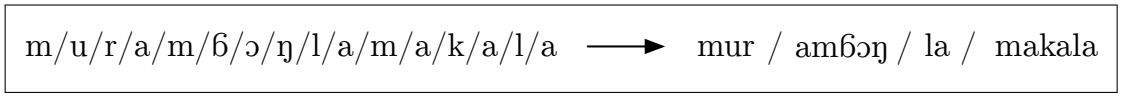
2.1.1 Two sides of the same problem

One key step in documenting an UL is to identify (parts of) the lexicon, a central problem addressed in this work. However, to be fully usable by linguists, language learners, ethnologists, etc., discovered lexical units in the UL need to be associated with their counterpart in the WL, and therefore with of proxy of their meaning. We are thus facing two problems:

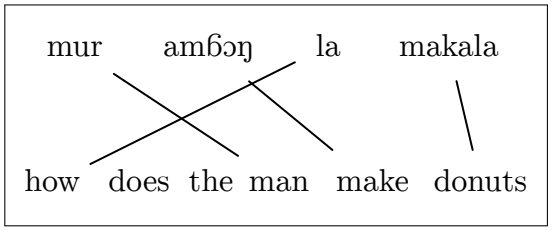
- A *segmentation* problem, as we need to transform a continuous sequence of units $\boldsymbol{\pi}$ in the UL into words or subword units (see Figure 2.1a).
- An *alignment* problem, as we need to map unknown discovered units in the UL with known units in the WL (see Figure 2.1b).

It is natural to think of the segmentation problem for the UL side as a preprocessing task before one can perform an alignment to the word units in the WL. This approach, depicted Figure 2.2a, is indeed taken by many researchers in order to align comparable

[1] Mboshi, Myene, and Basaa in the BULB project.
[2] French, in the BULB project.
[3] When the units are the results of (unsupervised) acoustic units discovery.



(a) The segmentation problem: transforming a continuous sequence of units in the UL into words or subword units.



(b) The alignment problem: mapping units in the UL with known units in the WL.

Figure 2.1: A first view of the word segmentation and alignment problems.

units (Section 2.6.1). Conversely, alignment between units in the UL and words in the WL can help inferring a segmentation on the UL side, as depicted in Figure 2.2b, although, alone, this approach is less practical for reasons explained in Section 2.6. Lastly, segmentation and alignment can be jointly modeled, in the hope that refined information regarding segmentation during training will inform the alignment decisions the model makes, while refined alignment will guide towards better segmentation of the UL (Figure 2.2c and Section 2.6.2).

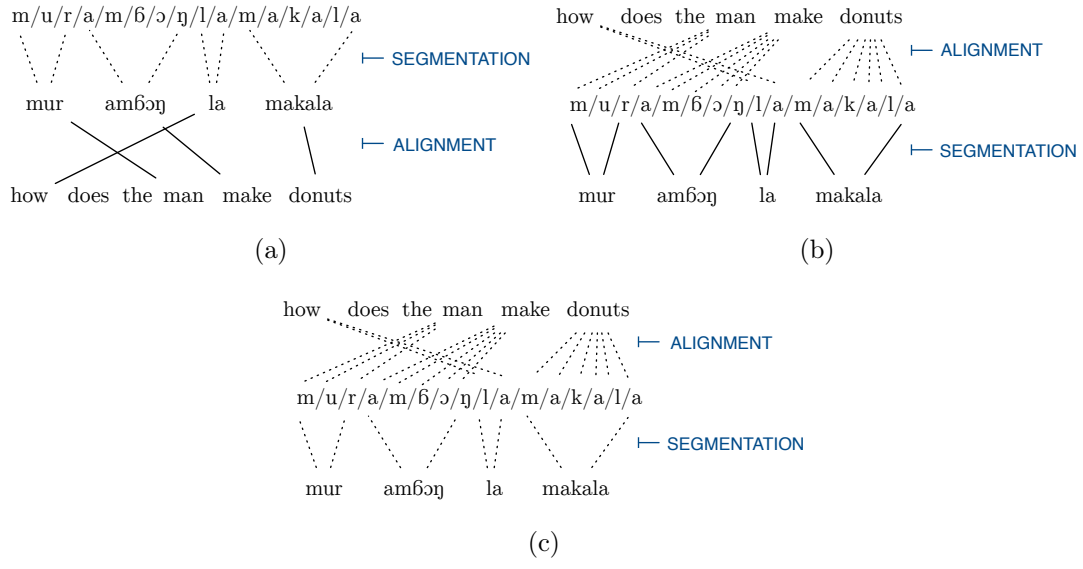


Figure 2.2: The segmentation and alignment tasks in relationship to each other; example from Bàsàá by Fatima Hamlaoui. Segmentation can serve as a preprocessing step for alignment (top left), while alignment can guide segmentation (top right). Both segmentation and alignment can also be learnt jointly (bottom).

In the remainder of this thesis, we will refer to the unsupervised word segmentation task indifferently as *word segmentation* or *word discovery*, and to the automatic word alignment task as *word alignment*. Even though our work focuses mainly on the word segmentation task and its evaluation, the entangled nature of these two problems leads us to also briefly review the literature related to automatic word alignment. We now introduce both tasks more formally.

2.1.1.1 Word segmentation

The word segmentation (or discovery) task is, *per se*, a monolingual task consisting in identifying boundaries around word units from an unsegmented stream of symbols in a given language. Formally, it consists in defining a function associating the sequence $\boldsymbol{\pi} = \pi_1, \ldots, \pi_l, \ldots, \pi_L$ to a sequence $\boldsymbol{\omega} = \omega_1, \ldots, \omega_j, \ldots, \omega_J$, where ω_j, for $j \in [1, J]$, is a *word*. It is well known that a formal definition of the word *word* is hard to produce in many languages. This discussion would reach far outside of the scope of this work, so we will somewhat dodge the problem by using *word* either to relate to what a linguist would call (or has called in its annotations) a word, or to refer to a *token*, the output of a tokenizer for a particular language.

An equivalent definition of the word segmentation task is to associate the sequence $\pi_1, \ldots, \pi_l, \ldots, \pi_L$ to a sequence of binary decisions $b_1, \ldots, b_l, \ldots, b_{L-1}$, with each b_l corresponding to the presence (1), or absence (0), of a word *boundary* after unit π_l in the original sequence. Note that sentence boundaries are known in our scenario, and that they are also word boundaries for the first and last words in the sentence.

In that respect, the word segmentation task presents strong links with unsupervised morphology learning. This is because, from an abstract point of view, morphology learning and lexical acquisition problems can be viewed as instances of a same generic task, which is to learn to segment an input stream of symbols in an unsupervised way, and to extract a minimal inventory of units, be they called words or morphemes. Some of the background work we review in this chapter will therefore be concerned with learning morphology (learning a list of morphemes in particular) rather than a lexicon. Moreover, the question of the segmentation *granularity*, i.e. whether we, in effect, segment at the word or subword level (or multi-word level for that matter), will recur in this work.

Another line of research addresses the task of segmenting sentences into words in languages having no overt word separator in their orthography (Chinese, Japanese, Thai, etc.) without supervision. As it is formally identical to the task we have just defined, this will also be relevant to our study, although the purpose remains often distant to the language documentation goal we pursue. In particular, many studies in this line of research approach the word segmentation task with machine translation in mind: in this context, the right segmentation granularities for the source and target sides of the corpus are determined by the translation performance achieved for *each particular language pair*. Rather than the linguistic soundness of the decomposition of, say, the source language, researchers aim at finding its right decomposition when translated into a particular target language.

The word segmentation problem is tightly coupled with the design of *language models*, i.e. probabilistic models assigning a probability distribution $P(w_1, \ldots, w_I)$ to a

sequence of words $w_1, \ldots, w_I$. Without loss of generality, this probability distribution can be rewritten, using the chain rule, as

$$P(w_1, \ldots, w_I) = P(w_1) \prod_{i=2}^{I} P(w_i \,|\, w_1, \ldots, w_{i-1}) \,. \tag{2.1}$$

Except for some of the early approaches to word segmentation (Section 2.2), and most paradigmatic approaches (Section 2.3), language models are the theoretical backbone for word segmentation. Computational modeling of child language acquisition, for instance, has been heavily relying on such models. We review various studies related to the word segmentation task in Sections 2.2, 2.3, and 2.4.

2.1.1.2 Word alignment

The automatic word alignment task, contrary to the segmentation task we just defined, is a bilingual task in essence. Informally, it consists, given a parallel corpus aligned at the sentence level, to identify links between words (or more generally "units") that are mutual translations of each others.

More formally, this can be viewed as learning a symmetrical *binary relation* $\mathcal{R}$ over the sets V_S and V_T indexing word positions in the *source* and *target* parts of each sentence pair.[4] For reasons that will become clearer in Sections 2.5 and 2.6,[5] we assume here that our source sentence is the WL sequence $\mathbf{w}$, and that the target sentence is the UL sequence $\boldsymbol{\omega}$, but this could be reversed. Learning this binary relation corresponds to learning, for each sentence pair, a subset of the Cartesian product $V_S \times V_T$. A word alignment can, therefore, equivalently be represented by a *simple bipartite graph*,[6] making links more explicit. Both mathematical objects can be represented as matrices, in which binary values indicate the presence or absence of a link; the search space $\mathcal{A}$, hence, will correspond to all binary matrices $A = (a_{ij})$, with $a_{ij} = 1$ if source word w_i is aligned to target word ω_j, and $a_{ij} = 0$ otherwise.

For computational reasons, however, the search space $\mathcal{A}$ can be restricted to binary vectors $\mathbf{a} = a_1, \ldots, a_J$, in which each $a_j \in [1, I]$ indicates the word position in the source sentence $\mathbf{w}$ to which target word ω_j is aligned to. This drastically reduces the size of the search space, from $2^{I \times J}$ to I^J, and likens the word alignment task to a sequence labeling task, in which word ω_j, aligned to word w_{a_j}, is labelled a_j. Figure 2.3 depicts the various representations we just discussed. Other representations for alignments can also be found in the literature, especially when trying to align units of different granularities, introducing for instance the concept of *spans*.

As we note in Section 2.5, where we provide the theoretical foundations for statistical word alignment, the concept of word alignments emerged in the first word-based

[4] The *source* and *target* terminology is standard in the machine translation (MT) literature. It identifies the direction of the translation, from source to target language. However, due to the use of the *noisy channel* model, this terminology can sometimes become confusing.

[5] Succinctly, if $\boldsymbol{\omega}$ is replaced by its unsegmented counterpart $\boldsymbol{\pi}$, and since standard alignment models allow only for one outbound link per target units, this direction will better accommodate the alignment between, say, words and phonemes.

[6] An undirected graph, without multiple edges or loops, in which every edge connects a vertex in V_S to one in V_T.

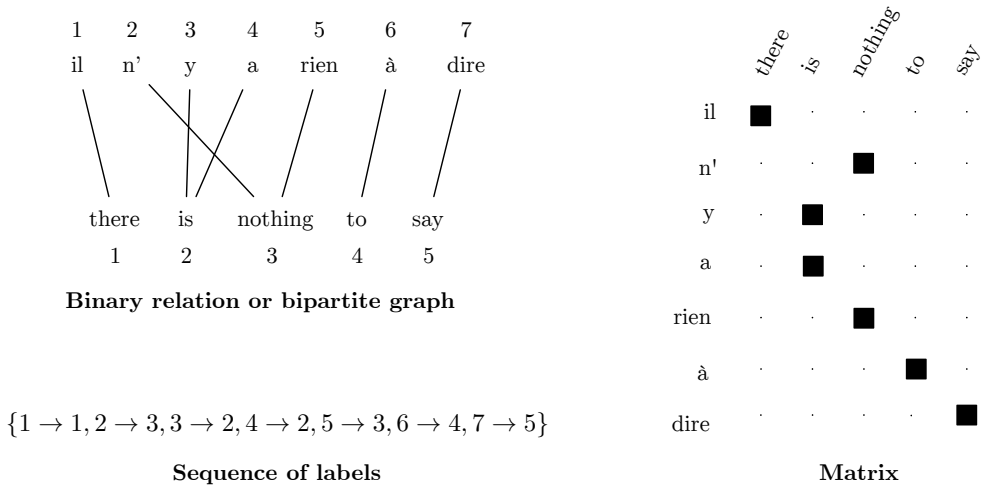


Figure 2.3: Various representations for word-to-word alignment. (English is the source language, and French, the target language.)

models for machine translation. Alignments have subsequently been the foundation of statistical machine translation (SMT) (Lopez, 2008; Koehn, 2010), and specifically phrase-based SMT, as they allow for the extraction of relevant phrase pairs used to build translation tables. In recent years, and as neural machine translation (NMT) has gradually superseded SMT in machine translation, there has been significantly less work on word alignment; NMT indeed, in its current form, does not rely crucially on such a concept.

2.1.2 Evaluation

We introduce now the main metrics generally used to evaluate the tasks defined in the preceding section. More specific quantitative indicators will be defined when they are needed in ensuing chapters.

Segmentation To evaluate word segmentation accuracy, we will resort to orthographic transcriptions manually segmented by linguists. These transcriptions, for languages known to be unwritten or rarely written, correspond to non-standard transcriptions, and reflect idiosyncratic choices made by a particular linguist. It is important to keep in mind that the choice, for instance, to attach or detach a particular prefix from the word stem, promoting it in the latter case to its own word form, can be at times controversial among linguists studying the language.

The main metrics we use in this work, following Goldwater (2006), are:

- BP, BR, and BF: precision, recall, and F-measure on word boundaries, excluding the sentence boundaries which are already known.
- WP, WR, and WF: precision, recall, and F-measure on words. Both boundaries delimiting the word need to be correctly identified to constitute a correct decision.

- LP, LR, and LF: precision, recall, and F-measure on the lexicon (word types as opposed to word tokens in the previous measure).

In each case, precision corresponds to the proportion of correct decisions (e.g. introduce a word boundary) amongst all decisions made by the system. Recall corresponds to the proportion of correct decisions amongst all correct decisions in the reference, and F-measure is the harmonic mean of the precision and the recall (F-measure $= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$).

As noted in Section 2.1.1.1, and related to the arbitrary choices made in the gold transcriptions that we just mentioned, one important problem with these metrics is their inability to mix word and subword levels in the evaluation. If some automatic process over-segments the data, such metrics will not be able to discriminate linguistically sound over-segmentations from mere incorrect ones. In other words, adding boundaries between morphemes, rather than randomly inside word forms, will not be captured by the metrics introduced here. We have not been able to devise a metric mixing both levels, one main difficulty being that it would require a segmentation reference at the morpheme level. Consequently, however imperfect, the metrics introduced above are the most solid quantitative metrics at our disposal. They will be complemented, when it is useful in the discussion, by other statistics, for instance the number of predicted tokens or types, or the average length of predicted tokens.

Alignment To evaluate word alignment accuracy, a common measure introduced by Och and Ney (2003) is the AER (*Alignment Error Rate*). Links are labeled by a human annotator as either "sure" (these links belong to a set noted S) or "possible" (these links belong to a set noted P, and correspond to links that are ambiguous for the annotator; $S \subseteq P$). On the other end, the alignment A is the set containing all the links automatically discovered by a given system. The AER is then defined by:

$$\text{AER}(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \,. \tag{2.2}$$

This metric has been criticized by Fraser and Marcu (2007) for not satisfying an important property of standard F-measure, i.e. to penalize important disequilibria between precision ($\frac{|A \cap P|}{|A|}$), and recall ($\frac{|A \cap S|}{|S|}$). The main difficulty for us, however, is that the metric requires a gold standard parallel corpus aligned at the word level. Such resource is notoriously difficult and costly to build (Melamed, 1998), and unfortunately not available for our data. Some work is being currently undertaken, however, to produce such references for the BULB project (Section 1.2.2).

2.1.3 Remarks

We conclude the first section of this chapter with several remarks, meant to provide some orientation to the reader in the vast body of research addressing questions related to word segmentation and alignment, and in the present work. We first put unsupervised morphology learning in perspective with our own research. We then point out to the diversity of language typologies, a fact sometimes understated in unsupervised approaches, when these are assumed to be linguistically agnostic. We conclude

with a discussion about the purpose of the evaluation of the tasks we defined in Sections 2.1.1.1 and 2.1.1.2.

Learning morphology The research on word segmentation presents strong links with unsupervised morphology learning, and often tackles conceptually equivalent goals, as we noted in Section 2.1.1.1. Therefore, it should come as no surprise that some of the work we discuss in the remainder of this chapter is concerned with segmenting words into subword units, rather than segmenting sentences into word forms.

One approach is to consider morphology learning as a segmentation task using distributional cues, e.g. context. This approach is particularly suited for analytic or agglutinative languages (see next paragraph for precisions about typology), which tend to exhibit a one-to-one correspondence between meaning and form. Hammarström and Borin (2011) provide a high-level comprehensive survey of the unsupervised learning of morphology, and distinguish between two trends for that approach:

1. to provide an operational *description of morphological phenomena*, usually restricted to simplistic forms of concatenative processes. The input here is typically a set of raw word forms that need to be explained in the most economic way possible – for instance using the minimum description length principle (MDL) (Rissanen, 1989);
2. to model *language acquisition*. This trend attempts to provide plausible models to account for the way infants learn language using, in particular, child-directed speech corpora. Researchers also often try to consider supplementary information to raw phoneme strings, e.g., intonation and other prosodic cues, but also semantics, pragmatics, etc.

A different approach is to model the relations between word forms, and build *paradigmatic* structures from those relations; this last trend will likely better accommodate the morphological behavior of flective languages (see next paragraph also). Some work, for instance the work of Goldsmith (2001), lies between these two poles, involving a concept of *signature* akin to a paradigm, yet relying mostly on a segmenting approach.

Language typologies As we just pointed out, the diversity of approaches to the unsupervised learning of morphology (segmentation vs. paradigms) is largely induced by the broad variety of natural languages. Since phenomena occurring at the subword level also influence word segmentation,[7] typological considerations will also inform the word segmentation task. Borrowing from (Eifring and Theil, 2004), we distinguish between *analytic*, *synthetic*, and *polysynthetic* languages.[8] The latter two classes of languages may also be further divided into so-called *agglutinative* and *flective* (or *fusional*) languages.

[7] For instance, vowel deletion in Mboshi (see Section 3.2.1).

[8] Analytic, or isolating, languages tend to feature words corresponding to only one morpheme, whereas in synthetic (resp. polysynthetic) languages, words correspond to more than one morpheme (resp. several morphemes constituting sometimes up to an entire clause).

Agglutinative languages have the following properties: i) they make use of morphemes that express only one meaning element; ii) their morphemes have clear boundaries; iii) grammatical processes adjoining prefixes and suffixes do not modify morphemes' forms. Flective languages, conversely, display the opposite properties, called *cumulation*, *fusion*, and *introflection*. It is important to note that this typology describes the extremes of a continuum, and that natural languages can display both flective and agglutinative phenomena, or be mostly analytic despite having some words with multiple morphemes in their lexicon.

One may hope that an unsupervised learning approach will be linguistically agnostic, but many experiments, including ours (see for instance the contrast between segmentation results in Mboshi and English in Section 3.3.2), prove otherwise. In that respect, the recent work of Vania and Lopez (2017) is a rare occurrence of a systematic study across language typologies (albeit not in the context of unsupervised word discovery, but on a language modeling task comparing word representations). As we mostly study word segmentation for two Bantu languages (Mboshi and Myene), which display a mostly agglutinative morphology, we will naturally be drawn, in our experiments, towards the segmentation approach rather than the paradigmatic approach.

Intrinsic or extrinsic evaluation? The tasks defined in Section 2.1 can be evaluated with the metrics introduced in Section 2.1.2 for their own sake. Word segmentation aims at building a lexicon for an unknown language, and word alignment aims at providing meaning to the discovered units. Combined together, and provided they are performed with enough accuracy, these two tasks can help building a bilingual dictionary automatically. This is valuable for linguists and for the overarching goal of language documentation and preservation.

We will also consider however, in Chapter 4, the word segmentation evaluation as an extrinsic measure akin to evaluate the correctness of a linguistic hypothesis. This could also be done, in theory, with the word alignment metric, AER, to assess the "compatibility" of the segmentation on the source and target sides. It is reasonable to posit that the best segmentations in the UL and in the WL for the alignment task are those that will allow for a one-to-one correspondence. As noted earlier, this does not guarantee a linguistically sound segmentation in general, but if the two sides are known to be typologically close, one could imagine to use AER as an extrinsic measure for segmentation on the language where a reference segmentation is lacking. In practice, however, creating manual word alignments is time-consuming. Moreover, in order to compute AER we would need for automatically segmented units to be comparable to reference (annotated) units in the UL; this would not be the case at the word level (although we could extract phoneme-to-word alignments from word-to-word alignments).

2.2 Early models for unsupervised string segmentation

We start our review of previous work related to word segmentation with three early approaches: the first using transition statistics, the second introducing a particular use of HMM models, and the third relying on ideas related to data compression. Many vari-

ants of these approaches have been subsequently proposed for unsupervised morpheme analysis in the context of the Morpho Challenge (Kurimo et al., 2010).

2.2.1 Pioneer work

Harris (1955) pioneered automatic morphology discovery, observing that transitions between morphemes inside a word are less predictable than transitions between phonemes within a morpheme. Counting the number of phonemes that could extend any prefix into another legal prefix in the language – the *successor frequency* – it is possible to introduce without any supervision a boundary, within a word, at positions that correspond to peaks of that frequency.[9] This approach, and its information-theoretic interpretations using mutual information and entropy measures, proved to be extremely influential. Déjean (1998), in particular, devised an unsupervised morpheme discovery procedure using Harris' local association statistics during a bootstrapping step, subsequently expanding the morpheme list with morphemes appearing in similar contexts to the ones already discovered.

This strategy can be applied to the word segmentation task, observing that transitions between phonemes at word boundaries are also less predictable than within words. A variant (Saffran et al., 1996) uses "transitional probabilities" between syllables, i.e. the conditional probability of a syllable given the previous syllable, to identify word boundaries. This leads, however, to poor results on realistic corpora, as demonstrated by Brent (1999). Another application of this principle, this time as a preprocessing step, can be found in (Besacier et al., 2006), in an early work pursuing translation from speech (see Section 2.6.1).

2.2.2 Multigrams

The transitional methods we just mentioned focus locally on particular statistics; they do not represent explicitly the words, nor the structure of their sequence, to place boundaries. In contrast, Deligne et al. (2001) expose different concepts introduced through a series of papers addressing the problem of segmenting one or multiple streams of symbols in a non-supervised manner. Here, the focus is to learn variable-length dependencies inside strings of phonemes or graphemes.

Deligne and Bimbot (1995) present the "multigram model" in which a sentence (or more generally a stream of symbols) is seen as the concatenation of independent sequences of words (resp. symbols). Those sequences' length can vary up to a maximum length n. One way to look at the n-multigram model is to think of it as a n-state Hidden Markov Model with state i emitting only sequences of length i and all transition probabilities being equal (see Figure 2.4). This allows for an efficient training of the model using the EM algorithm and a forward-backward procedure so as to avoid enumerating all the possible segmentations.

This forward-backward procedure slightly differs from the standard approach used for HMM training, in order to take into account the dependency of the number of emissions with respect to the segmentation. The multigram model is compared successfully,

[9] The equivalent reasoning for suffixes makes use of the *predecessor frequency*.

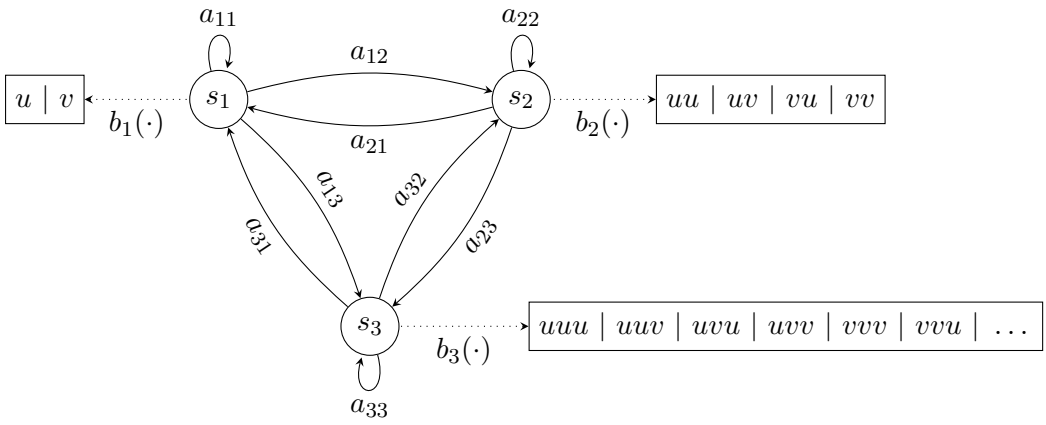


Figure 2.4: A HMM corresponding to a 3-multigram model with 3 states s_1, s_2, s_3 which can emit respectively 1, 2 or 3 symbols from the vocabulary $\{u, v\}$. The transition probabilities a_{ij} from state s_i to state s_j are uniform; $b_i(x)$ represents the probability to emit symbol x in state s_i.

as a language model, to n-gram models of different order. It is also subsequently implemented in the context of the unsupervised segmentation of phoneme strings (Bimbot et al., 1995).

The model is then extended in (Deligne et al., 1995) to a *joint* version that is able to deal with two or more streams of symbols, which are themselves seen as different transcodings of the same higher level symbol stream. To make the model tractable, it is again necessary to limit the maximum length of units in both streams. A m, n joint multigram model can then be identified to a HMM where each of the $m \times n$ states emits pairs of i symbols in one stream, and j symbols in the other. This new joint multigram model is able to learn joint segmentations through many-to-many sequential pairings. In this respect, this last extension of multigrams belongs to the family of joint models studied in Section 2.6. The authors remark that when using this model to learn graphemic and phonetic pairings from words, the extracted joint units often have a morphological interpretation.

2.2.3 Minimum description length principle

The minimum description length principle (MDL) was introduced by Rissanen (1989). It rests on the assumption that the model M^* (among all models in a given set $\mathcal{M}$) that best explains the regularities in some data D will also be the model that allows for the highest compression of the data.[10] More formally,

$$M^* = \underset{M \in \mathcal{M}}{\operatorname{argmin}} \left(L(D \mid M) + L(M) \right), \tag{2.3}$$

[10] In Rissanen's words: *"In our thinking, the main objective in statistics is to learn the constraints in the observed data which permit the shortest encoding both of the observations and the constraints."* (Rissanen, 1989)

where $L(M)$ is the code length (or "description length") needed to specify model M, and $L(D \mid M)$ is the code length of data D using the code provided by M.

De Marcken (1996) proposes that, if a lexicon is given, it is possible to learn a locally optimal parsing of the input data via the EM algorithm. In the case of a parsing that involves only concatenation and according to the work of Deligne and Bimbot (1995) (Section 2.2.2), this corresponds to an implementation of the Baum-Welch algorithm, that is, EM with a forward-backward procedure. From there, it is possible to infer probabilities for the lexicon, hence its code length.[11] In MDL terms, the main idea consists in positing that, if a lexicon (the model) minimizes both its own description length (i.e. the space needed to encode it) and the description length of the data, then that lexicon is the theory that best explains the data, and should be able to capture some of the principles at work in the language that originated the data.

The difficulty, as with most approaches encountered in the MDL framework, is that the search space $\mathcal{M}$ (here the set of possible lexicons) can be very large, if not infinite. This makes search heuristics necessary, which will be likely to harm the principled nature of the model, and its interpretability. A strong case, in that respect, is made by the work of Goldwater (2006), discussed in Section 2.4.3. Interestingly, Goldwater (2006) also notes that Equation (2.3) can then be seen as a maximum *a posteriori* (MAP) Bayesian inference scheme, because

$$\begin{aligned} M^* &= \operatorname*{argmin}_{M \in \mathcal{M}} \left(-\log P(D \mid M) + L(M)\right) \\ &= \operatorname*{argmax}_{M \in \mathcal{M}} \left(P(D \mid M) \times 2^{-L(M)}\right). \end{aligned} \tag{2.4}$$

The first line ensues from the fact that, under M, the optimal code length in bits for D is approximately $-\log P(D \mid M)$.[12] Therefore the MDL approach is equivalent to assuming a prior probability for the hypothesized model M, $L(M)$, that decreases exponentially with its code length, while $L(D \mid M)$ corresponds to the likelihood of the data given this model.

Described in an influential series of papers spanning over a decade, Morfessor[13] (and all its variants) has been established as a *de facto* standard for unsupervised morphology learning, especially for modeling agglutinative morphology. Its reliance on the MDL principle is described in (Creutz and Lagus, 2002), where the corpus code length is decomposed, similarly to De Marcken's approach with words mentioned above, into the code length of a morph[14] dictionary, computed as the sum of the morph lengths, and the code length of the corpus with each morph m coded with $-\log P(m)$ bits. This model is refined in (Creutz, 2003), where the morph generation model is replaced by a unigram of characters, and where a more complex prior on the morph codebook, integrating both length and type distributions, is used. Creutz and Lagus (2004) introduce some morphotactics, with distinct hidden states for prefix, stem and suffixes,

[11] The correspondence between code length functions and probability distributions is central to MDL.

[12] More precisely, $\lceil -\log P(D \mid M) \rceil$, as the logarithm may not be equal to an integer. The building of such optimal codes is detailed in, e.g., (Cover and Thomas, 2006).

[13] See `http://www.cis.hut.fi/projects/morpho` for open source implementations of Morfessor.

[14] This term is commonly used to refer to "pseudo-morphemes", i.e. automatically discovered morpheme-like units.

while Creutz and Lagus (2005) (see also (Creutz and Lagus, 2007) for a more comprehensive presentation), allow for segmentation to be performed recursively.[15] Kohonen et al. (2010) first attempt to introduce annotated (i.e. segmented) data in conjunction with non-annotated data, using the model of Creutz and Lagus (2005); as in most semi-supervised approaches, the objective function combines two likelihood terms that need to be carefully weighted. Grönroos et al. (2014) describe the latest evolution of Morfessor, trying to better combine the benefits of semi-supervision and richer morphotactics.

2.3 Learning paradigms

In Section 2.1.3, we mentioned that morphology learning can be seen as building *paradigms*. This is an important approach that we account for briefly here, as this will not be an approach we pursue.

2.3.1 Signatures

The work of Goldsmith (2001), as stated in Section 2.1.3, lies somewhere between the learning of morphological segmentation procedures, as well as the identification of morpheme inventories, and a more paradigmatic approach. Signatures, the key concept in this work, can be viewed as a weaker form of linguistic paradigms, consisting of sets of suffixes that systematically alternate with a set of stems (see Figure 2.5). The approach is based again on the MDL principle (Section 2.2.3), which is instantiated here as follows: the model is made of sets of stems, suffixes, and *signatures*, which record the possibility that a stem and a suffix can actually co-occur. Denoting t a stem, f a suffix, w a word, and σ a signature, the probabilistic model which underlies the compression algorithm can be expressed as:

$$P(w = tf) = \sum_{\sigma} P(\sigma)P(t\,|\,\sigma)P(f\,|\,\sigma)\,. \tag{2.5}$$

Equation (2.5) serves to compute the size of the data, given the model; the model size takes into account the length of the encoding of the lists of stems, suffixes, and signatures.

$$\left\{\begin{matrix}\text{jump}\\ \text{laugh}\\ \text{walk}\end{matrix}\right\}\left\{\begin{matrix}\text{NULL}\\ \text{ed}\\ \text{ing}\\ \text{s}\end{matrix}\right\}$$

Figure 2.5: An example of signature which covers the words *jump, jumped, jumping, jumps, laugh, laughed, laughing, laughs, walk, walked, walking, walks.*

If the underlying principles behind this model (implemented in the Linguistica system) are well motivated, and based on notions of text compression, the algorithmic part

[15] Which means that a morph can itself be decomposed (e.g. in a word such as *creations* where the suffix *-ations* can be further decomposed into *-ation+s*).

is more *ad hoc*, a problem often seen in MDL approaches, as mentioned in Section 2.2.3. An interesting question also raised by this work is the completeness of signatures: for long signatures, e.g. a complete verbal paradigm, many sub-signatures also exist in the data (corresponding to partial paradigms); how can they be merged since there is no way to "hallucinate" forms that are not in the original list, as this would not help to compress the data?

2.3.2 Signatures as finite state automata

Hu et al. (2005) further expand this trend of research, interpreting Goldsmith's signature as a finite-state automaton (FSA) (see Figure 2.6) built from character-based alignments between pairs of word forms. These alignments are established using the string edit distance (SED), identifying perfect and imperfect character's spans, corresponding respectively in the FSA to adjacent states with either one transition or two transitions. The FSAs extracted from pairs of words in the corpora (here a Swahili translation of the Bible) are then collapsed, disambiguated, and scored in a manner reminiscent of the MDL.[16] The most robust FSAs are finally used to hypothesize stems heuristically from words not yet analysed in the corpus

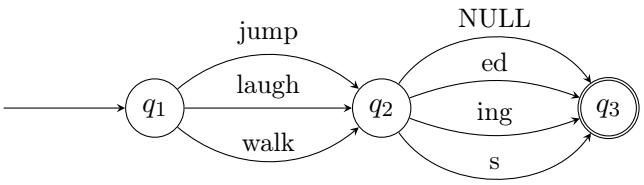


Figure 2.6: The signature presented in Figure 2.5 now seen as a 3-state finite-state automaton

2.3.3 Paradigms

The work of Dreyer and Eisner (2011) is the culmination of a series of papers (Dreyer et al., 2008; Dreyer and Eisner, 2009) aimed at learning *word-based models of morphology* (Aronoff, 1976; Blevins, 2006). Under this view of morphology, morphological processes cannot be reduced to the concatenation of segmental strings to a stem; instead, morphology should attempt to model the relations between forms within *paradigms* – a notion that should therefore be given a first class status in this theory.

From a computational perspective, segmentation is no longer the main target. Instead, the model should be able to: i) cluster related forms within paradigms; ii) learn the mapping between forms and slots in the paradigm; iii) predict the realization of slots that are not observed in the corpus, all this in an almost unsupervised fashion. In the work of Dreyer and Eisner (2011), the supervision consists mostly of an abstract description of the paradigm's cells and of a handful of exemplar paradigms. In a nutshell, this work relies on two main components. The first component is a finite-state probabilistic

[16] This score is based on the number of letters "saved" by the FSA template when generating the corresponding words.

model for morphologically related forms, which should capture the surface similarity and systematic alternations between forms within a paradigm. The second component is a nonparametric Bayesian model, which takes care of the statistical regularities of the distribution of types, inflections, and forms.

2.4 Nonparametric Bayesian models

We just mentioned the use of nonparametric Bayesian models to learn morphological paradigms. In another context, the modeling of language acquisition, especially for infants, the seminal work of Goldwater (2006) initiated a rich line of research using nonparametric Bayesian language models. This work demonstrates that these models lead to better word segmentation performance when compared to older unsupervised techniques. They are also attractive for several other reasons:

- they are well-formed probabilistic generative models, and therefore interpretable;
- they define distributions with a non-finite number of possible outcomes, a desirable property when modeling natural language lexicons: these are not closed sets, and it is difficult to make assumptions regarding the size of a lexicon for an unknown language;
- they are crucially able to produce power law ("Zipfian") distributions over words, a universal prior for natural languages; this is achieved using "rich-get-richer" stochastic processes, in which the more frequent an outcome of the process is, the more likely it is to be generated again in the future.
- they are able to adapt the number of their parameters[17] to the quantity of data available; in other words, these language models are naturally smoothed and less prone to overfitting;
- they benefit from robust inference schemes (Gibbs sampling, Variational Inference) and do not require the design of *ad hoc* search heuristics to be computationally tractable.[18]

2.4.1 Stochastic processes

We introduce, in this section, a series of stochastic processes necessary to define several nonparametric Bayesian language models we use later on in this thesis. We keep notations from Goldwater (2006), our main source for this presentation.

Chinese restaurant process An important stochastic process for nonparametric Bayesian language models is the so-called *Chinese restaurant process* (CRP), which

[17] The term *nonparametric* refers to that ability, and does not mean that these models have no parameters.

[18] The training of these models, however, often requires a lot of computation, a fact that – similarly to artificial neural networks – long hindered the applicability of these models for realistic corpora. This was only overcome in the past two decades with the advent of faster computing units, but most of the theoretical foundations for these methods arose the 1970s or earlier.

generates partitions of integers. The analogy goes as follows: each customer i (represented as an integer) sequentially enters a restaurant with an infinite number of tables, each table accommodating a potentially infinite number of customers. When customer i enters, an arrangement $\mathbf{z}_{-i}$ of the previous customers is observed, with $K(\mathbf{z}_{-i})$ non-empty tables, each already accommodating $n_k(\mathbf{z}_{-i})$ customers for $k \in [1, K(\mathbf{z}_{-i})]$. The customer either seats at a non-empty table with probability $P(z_i = k \,|\, \mathbf{z}_{-i})$, or chooses a new one with probability $P(z_t = K(\mathbf{z}_{-i})+1 \,|\, \mathbf{z}_{-i})$. These terms are defined as follows:

$$P(z_i = k \,|\, \mathbf{z}_{-i}) = \begin{cases} \frac{n_k(\mathbf{z}_{-i})}{i-1+\alpha} & \text{if } 1 \leq k \leq K(\mathbf{z}_{-i}) \\ \frac{\alpha}{i-1+\alpha} & \text{if } k = K(\mathbf{z}_{-i}) + 1\,, \end{cases} \tag{2.6}$$

with $\alpha \geq 0$, a parameter of the process called the concentration[19] parameter. Larger values for this parameter result in a tendency towards opening more new tables, hence a more uniform distribution of customers across the tables, and more clusters in the partition produced. It is also clear that a "rich-get-richer" effect will ensue from this definition of the CRP, and that this effect will get stronger as α gets smaller.

The probability of a given sequence of table assignments $\mathbf{z}$ for n customers is given by:

$$P(\mathbf{z}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \cdot \alpha^{K(\mathbf{z})} \cdot \prod_{k=1}^{K(\mathbf{z})} (n_k(\mathbf{z}) - 1)!\,, \tag{2.7}$$

with $K(\mathbf{z})$ the total number of tables in the arrangement $\mathbf{z}$, and $n_k(\mathbf{z})$ the number of customers at table k in this arrangement. The Gamma function is defined by $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$ for $x > 0$.

Dirichlet process A Dirichlet process $\mathrm{DP}(\alpha, G_0)$, with a concentration parameter α and a base distribution G_0, is a stochastic process whose sample path is a probability distribution over a measurable set S. For any partition $B_1, \ldots, B_n$ of S, if $X \sim \mathrm{DP}(\alpha, G_0)$ then

$$(X(B_1), \ldots, X(B_n)) \sim \mathrm{Dir}(\alpha G_0(B_1), \ldots, \alpha G_0(B_n))\,, \tag{2.8}$$

where $\mathrm{Dir}(\cdot)$ is the Dirichlet distribution.

In an alternative view of the Dirichlet process, a *stick-breaking process* makes the fact that the DP generates discrete distributions with a countably infinite support more explicit. In this view, the base distribution of the DP distributes independently the *locations* of the probability mass function. The α parameter, on its end, influences the probability of each of these locations: a series of independent random variables β_k are drawn sequentially from a $\mathrm{Beta}(1, \alpha)$ distribution; β_1 breaks the unit "stick", and this portion is the probability mass for the first location drawn from the base distribution; β_2 breaks the remaining portion of the stick, and this defines the probability mass of

[19] Maybe the term *dispersion*, sometimes used in place of *concentration*, can seem more natural since higher values of this parameter lead to a higher dispersion of the customers across the tables. The standard terminology (that we keep) comes from the fact that in a Dirichlet process $\mathrm{DP}(\alpha, G_0)$, G_0 corresponds to the mean of the process, and higher values of α lead to distributions that are closer to, or more *concentrated* around, G_0.

the second location, etc. This ensures that the total probability mass will be 1. This also gives another intuition as to why small values of α will lead to more "concentrated" probability mass and sparser distributions.

Another intuitive way to understand the Dirichlet process is to look at it as a "two-stage" CRP model, in which 1) customers are seated according to a CRP process with a certain concentration parameter α as defined by Equation (2.6), and 2) each new opened table is then labelled with a draw from a distribution G_0. This two-stage CRP model is equivalent[20] to a Dirichlet process with concentration parameter α and base distribution G_0. Goldwater (2006) calls the CRP in the first step the **adaptor**, and the distribution G_0 in the second step the **generator**.

Pitman-Yor process A Pitman-Yor process $\text{PYP}(\alpha, \beta, G_0)$ is a generalization of the Dirichlet process $\text{DP}(\alpha, G_0)$. In the two-stage view of the DP, the CRP adaptor is modified in such a way that the conditional probability for the i^{th} customer to seat at table k is defined by:

$$P(z_i = k \,|\, \mathbf{z}_{-i}) = \begin{cases} \frac{n_k(\mathbf{z}_{-i}) - \beta}{i - 1 + \alpha} & \text{if } 1 \leq k \leq K(\mathbf{z}_{-i}) \\ \frac{K(\mathbf{z}_{-i})\beta + \alpha}{i - 1 + \alpha} & \text{if } k = K(\mathbf{z}_{-i}) + 1\,, \end{cases} \tag{2.9}$$

with $0 \leq \beta < 1$, $\alpha > -\beta$ (α corresponds to the concentration parameter in the standard CRP), and $K(\mathbf{z}_{-i})$ the number of tables already occupied when the i^{th} customer enters the restaurant. The new parameter, β, of the Pitman-Yor process, gives more control over the shape of the tail of the distributions generated by the process. It allows to "save" some probability mass to augment the likelihood of opening new tables, even as the number of customers grows and tends to decrease the probability to open a new table in the equivalent DP. Hence, the β parameter is often called the *discount* parameter of the PYP.

Application to word generation If G_0 (the *generator* in the two-stage view of the DP) corresponds to a distribution defined over a lexicon, this means that tables in the restaurant will be labeled with words and that each customer entering the restaurant will represent a word token. The CRP (the *adaptor*, responsible for assigning customers to tables) will, on the other hand, control word frequencies according to a power-law distribution. We will define such language models more formally in Section 2.4.3. Note that PYP language models differ from DP language models only in the definition of the adaptor (Equation (2.6) vs. Equation (2.9)).

Sometimes confusing is that the generator in a DP (or a PYP) can generate duplicated labels, in other words multiple tables can share the same label in the Chinese restaurant analogy. In fact, the expected number of CRP tables (in a DP) for a type corresponding to n tokens is $\alpha \log \frac{n+\alpha}{\alpha}$ (Antoniak, 1974, cited by Goldwater (2006)). For a given number of tokens corresponding to a particular type, the average number of tables labelled by that type grows with α. This intuitively agrees with the dispersion effect of α already mentioned, with greater values of α leading to the opening of more tables.

[20] The technical explanation can be found in Section 3.6 of (Goldwater, 2006).

2.4.2 Sampling

Bayesian approaches are interested in taking into account the full posterior distributions for the parameters of the model instead of a point-wise estimate (like MAP or ML estimates) usually obtained via an EM procedure when supervision data are lacking. Since the posterior distribution is usually impossible to express analytically, sampling algorithms known as Markov Chain Monte Carlo (MCMC) methods are used. Following (Goldwater, 2006), we give the main ideas these sampling methods are built upon.

MCMC methods Such sampling algorithms involve building a Markov chain, with random states $Y^1 \dots Y^T$ and transition matrix $\mathbf{P}$, with proper conditions ensuring its convergence to a unique stationary distribution Φ over the states satisfying $\Phi\mathbf{P} = \Phi$. The states of the Markov chain correspond to an assignment of values to the random variables we want to sample from, and the state space of the Markov chain corresponds to the hypothesis space of the model.

Proper construction of $\mathbf{P}$ guarantees Φ will be the distribution we are interested to infer. After convergence at time T_c, $\forall t \geq T_c$, Y^t will be a sample of the distribution of interest. The conditions on the Markov chain are the following:

- the chain needs to be irreducible (existence of a finite path with non-zero probability between all pairs of states);
- the chain also needs to be aperiodic (together with the preceding condition, this defines an "ergodic" chain);
- $\mathbf{P}$ needs to satisfy $\Phi\mathbf{P} = \Phi$ when Φ is the distribution we want to sample from. This is the "general balance" condition.

Gibbs sampling A particular algorithm build on these principles is Gibbs sampling. If we decompose each state variable Y^t into its K components $Y_1^t \dots Y_K^t$ corresponding to different variables in the model, each iteration of this sampler corresponds to K steps. In each of these steps, the k^{th} component Y_k^t is sampled from its conditional distribution given the current values of all the other components.

To guarantee ergodicity, we need to avoid cases where the conditional probabilities take null values – for example when changing the value of a random variable without also changing the value of another variable could lead to a state with zero probability. To address this problem, it is possible to "block" a Gibbs sampler, sampling a block of variables at once instead of separately, which can also improve convergence speed.

A number of samples are typically ignored at the beginning of the sampling process, during the "burn-in" period necessary for the Markov chain to converge. Importantly, successive samples will be correlated and researchers often retain only a fraction of non-neighbouring samples, or average their results over distinct runs of the Gibbs sampler.

Exchangeability A sequence of random variables $Z_1, \dots, Z_n$ is exchangeable if the joint probability of the sequence is not changed by a permutation of the indices of the sequence, in other words if, for any permutation σ,

$$P(Z_1, Z_2, \dots, Z_n) = P(Z_{\sigma(1)}, Z_{\sigma(2)} \dots, Z_{\sigma(n)}).$$

Note that exchangeability is related, but distinct, to the concept of a series of independent and identically distributed (i.i.d.) random variables.[21]

A sequence of variables distributed according to a distribution itself drawn from a Dirichlet Process has this property of exchangeability, which is crucial to perform inference using Gibbs sampling efficiently: any assignment of a component can be made under the assumption that this component is the last one in the sequence. This way, one can avoid recomputing counts for the part of the sequence occurring after the currently assigned component.

2.4.3 Goldwater's language models

In this section, we formally present two language models based on the Dirichlet process and introduced by Goldwater et al. (2006a). These models will turn out to be very strong baselines for word segmentation in our experiments (see Chapter 3).

Unigram model The first model proposed by Goldwater et al. (2006a) is a *unigram language model*. The unigram assumption usually means that the terms from the product in Equation (2.1) are approximated by $P(w_i \,|\, w_1, \ldots, w_{i-1}) \triangleq P(w_i)$, and that the probabilities of words appearing in a sequence are independent of each other. In the present case, the independence assumption is conditional to a given draw from a DP.

In Goldwater's unigram model indeed, the words are distributed according to a draw G_1 from a Dirichlet process $\mathrm{DP}(\alpha_1, G_0)$,[22] with $\alpha_1 \geq 0$ a parameter of the model, and G_0 a (uniform) unigram distribution over characters (or phonemes).[23] We can write these assumptions as:

$$\begin{aligned} w_i \,|\, G_1 &\sim G_1 \\ G_1 \,|\, \alpha_1, G_0 &\sim \mathrm{DP}(\alpha_1, G_0)\,, \end{aligned} \tag{2.10}$$

where "$A \,|\, B \sim C$" reads as "A given B is distributed according to C". The probability of a sentence under this language model is consequently given by:[24]

$$P_{\text{unigram}}(w_1, \ldots, w_I) \triangleq \prod_{i=1}^{I} P(w_i \,|\, G_1). \tag{2.11}$$

Unfortunately, as G_1 has an infinite support, it is not possible to sample w_i from G_1. It is possible, however, to integrate over G_1 to obtain the conditional probability of w_i given the previously generated words (Blackwell and MacQueen, 1973, cited by Goldwater (2006)):

$$P(w_i = w \,|\, \mathbf{w}_{-i}, \alpha_1, G_0) = \frac{n_w(\mathbf{w}_{-i}) + \alpha_1 G_0(w)}{i - 1 + \alpha_1}\,, \tag{2.12}$$

[21] De Finetti's theorem states that exchangeable observations are conditionally independent given some latent variable.

[22] Recall that a draw from a DP is a probability distribution.

[23] That is, $P_{\mathrm{G}_0}(w) = (1 - p_b)^{K-1} p_b \prod_{k=1}^{K} P(u_k)$, with $w = u_1, \ldots, u_K$, and p_b the probability to generate a word boundary. $P(u_k)$ is distributed uniformly over characters or phonemes.

[24] We ignore sentence boundaries in this presentation.

with $\mathbf{w}_{-i} = w_1, \ldots, w_{i-1}$, and $n_w(\mathbf{w}_{-i})$ the number of times the word w is seen in $\mathbf{w}_{-i}$. Note that, without conditioning on G_1, successive words are no more independent.

With these posterior distributions, it is possible to calculate a probability for the segmentation of an unsegmented sequence (of characters or phonemes). The Gibbs sampler proposed by Goldwater considers all possible word boundaries, and successively samples between two hypotheses for each boundary position: one where a boundary is present, and the other where the boundary is absent (all other boundaries in the corpus are kept identical). At the end of an iteration of the sampler, all boundary positions in the corpus have been considered once.[25]

Bigram model The limitations of the unigram model, and its tendency to undersegment an input character sequence, as it tries to capture frequently co-occurring bigrams as single words, led Goldwater to develop a *bigram language model* capturing the effect of the context on word generation. In the traditional language model formulation, a bigram dependency means that the terms in Equation (2.1) are approximated by $P(w_i \,|\, w_1, \ldots, w_{i-1}) \triangleq P(w_i \,|\, w_{i-1})$. In Goldwater's formulation, the generation of successive words w_i now relies on a more involved generative process involving a hierarchical Dirichlet process (Teh, 2006), or HDP, that we describe now.

The bigram distributions $P(w_i \,|\, w_{i-1} = w, G_w)$, for each word form w in the left context (or *history*), are distributed according to a draw G_w from a Dirichlet process $\mathrm{DP}(\alpha_2, G_1)$. The base distribution, G_1, of this DP is itself drawn from a Dirichlet process $\mathrm{DP}(\alpha_1, G_0)$. Lastly, identically to the unigram model, G_0 is a unigram distribution over characters. Summing it up, the generative process for words is defined by

$$\begin{array}{lcll} w_i \,|\, w_{i-1} = w, G_w & \sim & G_w & \forall w \\ G_w \,|\, \alpha_2, G_1 & \sim & \mathrm{DP}(\alpha_2, G_1) & \forall w \\ G_1 \,|\, \alpha_1, G_0 & \sim & \mathrm{DP}(\alpha_1, G_0)\,, & \end{array} \tag{2.13}$$

and the probability of a sequence of words is given by:[26]

$$P_{\mathrm{bigram}}(w_1, \ldots, w_I) \triangleq \prod_{i=1}^{I} P(w_i \,|\, w_{i-1}, G_{w_{i-1}})\,, \tag{2.14}$$

assuming w_0, a beginning-of-sentence symbol.

As with the unigram language model, G_1 and all G_w have non-finite supports and are integrated out, leading to the following conditional probabilities:

$$\begin{aligned} P(w_i \,|\, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \alpha_1, \alpha_2, G_0) &= \frac{n_{(w_{i-1}, w_i)}(\mathbf{w}_{-i}) + \alpha_2 P_{\mathrm{backoff}}(w_i \,|\, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \alpha_1, G_0)}{n_{w_{i-1}}(\mathbf{w}_{-i}) + \alpha_2} \\ P_{\mathrm{backoff}}(w_i \,|\, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \alpha_1, G_0) &= \frac{t_{w_i}(\mathbf{w}_{-i}, \mathbf{z}_{-i}) + \alpha_1 P(w_i \,|\, G_0)}{t_{all}(\mathbf{z}_{-i}) + \alpha_1}\,, \end{aligned} \tag{2.15}$$

[25] In the sketch for Gibbs sampling given in Section 2.4.2, each of the K components is an assignment to the random variable corresponding to the presence or absence of a word boundary at a particular position in the data.

[26] Ignoring again the sentence boundaries. Note that, for simplicity, we abuse the notation, as we denote by w_{i-1} both the random variable and its realization.

with $n_{(w_{i-1},w_i)}(\mathbf{w}_{-i})$ the number of bigram tokens (w_{i-1}, w_i), $n_{w_{i-1}}(\mathbf{w}_{-i})$ the number of tokens w_{i-1},[27] $t_{w_i}(\mathbf{w}_{-i}, \mathbf{z}_{-i})$ the number of bigram tables labelled w_i, and $t_{all}(\mathbf{z}_{-i})$ the total number of bigram tables. Here, $\mathbf{z}_{-i}$ corresponds to the seating arrangement in the bigram restaurants. P_{backoff} corresponds to the posterior estimate of base distribution G_1, and can be seen as a unigram backoff (see Goldwater, 2006, section 5.5.1).

To understand these equations, it is important to realize that in the specification of the model, each word type w has its own restaurant, and that this "bigram restaurant" corresponds to the distribution of tokens following w ($G_w \sim \text{DP}(\alpha_2, G_1)$ in Equation (2.13)). When a new table is opened in a given bigram restaurant, a label is drawn from G_1, which corresponds to the so-called "backoff" restaurant ($G_1 \sim \text{DP}(\alpha_1, G_0)$ in Equation (2.13)). All this means that customers in the bigram restaurants correspond to bigram tokens, and that customers in the backoff restaurant correspond to labels on bigram tables. Note that in the bigram model, and contrary to the unigram model, the seating arrangement $\mathbf{z}_{-i}$ matters to calculate the conditional probability of the words.

Similarly to the unigram model (yet with an increased complexity), a Gibbs sampler can be built to infer segmentations from an unsegmented corpus.

2.4.4 Nested language models

The hierarchical nature of Goldwater's bigram model can be further extended to n-gram dependencies. Mochihashi et al. (2009) adopt this stance, and replace the Dirichlet process by the more general Pitman-Yor process, proposing a model they call the *nested Pitman-Yor language model*. Under a n-gram version of this language model, the distribution of words follows the hierarchical scheme:

$$\begin{aligned}
w_i \,|\, w_{i-n+1}^{i-1}, G_{w_{i-n+1}^{i-1}} &\sim G_{w_{i-n+1}^{i-1}} && \forall w_{i-n+1}^{i-1} \\
G_{w_{i-n+1}^{i-1}} \,|\, \alpha_n, \beta_n, G_{w_{i-n+2}^{i-1}} &\sim \text{PYP}(\alpha_n, \beta_n, G_{w_{i-n+2}^{i-1}}) && \forall w_{i-n+1}^{i-1} \\
G_{w_{i-n+2}^{i-1}} \,|\, \alpha_{n-1}, \beta_{n-1}, G_{w_{i-n+3}^{i-1}} &\sim \text{PYP}(\alpha_{n-1}, \beta_{n-1}, G_{w_{i-n+3}^{i-1}}) && \forall w_{i-n+2}^{i-1} \\
&\dots && \\
G_{w_{i-2}^{i-1}} \,|\, \alpha_3, \beta_3, G_{w_{i-1}} &\sim \text{PYP}(\alpha_3, \beta_3, G_{w_{i-1}}) && \forall w_{i-2}^{i-1} \\
G_{w_{i-1}} \,|\, \alpha_2, \beta_2, G_1 &\sim \text{PYP}(\alpha_2, \beta_2, G_1) && \forall w_{i-1} \\
G_1 \,|\, \alpha_1, \beta_1, G_0 &\sim \text{PYP}(\alpha_1, \beta_1, G_0)\,, &&
\end{aligned} \tag{2.16}$$

with $w_{i-n+1}^{i-1} = w_{i-n+1}, \dots, w_{i-1}$, the $(n-1)$ words in the left context of w_i.

Additionally, Mochihashi et al. (2009) replace the base distribution on characters, G_0, by a second hierarchical Pitman-Yor process, a *spelling model*, whose m-gram structure is equivalent to that of their language model. To make their nested model tractable, the authors supplement it with a blocked Gibbs sampler, using an efficient forward filtering and backward sampling procedure, as well as a Poisson correction for word length. They report faster inference and better accuracy than Goldwater's bigram model.

A further extension can be found in (Neubig et al., 2010), in which a similarly nested Pitman-Yor language model is used to learn word segmentation from phoneme lattices.

[27] This is also the total number of customers in the bigram restaurant w_{i-1}, in other words all bigram tokens beginning with w_{i-1}.

The main novelty of this work is to reinterpret the model of Mochihashi et al. (2009), i.e. the hierarchical Pitman-Yor language and spelling models, in terms of a weighted finite state acceptor (WFSA) in charge of assigning the proper posterior probability to a given segmentation. This acceptor can be composed with a phoneme lattice encoding acoustic model scores and with a weighted finite state transducer (WFST) transducing any sequence of phonemes into all of its possible segmentations.

Another generalization (Löser and Allauzen, 2016) introduces some morphotactics through word classes: in this model, sentences are produced by a nonparametric Markov model, where both the number of states and the number of types are automatically adjusted based on the available data. Two hierarchical Pitman-Yor processes are also embedded in this architecture: one for controlling the number of classes (states) and one for controlling the number of words. As in Mochihashi et al. (2009), the base distribution is also a hierarchical PYP spelling model.

2.4.5 Adaptor Grammars

We conclude this section with the presentation of the Adaptor Grammar (AG) framework, built on the concepts of adaptors and generators introduced in Section 2.4.1 (the "two-stage" view of the Dirichlet process). The goal is to learn and infer structure both at word and subword levels, thus combining the learning of morphological structures with the learning of word dependencies, and ultimately learning how to segment sentences into words, and words into morphemes. Our presentation follows (Johnson et al., 2007b) and (Johnson and Goldwater, 2009).

Framework Adaptor Grammars (Johnson et al., 2007b) are an extension to probabilistic context-free grammars (PCFGs) relaxing the assumption that each subtree of a nonterminal node is generated independently from other subtrees rooted in the same nonterminal.

Formally, if we consider a PCFG defined by the quintuple (N, W, R, S, θ), where N is a finite set of nonterminal symbols, W a finite set of terminal symbols disjoint from N, R a finite set of rules of the form $A \to \beta$, with $A \in N$ and $\beta \in (N \cup W)^*$, S a particular nonterminal start symbol, and θ defining probabilities for production rules associated to each nonterminal, the distributions G_A over trees rooted in nonterminal[28] symbols A are defined through the following recursion:

$$G_A = \sum_{A \to B_1 \ldots B_n} \theta_{A \to B_1 \ldots B_n} \, \mathrm{TD}_A(G_{B_1} \ldots G_{B_n}) \tag{2.17}$$

with each particular tree distribution $\mathrm{TD}_A(G_{B_1} \ldots G_{B_n})$ defined by

$$\mathrm{TD}_A(G_{B_1} \ldots G_{B_n}) \left(\begin{array}{ccc} & A & \\ t_1 & \ldots & t_n \end{array} \right) = \prod_{i=1}^{n} G_{B_i}(t_i) \,, \tag{2.18}$$

with t_i a tree rooted in B_i.

[28] For terminal symbols A, G_A is the distribution putting all of its mass on the single node labelled A.

To define an Adaptor Grammar from this PCFG, we consider the *adaptors* C_A for $A \in N$, with each C_A defined as a function associating the distribution G_A to a distribution over distributions having the same support as G_A. The recursion defining the new distribution H_A over nonterminal symbol A is given by:

$$\begin{aligned} H_A &\sim C_A(G_A) \\ G_A &= \sum_{A \to B_1 \dots B_n} \theta_{A \to B_1 \dots B_n} \, \mathrm{TD}_A(H_{B_1} \dots H_{B_n}) \,. \end{aligned} \tag{2.19}$$

For example, the adaptor C_A can be the function associating a distribution G_A to the Dirichlet process $\mathrm{DP}(\alpha_A, G_A)$ (see Section 2.4.1) with concentration parameter α_A and base distribution G_A. It is also possible to define C_A as the identity[29] function for certain "non-adapted" nonterminals. If the set of adapted nonterminals is denoted by M, the Adaptor Grammar associating the distributions H_A over each nonterminal A is finally defined by:

$$\begin{aligned} &\begin{cases} H_A \sim \mathrm{DP}(\alpha_A, G_A) & \text{if } A \in M \\ H_A = G_A & \text{if } A \notin M \end{cases} \\ &G_A = \sum_{A \to B_1 \dots B_n} \theta_{A \to B_1 \dots B_n} \, \mathrm{TD}_A(H_{B_1} \dots H_{B_n}) \,. \end{aligned} \tag{2.20}$$

Using these adaptors in the recursion, it is possible to allow for a symbol's expansion to depend on the way it has been rewritten in the past. Informally, Adaptor Grammars are able to "cache" entire subtrees expanding nonterminals and provide a choice to rewrite each new nonterminal either as a regular PCFG expansion or as a previously seen expansion. In this respect, Adaptor Grammars can be seen as a nonparametric extension of PCFGs. Moreover, using adaptors based on the Dirichlet process or the Pitman-Yor process, one can build models capturing power-law distributions over trees and subtrees.

Inference The training data in the unsupervised context consists only in terminal strings (*yields* of trees rooted in the start symbol S). In order to sample posterior distributions over analyses produced by a particular Adaptor Grammar based on Pitman-Yor processes, Johnson et al. (2007b) devise a method relying on a Markov chain Monte Carlo algorithm (see Section 2.4.2) together with a PCFG approximation[30] of the Adaptor Grammar. The idea for this approximation is, for each analyzed string, to add to the rules of the "base" PCFG all the production rules corresponding to the yields of the adapted nonterminals in the Adaptor Grammar, given all the analysed strings in the data set, except the currently analysed string. One can sample analyses from this PCFG using the algorithm described in (Johnson et al., 2007a).

The inference procedure is described by the following main steps:

[29] More precisely, as a function mapping G_A to the distribution placing all its mass on G_A in the space of distributions.

[30] After relaxing the independence assumption made in PCFGs, there is, according to Johnson et al. (2007b), no efficient direct sampling procedure from $P(u_i|s_i, \mathbf{u}_{-i})$, with u_i the analysis of the i^{th} string s_i in the data, and $\mathbf{u}_{-i}$ the vector of all the analyses except analysis u_i, required to perform a MCMC procedure.

1. Initialize with a random tree generated by the grammar for each string,
2. Randomly select a string and sample a parse from the PCFG approximation,
3. Update the parse for this string if the Metropolis-Hastings procedure accepts the proposed analysis,[31]
4. Go back to step 2 until convergence. At convergence, the analyses are samples of the posterior distribution over analyses under the Adaptor Grammar, and can be used to compute the production probabilities θ.

Expressivity The Adaptor Grammar framework's main strength lies in its flexible and powerful expressivity. It is possible to replicate under this framework equivalent or similar nonparametric models for word segmentation (Goldwater et al., 2006a) or morphology learning (Goldwater et al., 2006b). For example, the unigram model of Goldwater (see Section 2.4.3) corresponds to the following production rules in the Adaptor Grammar (the adapted nonterminal is underlined):

Words → Word
Words → Word Words
Word → Phonemes
Phonemes → Phoneme
Phonemes → Phoneme Phonemes

More importantly, Adaptor Grammars allow us to infer, in a single procedure, over structures that are mutually dependent, for example word boundaries and word-initial syllable collocations. In other words, it is possible to learn simultaneously something about the structure of an utterance and the structure of the words composing it. This requires to specify more than one adapted nonterminal in the grammar, which turns out to be equivalent to implementing a hierarchical Dirichlet process (HDP). Hence, integrating a learning procedure for morphology into the unigram word-based grammar above, would consist in adding, for instance, rules of the form:

Word → Stem
Word → Stem Suffix
Stem → Phonemes
Suffix → Phonemes

while removing the "Word → Phonemes" production.

It should be noted however that the expressivity of this framework presents some limits, since the number of adaptors is required to be fixed in advance and corresponds to the number of nonterminals. Goldwater's bigram model for example, associates one Dirichlet process per word type, and the number of these types is not known in advance. Johnson (2008b) shows, nonetheless, that introducing an adapted nonterminal for word collocations allows to capture inter-word dependencies, and achieves similar performance to Goldwater's bigram model.

[31] This step corrects the probability approximation made using the PCFG "snapshot" of the Adaptor Grammar.

Applications of the framework Experiments on an English corpus of child-directed speech are performed in (Johnson, 2008b; Johnson and Goldwater, 2009; Johnson et al., 2014) with successive improvements relying on increasingly complex grammars, taking into account phonotactic constraints, and different levels of collocations, together with refined initialization, advanced sampling techniques,[32] and modeling of function words.

Of particular interest to us, Johnson (2008a) looks at unsupervised morphology learning for a Bantu language, Sesotho.[33] This is mostly an experimental work exploring various ways to express morpho-phonological knowledge into the formalism of Adaptor Grammars. One interesting outcome of this study is to show the effectiveness of having an explicit hierarchical model of word internal structure for Sesotho, a language with a complex morphology, and to demonstrate the applicability of the framework to various types of morphology.

Extensions of adaptor grammars More recent work has been building on the techniques presented in the preceding sections. O'Donnell et al. (2009), elaborating on the idea of a heterogeneous lexicon, introduce Fragment Grammars, a generalization of Adaptor Grammars in which fragments of subtrees can be adapted – and not only entire subtrees yielding terminal strings –, that is to say, the distribution of a subtree prefix can be learnt in this framework. It is not clear however to which degree this extension compares to state-of-the-art results on standard tasks. Botha and Blunsom (2013) propose another extension to the Adaptor Grammar formalism using a probabilistic and adapted version of Simple Range Concatenating Grammars (SRCGs) attempting to capture non-concatenative phenomena in morphology and obtaining improvements in a task of morphological segmentation for Semitic languages (in this case Arabic and Hebrew).

We should also mention the work of Cohen et al. (2010) who devise a variational inference algorithm which provides an alternative to the MCMC method used by Johnson et al. (2007b). Zhai et al. (2014) combine both methods in an online and hybrid fashion, significantly improving the inference's speed for Adaptor Grammars. Synnaeve et al. (2014), lastly, take advantage of non-linguistic context to improve word segmentation using an Adaptor Grammar. This non-linguistic context (activity at stake, visual cues) is approximated as a topic obtained via the training of a topic model (Blei et al., 2003). The most probable topic of each utterance is then added as a prefix and the grammar is modified to make use of them. This proves to be helpful for the task of word segmentation. Another attempt to guide the learning process can also be found in (Börschinger and Johnson, 2014), proposing to model the role of stress cues in language learning. Lastly, Lee et al. (2015) propose to use Adaptor Grammars in conjunction with an acoustic model to jointly learn phoneme and word-like units directly from speech.

[32] Resampling of the table labels within the Gibbs sampling procedure, sampling of the adaptor's hyperparameters, and integrating out the production rules' probabilities.

[33] The task is to segment in words children productions.

2.5 Automatic word alignment

Historically, the concept of alignment emerges at the beginning of the 1990s in the context of the first word-based probabilistic methods to be applied to machine translation (MT) by IBM researchers (Brown et al., 1993). Our presentation follows (Allauzen and Yvon, 2011) and (Brunning, 2010), and aims at providing the reader with the basic theory underlying word alignment.

2.5.1 Probabilistic formulation

With the notations introduced in Section 2.1, the probability $P(\boldsymbol{\omega} \,|\, \mathbf{w})$ to observe the target[34] word sequence $\boldsymbol{\omega} = \omega_1^J = \omega_1, \ldots, \omega_J$ conditionally to the observation of the source word sequence $\mathbf{w} = w_1^I = w_1, \ldots, w_I$ is modeled by marginalizing a hidden alignment variable A with possible outcomes in $\mathcal{A}$:

$$P(\boldsymbol{\omega} \,|\, \mathbf{w}) = \sum_{A \in \mathcal{A}} P(\boldsymbol{\omega}, A \,|\, \mathbf{w}) \,. \tag{2.21}$$

For computational reasons already invoked (see Section 2.1.1.2), Brown et al. (1993) restrict $\mathcal{A}$ to vectors of the form $\mathbf{a} = a_1^J = a_1, \ldots, a_J$, in which each a_j indicates the word position in $\mathbf{w}$ to which target word ω_j is aligned to. A special `null` word w_0 is added to the source sequence to account for unaligned target words.

Under this theoretical framework, it is only possible to produce "1-1" or "1-n" alignments from source to target,[35] but no "m-n" alignment type (see Figure 2.7 for an illustration of these types). Therefore, these alignments are often called asymmetrical. Machine translation systems[36] using word alignments produced by IBM models typically train an IBM model from source to target, and from target to source, in order to build symmetrical alignments using various heuristics (Och and Ney, 2000; Koehn et al., 2003).

2.5.2 A series of increasingly complex parameterizations

To parametrize $P(\boldsymbol{\omega}, \mathbf{a} \,|\, \mathbf{w})$, and make training and inference tractable, various parametrizations have been proposed, leading to specific word alignment models. To have a better intuition about these models, it is useful to consider the following decomposition into two sub-models. The first sub-model influences the *distortion* (i.e. the word order) in the translation process, while the second sub-model is in charge of the *translation* itself:

$$P(\boldsymbol{\omega}, \mathbf{a} \mid \mathbf{w}) = \underbrace{P(\mathbf{a} \,|\, \mathbf{w})}_{\text{distortion}} \underbrace{P(\boldsymbol{\omega} \,|\, \mathbf{a}, \mathbf{w})}_{\text{translation}} \,. \tag{2.22}$$

[34] As mentioned in footnote 4 on p. 19, this can be confusing as the literature on word alignment often refers to $\boldsymbol{\omega}$ as the *source* sequence. This is because, for the purpose of machine translation, the original probability $P(\mathbf{w} \,|\, \boldsymbol{\omega})$ is envisioned through a noisy channel model $P(\mathbf{w} \,|\, \boldsymbol{\omega}) \propto P(\boldsymbol{\omega} \,|\, \mathbf{w}) P(\mathbf{w})$, with $P(\boldsymbol{\omega} \,|\, \mathbf{w})$ becoming the distribution of interest for automatic alignment. However, in the absence of a noisy channel model here, we choose the natural terminology.

[35] At most one outbound link per target word.

[36] In particular phrase-based machine translation systems, nowadays largely supplanted by neural machine translation (NMT).

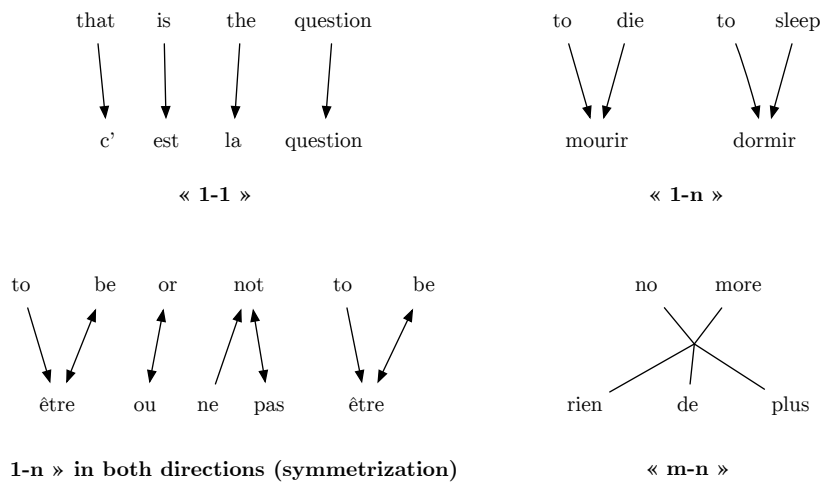


Figure 2.7: Alignment types for IBM models. The "m-n" type cannot be produced by these models. For each alignment, French source is represented at the bottom, and English target, at the top; arrows indicate target to source links, and in the symmetrization case, alignments are computed in both directions.

Additionally, it is possible, without any loss of generality, to rewrite this probability using the chain rule:

$$P(\boldsymbol{\omega}, \mathbf{a} \mid \mathbf{w}) = P(J \mid \mathbf{w}) \prod_{j=1}^{J} P(a_j \mid a_1^{j-1}, \omega_1^{j-1}, J, \mathbf{w}) P(\omega_j \mid a_1^{j}, \omega_1^{j-1}, J, \mathbf{w}) \,. \tag{2.23}$$

The models we present next correspond to hypotheses akin to simplify this factorization.

IBM Model 1 In this model, the hypothesis made is that each a_j is chosen independently from the others (a), and drawn uniformly (b) amongst all possible positions in the source sentence:

$$P(\mathbf{a} \,|\, \mathbf{w}) \overset{\text{(a)}}{\triangleq} \prod_{j=1}^{J} P(a_j \,|\, \mathbf{w}) \overset{\text{(b)}}{\triangleq} \frac{1}{(I+1)^J} \,. \tag{2.24}$$

Once alignment links are determined, target words only depend on the source word they are aligned to:

$$P(\boldsymbol{\omega} \,|\, \mathbf{a}, \mathbf{w}) \triangleq \prod_{j=1}^{J} P(\omega_j \,|\, w_{a_j}) \,, \tag{2.25}$$

from which we deduce, ignoring the probability[37] of target sentence length J, the equation describing IBM Model 1:

$$P(\boldsymbol{\omega}, \mathbf{a} \,|\, \mathbf{w}) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} P(\omega_j \,|\, w_{a_j}) \,. \tag{2.26}$$

The probability to generate $\boldsymbol{\omega}$ from $\mathbf{w}$ under this model is subsequently derived from Equation (2.21) by marginalizing the alignment variable.

A problem, known as the *garbage collector* problem, is that this model tends to align rare words from the source sentence to too many words on the target side. Some solutions to this problem are proposed by (Moore, 2004). This model also suffers from not capturing any dependency between the positions of aligned words, which leads to uncontrolled distortions. Word order for source and target words is not taken into account.

IBM Model 2 To remedy this last defect of IBM Model 1, IBM Model 2 introduces a dependency between the value of a_j and the position j of corresponding target word ω_j, by rewriting the first term under the product in Equation (2.23) according to:

$$P(a_j \,|\, a_1^{j-1}, \omega_1^{j-1}, J, \mathbf{w}) \triangleq P(a_j | j, J, I) \,, \tag{2.27}$$

As a result, the model can learn, for instance, that alignment links are located along the diagonal in a matrix representation. Assuming lexical dependencies identically to IBM Model 1, the complete description of IBM Model 2 follows:

$$P(\boldsymbol{\omega}, \mathbf{a} \,|\, \mathbf{w}) = \prod_{j=1}^{J} P(a_j \,|\, j, J, I) P(\omega_j \,|\, w_{a_j}) \,. \tag{2.28}$$

Other IBM models Subsequent models (IBM Model 3, IBM Model 4 and IBM Model 5), also introduced by (Brown et al., 1993), correct – at the cost of a much higher complexity – other limitations of IBM Model 2.

Of particular interest is the introduction in IBM Model 3 of a new latent variable, *fertility*, which allows to model the propensity of certain source words to be aligned with many target words. This concept will be extended by Stahlberg et al. (2012) to better deal with the situation where one tries to align words to phonemes.

More sophisticated position dependencies are handled by IBM Model 4, with some deficiencies further corrected in IBM Model 5.

Reparameterization of IBM Model 2 In IBM Model 2, the definition of the distortion distribution in Equation (2.27) as $P(a_j \,|\, j, J, I)$, results in the estimation of independent parameters for all distributions of alignment positions on the source side, given the current target position, target sentence length, and source sentence length. This leads to sparse observations and less robust estimates.

[37] If the model is used for word alignment and not for translation, the length of target sentences is observed.

To remedy this problem, (Dyer et al., 2013) propose a log-linear reparameterization in which the distortion term is redefined as:

$$P(a_j = i \,|\, j, J, I) = \begin{cases} p_0 & \text{if } i = 0 \\ (1 - p_0) \times \dfrac{e^{\lambda h(i,j,I,J)}}{Z_\lambda(j, I, J)} & \text{if } 0 < i \leq I\,, \end{cases} \tag{2.29}$$

with p_0 a null alignment probability (assuming w_0 a `null` symbol), $\lambda \geq 0$ a "precision" parameter controlling how much the model will favor alignment points close to the diagonal (in a matrix representation of the alignment), and $h(i, j, I, J) = -|\frac{i}{I} - \frac{j}{J}|$. $Z_\lambda(j, I, J)$ is the normalization factor equal to $\sum_{i'=1}^{I} e^{\lambda h(i',j,I,J)}$. When $\lambda \to 0$, the distortion will be almost uniform, and therefore comparable to that of IBM Model 1.

The authors also develop an efficient training scheme for this model, and show its competitiveness with IBM Model 4 – still widely used as a state-of-the-art word alignment model – both in terms of AER (see 2.1.2), and of translation quality (measured as a BLEU score (Papineni et al., 2002)).

HMM model Another appealing parametrization, building on IBM models, was proposed by Vogel et al. (1996), with the intention to capture the local monotonicity[38] usually observed in alignments. This is achieved by re-parametrizing the distortion term (Equation (2.22)) with a first-order Markovian dependency hypothesis:

$$P(\mathbf{a} \,|\, \mathbf{w}) \triangleq \prod_{j=1}^{J} P(a_j \,|\, a_{j-1}, \mathbf{w})\,. \tag{2.30}$$

With an identical hypothesis for the translation term, this gives the following definition for the HMM model:

$$P(\boldsymbol{\omega}, \mathbf{a} \,|\, \mathbf{w}) = \prod_{j=1}^{J} P(a_j \,|\, a_{j-1}, \mathbf{w}) P(\omega_j \,|\, w_{a_j})\,. \tag{2.31}$$

In practice, the term $P(a_j \,|\, a_{j-1}, \mathbf{w})$ is a function of the jump width $|a_j - a_{j-1}|$.

2.5.3 Parameters estimation

Parallel corpora very rarely contain annotations of alignments at the word level. To train a word alignment model, i.e. estimate its parameters, the iterative EM algorithm (Dempster et al., 1977) is used to find the values of the parameters that maximize the likelihood of the data in the absence of observation of the latent variables. The intuition is that when alignments are available, it is possible to estimate the parameters of the model with the method of Lagrange multipliers. Conversely, the maximum-likelihood estimates for the parameters allow to calculate the posterior probability of the alignment links. Initializing randomly the parameters, these posterior probabilities can be used as "pseudo-counts" to re-estimate the parameters of the model so that they maximize the

[38] The fact that consecutive words in the source sequence are likely to be aligned to consecutive words in the target sequence.

(log-)likelihood of the data. With $\mathcal{C} = \{(\mathbf{w}^{(n)}, \boldsymbol{\omega}^{(n)}), n = 1 \dots n_D\}$ a collection of n_D mutually translated sentence, $\mathcal{A}^{(n)}$ the set of possible alignments for the n^{th} pair, and $\boldsymbol{\theta}$ the parameters of the model, the log-likelihood EM seeks to maximize is given by:

$$\ell(\boldsymbol{\theta}) = \log \left(\prod_{n=1}^{n_D} P(\boldsymbol{\omega}^{(n)} \mid \mathbf{w}^{(n)}; \boldsymbol{\theta}) \right) . \tag{2.32}$$

2.5.4 Alignments extraction

Once the EM algorithm has converged, the parameters of the model maximize the likelihood of the observed data. This maximum, however, can be a local optimum as the log-likelihood most often is not concave. IBM Model 1, nonetheless, enables the computation of a global[39] optimum, and its parameters are frequently used to initialize the parameters of more complex models. After training, it is then theoretically possible to extract the alignments using

$$\hat{\mathbf{a}}^{(n)} = \underset{\mathbf{a}^{(n)} \in \mathcal{A}^{(n)}}{\operatorname{argmax}} P(\mathbf{a}^{(n)} \mid \mathbf{w}^{(n)}, \boldsymbol{\omega}^{(n)}) , \tag{2.33}$$

although difficult in practice: simplifying hypotheses or heuristics are required to explore the search space.

Examining all the variants that have been proposed for automatic word alignment is beyond the scope of this brief presentation. It is possible, for instance, to predict alignment links corresponding to a posterior probability greater than a threshold, as in (Liang et al., 2006). In this work, two alignment models, one from source to target and the other from target to source, are learnt jointly; the goal is to make predictions in both directions agree, in order to produce a symmetrical alignment without resorting to heuristics *ex post*. Another example is *posterior regularization* (Graça et al., 2007, 2010; Ganchev et al., 2010), a framework in which certain constraints can be incorporated (symmetry again, but also bijectivity[40]). An in-depth study of early word alignment models is in (Och and Ney, 2003), and a comprehensive survey including many more recent approaches can be found in (Tiedemann, 2011).

2.6 Joint models for segmentation and alignment

Morphological divergences between languages are a major issue for word alignment algorithms, which assume similar concepts of words on both sides of the alignment, hence making the identification of one-to-one correspondences a reasonable goal. When this hypothesis is violated, which happens for instance when one attempts to align an analytic language such as Chinese with a synthetic language, for instance Turkish, alignment performance decreases significantly. This is because alignments from the synthetic to the analytic language tend to leave too many words unaligned on the analytic side. Additionally, rich morphological variations multiply the number of word forms, which

[39] Yet not unique, see (Toutanova and Galley, 2011).

[40] Symmetry means that alignments need to agree in both directions, while bijectivity constrains a word to translate to a single word.

hurts the robustness of the statistical procedures involved in the alignment process. In this section, we survey works aimed at combining ideas from the segmentation and from the alignment literature to mitigate these issues. Note though that the goal here remains to align word and subword units, while in our work we will rather be interested in using alignment as an auxiliary (weak) supervision for segmentation.

Importantly, a lot of theses works do not have computational language documentation in mind, and most of them do not operate on a low-resource scenario;[41] machine translation,[42] in fact, is the inspiration for most of this research, in particular when translating language pairs that are typologically different, or when one or both languages' spelling rules do not overtly mark word boundaries. As we noted in Section 2.1.1.1, the right segmentation granularities for the source and target sides of the corpus are consequently determined for each particular language pair, and on the basis of translation performance rather than linguistic validity.[43]

The most obvious approach, "segment, then align", that we review first, consists in using segmentation or, more generally, morphological analysis, as a preprocessing step before computing alignments. The reverse approach, "align to segment" (see Section 2.1.1 and Figure 2.2b) is rarely practical alone, because units of different granularity (for instance phonemes and words) are extremely difficult to align accurately; the information content per fine-grained unit (e.g. phonemes) is too limited. The second part of this section then focusses on models jointly learning segmentations and alignments. If the influence of machine translation is still pervading, these models also enable, for instance, the extraction of a bilingual lexicon and the identification of morpheme boundaries in one of the languages in the bilingual pair. In other words, as segmentation is often used to alleviate the difficulty to align words, alignments can also be simultaneously used to guide segmentation.

2.6.1 Segment, then align

To reconcile the source and target side notions of words when languages are typologically different, the more synthetic[44] language can be preprocessed, so as to decompose complex lexical forms into shorter segments, or to neutralize morphological variations that are not marked in the other, morphologically simpler, language. Forms that only differ in their case mark can, for instance, be collapsed into one non-marked version for the purpose of aligning to English, where case is not marked.

[41] Although some of these works do address the segmentation (and translation) of phonetic transcripts, or seek the extraction of a bilingual lexicon, fitting more closely to our own goals (Stahlberg et al., 2012; Adams et al., 2015).

[42] More specifically, phrase-based SMT or n-gram-based SMT. NMT also uses various segmentation strategies to reduce sparsity (Costa-jussà and Fonollosa, 2016; Sennrich et al., 2016), but without explicitly performing jointly an alignment.

[43] One could, however, speculate as to whether an equivalent bias exists when two linguists with typologically distant mother tongues document the same language.

[44] The term "morphologically rich" is often used to denote a language in which a large quantity of information is encoded in the morphology of its words (rather than using prepositions, word order, etc.), which leads to large lexicons. We prefer the less ambiguous term *synthetic*, already defined in Section 2.1.3.

Popularity of the approach This strategy has been successfully applied to many language pairs in the context of machine translation applications: (Nießen and Ney, 2001) is a first attempt to cluster morphological variants when translating from German into English, while (Koehn and Knight, 2003) and (Dyer, 2009) are early attempts at splitting German compounds; see also (Durgar El-Kahlout and Yvon, 2010; Fraser et al., 2012) for other studies of translation from or into German. By the same token, Fishel and Kirik (2010) use unsupervised morphological analysis on an Estonian-English corpus to perform alignment on several variants of the lemmatized Estonian part of the corpus. Similar techniques have been proposed for other language pairs such as, for instance, Czech (Goldwater and McClosky, 2005), Arabic (Habash and Sadat, 2006), Spanish (de Gispert and Mariño, 2008), Finnish (Virpioja et al., 2007), Turkish (Oflazer and Durgar El-Kahlout, 2007) to name a few studies. Alternatively, (Burlot and Yvon, 2015) propose a factored model where morphological features (on the Czech side) can be aligned to (English) words; Burlot and Yvon (2017) then proceed to normalize the synthetic (Czech or Russian) side automatically by clustering forms that translate into the same target (English) word or words. Note that the reverse approach has also been attempted, splicing English words into complex forms in (Ueffing and Ney, 2003), or removing function words on the English side to attach them, as syntactic tags, to the corresponding content words (Yeniterzi and Oflazer, 2010).

With a slightly different goal in mind – a proof-of-concept speech translation system designed for unwritten languages – Besacier et al. (2006) follow a similar approach: in their system, the target sequence consists in phones, which are segmented without any supervision into multi-phone units, using mutual information between successive phones. When mutual information reaches a local minimum, a morph boundary is detected.[45] Discovered morphs are then aligned to source words using an instance of IBM Model 4, and subsequently translated into the source language via a phrase-based SMT system.

Challenges and improvements As noted by several authors, decomposing word forms into morphemes goes against the main intuition of phrase-based SMT (Koehn, 2010), which favors the translation of large units. It additionally reduces the effectiveness of language models, as it decreases the size of the context, and the benefits in terms of translation quality can be limited, except for the translation of out-of-vocabulary words. To mitigate these potentially negative effects, several authors have proposed to simultaneously consider multiple decomposition schemes, which are then recombined using system combination techniques (e.g. Minimum Bayes Risk decoding) as in (Dyer, 2007) (for German), (de Gispert et al., 2009) (for Finnish), and (Virpioja et al., 2010) for German and Czech. This way, it is possible to get the benefits of using large units in translation, when they are found in the training data, while still being able to produce unseen forms through morphological decomposition.

Another pitfall of the "segment, then align" strategy is its linguistic bluntness: posterior to splitting, morphs of the same words behave as if they were completely unrelated, and can align to arbitrarily remote units in the other language. The model of Eyigöz

[45] This is similar to the method using successor and predecessor frequencies introduced by Harris, see Section 2.2.1.

et al. (2013) is intended to address these issues and develops a hierarchical view of alignment, which reintroduces the distinction between words and morphemes. Assuming the availability of a morphological decomposition in both source and target, their model extends IBM Model 1 (and the HMM model) so as to constraint morpheme alignments with word alignments: if a source morph aligns with a target morph, then the corresponding word forms must also align. From a technical viewpoint, this corresponds to a two-level IBM Model 1, where word alignments decompose into morpheme alignments.

Fundamental limits Although preprocessing can sometimes rely on unsupervised techniques, as in the work of Besacier et al. (2006), Virpioja et al. (2007), or Fishel and Kirik (2010) mentioned above, it nevertheless typically implies external resources and tools for morphological analysis in the source and/or the target languages. In the low-resource setting motivating the present work, these resources will not be available. Moreover, if these approaches provide practical means to improve alignments involving one (or two) synthetic languages, they remain unsatisfactory for lack of modeling the dependency of the alignment process to the granularity of aligned units. Joint models of segmentation and alignment, discussed next in Section 2.6.2, try to capture this dependency.

2.6.2 Jointly segment and align

In this section, we review attempts to develop joint models of segmentation and alignment. We start with asymmetric approaches, where the segmentation in words of one side of the bitext is known and kept fixed; we then briefly review attempts to learn the segmentation simultaneously on both sides, as well as more recent work addressing the particular challenge of segmenting noisy inputs in the context of bilingual lexicon extraction. Table 2.1 summarizes inputs and outputs for a choice of models described in this section.

2.6.2.1 Asymmetric approaches

An early joint model is the proposal of Deng and Byrne (2005), which extends the HMM model of Vogel et al. (1996) (see Section 2.5.2) into a word-to-phrase alignment model, by allowing a source word to generate multiple target words (i.e. a phrase) on the target side. The authors develop a model analogous to IBM Model 4, where the fertility of source words is explicitly controlled, while preserving the desirable properties of the HMM model (efficient decoding algorithm, exact computation of posteriors, well-understood learning procedure). In this approach, an alignment is composed of a set of random variables $\{(a_n, h_n, \phi_n), n \in [1, N]\}$, where a_n is the index of the source equivalent of the n^{th} target phrase, h_n a binary indicator for null alignment, and ϕ_n the length of the n^{th} phrase. Given these, the alignment probability of the n^{th} target phrase ν_n, $P(\nu_n|a_n, h_n, \phi_n, \mathbf{w})$, is essentially a product of the standard translation model parameters $P(\omega_{j,n} \mid w_{a_n})$ (the phrase ν_n is composed by the sequence of words $\omega_{1,n}...\omega_{\phi_n,n}$). Another variant is also considered, where a bigram model is used for $P(\nu_n|a_n, h_n, \phi_n, \mathbf{w})$. The applications studied in this paper are phrase extraction and automatic alignment of Chinese to English, but this model could be used for the segmen-

tation of a target sequence of phonemes into "phrases" (pseudo-morphemes or words), with joint alignment to a source sequence of words.

Among the early approaches using bilingual information to segment text, Xu et al. (2008) present a Bayesian model able to learn a Chinese text segmentation suitable for machine translation. It assumes that the corpus of parallel sentences ($\mathbf{u} = u_1^K, \boldsymbol{\omega} = \omega_1^J$), with $\mathbf{u}$ a sequence of Chinese characters and $\boldsymbol{\omega}$ a sequence of English words, is generated in parallel to a hidden sequence of Chinese words $\mathbf{w} = w_1^I$ and a hidden alignment $\mathbf{a} = a_1^J$. The joint probability of a sentence pair and its hidden variables then factorizes to:

$$P(\mathbf{u}, \boldsymbol{\omega}, \mathbf{w}, \mathbf{a}) = P(\mathbf{w})\delta(\mathbf{w}, \mathbf{u})P(\boldsymbol{\omega}, \mathbf{a} \,|\, \mathbf{w})\,, \tag{2.34}$$

where $\delta(\mathbf{w}, \mathbf{u}) = 1$ if $\mathbf{w}$ corresponds to the character sequence $\mathbf{u}$, and 0 otherwise. $P(\mathbf{w})$ is specified by the monolingual unigram model of (Goldwater et al., 2006a) (see Section 2.4.3), with a slightly modified spelling model incorporating a Poisson prior on word length. The translation probability $P(\boldsymbol{\omega}, \mathbf{a} \,|\, \mathbf{w})$ is specified by the IBM Model 1, modified so that a Dirichlet process prior (see Section 2.4.1) is placed on the distributions over English words depending on the Chinese word it is aligned to:[46]

$$P(\boldsymbol{\omega}, \mathbf{a} \,|\, \mathbf{w}) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} P(\omega_j \,|\, G_{w_{a_j}})\,, \tag{2.35}$$

with $G_{w_{a_j}} \sim \mathrm{DP}(\alpha, P_0(\omega))$, and $P_0(\omega)$ the empirical distribution over English words in the data. The final model involves an equivalent factor for the other direction (English to Chinese), with a subsequent weighting of both components. Inference is performed using a Gibbs sampler considering only the alignment hypotheses that are close to the current alignment. The Gibbs sampling procedure is also combined iteratively with a realignment step using the `giza++` toolkit (Och and Ney, 2003).

With a similar goal in mind, finding the best segmentation[47] of an unsegmented sequence $\boldsymbol{\pi}$ (here Chinese or Korean), in order to train an MT system with English as the source language, Chung and Gildea (2009) propose another extension of IBM Model 1. Contrary to Xu et al. (2008), no language model is made explicit, and segmentation is essentially a by-product of learning the alignment. In IBM Model 1, posterior probabilities $P(\mathbf{a} \,|\, \boldsymbol{\omega}, \mathbf{w})$ for the alignments are computed during EM training in order to learn the translation parameters of the model. As the target sequence $\boldsymbol{\pi} = \pi_1, \ldots, \pi_L$ is not segmented into $\boldsymbol{\omega}$ here, the authors introduce a binary vector $\mathbf{b}$ indicating the presence or absence of a word boundary after each character in $\boldsymbol{\pi}$, and a hidden alignment variable a between a target subsequence $\pi_s^t = \pi_s, \ldots, \pi_t$ and a source word. Posterior probabilities are computed using dynamic programming:

$$P(b_s^t = (1, 0, \ldots, 0, 1), a = i \,|\, \mathbf{w}) = \frac{\alpha(s)P(\pi_s^t \,|\, w_i)P(a = i)\beta(t)}{P(\boldsymbol{\pi} \,|\, \mathbf{w})}\,, \tag{2.36}$$

where $\alpha(s) = P(\pi_1^s, b_s = 1 \,|\, \mathbf{w})$, and $\beta(t) = P(\pi_{t+1}^L, b_t = 1 \,|\, \mathbf{w})$; expected counts for individual word pairs (π_s^t, w_i) are subsequently cumulated over the data using these

[46] To compare to Equation (2.26).

[47] In the context of machine translation, the authors rather use the term *tokenization*.

	input		output	
	source	target	source	target
asymmetrical methods				
(Deng and Byrne, 2005)	words	words	words	phrases
(Xu et al., 2008)	characters	words	words	words
(Chung and Gildea, 2009)	words	characters	words	words
(Nguyen et al., 2010)	characters	words	words	words
(Naradowsky and Toutanova, 2011)	morphs+words	words	morphs+words	morphs+words
(Stahlberg et al., 2012)	words	phonemes	words	words
(Adams et al., 2016b)	phoneme lattices	words	words	words
symmetrical methods				
(Marcu and Wong, 2002)	words	words	phrases	phrases
(Zhang et al., 2003)	words	words	phrases	phrases
(DeNero et al., 2006)	words	words	phrases	phrases
(DeNero et al., 2008)	words	words	phrases	phrases
(Snyder and Barzilay, 2008a,b)	words	words	morphs	morphs
(Neubig et al., 2011)	words	words	phrases	phrases
(Neubig et al., 2012)	characters	characters	morphs	morphs

Table 2.1: Granularity of the inputs and outputs of various joint models of segmentation and alignment. We underline the new representation of the source and/or target produced by the system in the output, and use generically the terms "morphs" to denote subword units, "words" to denote words or pseudo-words, and "phrases" to denote multi-word units.

posteriors during the E step, and normalized during the M step to update lexical probabilities $P(\pi_s^t \mid w_i)$. This approach proves to be successful for machine translation in terms of BLEU score, while interestingly confirming that the best segmentation (in terms of F-measure with respect to a gold segmentation) does not always agree with the best tokenization for MT.

The work of Nguyen et al. (2010) can be viewed as a bilingual version of (Mochihashi et al., 2009) presented in Section 2.4.4: here, the generative story for a pair of unsegmented source and segmented target sentences follows two steps: first an unsupervised segmentation using the nested Pitman-Yor language model of (Mochihashi et al., 2009)) takes place; then, after defining unaligned source words and unaligned target words, one-to-one word pairs are aligned conditionally to the source segmentation. Note however that the alignment is never generated explicitly, as word order is not modeled on the target side: rather, a bag of target words is "aligned" to a bag of source words, and no distortion model is defined; lexical translation probabilities, finally, are distributed according to a draw from a Pitman-Yor process. Inference relies on sampling segmentation points, and computing the joint posterior probability. In order to speed-up learning, likely segmentation points in the source are pre-computed with a rule-based system.

Also asymmetrical, as the segmentation of the source into morphemes is assumed to be known in a setup where source and target words are also known, the work of Naradowsky and Toutanova (2011) presents an extension of the conventional HMM model for word alignment, with the goal of identifying morphemes on the target side.

The HMM model is improved in several different ways:

1. the source is a sequence of morphemes w_1^I;[48] source states are pairs $y = [a, t]$ where a is the index of the source morpheme in $[0, I]$, and t is the emitted target morpheme type (prefix, suffix, or stem). The authors propose to use a log-linear parameterization (Berg-Kirkpatrick et al., 2010) which enables them to include rich-features (e.g. dependency and POS information) on top of the conventional distortion-based model;
2. source states emit target morphemes; the corresponding distribution includes a first order dependency over past morphemes, and an additional conditioning over word boundary indicators;
3. to account for word boundaries in the target, even if they are observed, each HMM state also emits a word boundary Bernoulli variable b_j with Markovian dependencies.

Denoting $\mathbf{ta}$ the generalized alignment vector, $\boldsymbol{\omega}$ the target morpheme sequence, and $\mathbf{b} = b_1^J$ the vector of word boundary variables ($b_j = 1$ at word endings), the overall model defines $P(\boldsymbol{\omega}, \mathbf{b}, \mathbf{ta}|\mathbf{w})$ as a product of three terms: the first accounts for the translation model and is essentially a product over morphemes of terms $P(\omega_j \,|\, ta_j, \mathbf{w})$; the second is the distortion model $P(ta_j \,|\, ta_{j-1}, \mathbf{w})$; and the third is needed to model word endings $P(b_j \,|\, b_{j-1}, \mathbf{w})$.[49]

Lastly, most relevant to our research, Stahlberg et al. (2012, 2014) consider the problem of aligning a sequence of words (source) to a sequence of phonemes or characters (target). The authors develop a variant of IBM Model 3, "Model 3P", in which an additional level modeling target words' lengths is added to the generative story of IBM Model 3, namely: choosing a *fertility* for each source word, *translating* each (possibly duplicated) source word into a target word, reordering the target words (*distortion*).[50] In Model 3P, distortion and lexical translation are performed in reverse order, and the latter is preceded by the generation of a length for (abstract) target words, corresponding to the number of phonemes in each of those word. In both models, each symbol in the target aligns with exactly one source word, but the additional modeling level, which can be viewed as a target fertility, serves as a granularity bridge between words and phonemes. Targeting the related task of bilingual lexicon extraction, Adams et al. (2015) contrast Model 3P with a "segment, then align" approach relying on the (monolingual) nested Pitman-Yor language model of Mochihashi et al. (2009) and the `giza++` toolkit (IBM Model 4), as well as with a symmetrical joint approach introduced by (Neubig et al., 2011) that we review in the next section. In these experiments on bilingual lexicon extraction, Model 3P performs only slightly better than the direct alignment of source phonemes to target words with `giza++`, and is outperformed by the two other methods.

[48] We employ the notations defined for words in Section 2.1.1.2 to represent morphemes in this context.

[49] Additional dependencies in these three terms to previous boundary variables b_j and morphemes ω_j are also considered.

[50] We ignore here that target words can also be generated by the `null` source word.

2.6.2.2 Symmetric approaches

We now turn to approaches aimed at simultaneously segmenting the source and the target side. An early line of work following this path can be identified in attempts to directly extract bilingual phrases for phrase-based SMT, instead of relying on the usual heuristic pipeline (Koehn, 2010).

A first model is introduced in (Marcu and Wong, 2002), where a pair $(\mathbf{w}, \boldsymbol{\omega})$ of sentences is generated in two steps: i) generate K hidden concepts, each generating in turn a phrase-pair (v, ν), and ii) reorder the phrases on each side to recover $(\mathbf{w}, \boldsymbol{\omega})$. Note that this, as well as many other approaches along these lines, prevents to extract discontinuous phrases. Two models of increasing complexity are considered, depending on how they model the reordering component; in both cases estimation is computationally challenging and requires both to heavily filter the repertoire of possible phrase pairs, and to develop approximate estimation techniques. A conditional version of this model, analogous to an IBM Model 3 operating on phrases rather than words, is in (DeNero et al., 2006): learning its parameter with EM is however intractable, due to the need to sum over all segmentations and alignments, and leads to an approximation using constraints derived from word alignments. DeNero et al. (2008) note that these approaches are plagued by a clear tendency to undersegment the corpus, and propose to introduce priors in order to constraint the model in a principled way. The authors develop an algorithm based on Gibbs sampling to compute count expectations over the posterior distribution of latent segmentation and alignment variables, while taking prior information into account; sampling considers various simple operators to move from one assignment of these variables into another. Upon convergence, the expectations can be plugged into the M step of the EM algorithm to derive phrase translation probabilities.

Similar techniques lie at the core of the work reported in (Snyder and Barzilay, 2008a,b), which however considers the segmentation of words in character substrings, rather than of sentences into phrases. In this work, parallel sequences of words (phrases) are obtained by sampling from two monolingual distribution of morphemes, plus one bilingual distribution over abstract morphemes. This model is based on the Dirichlet process, and can also integrate prior information regarding abstract morphemes: for instance using string similarity when the two languages are orthographically or phonetically related.

A series of papers (Zhang et al., 2003; Zhang and Vogel, 2005a; Vogel, 2005; Zhang and Vogel, 2005b) develop an alternative approach to phrase alignments, which uses phrase-to-phrase association scores, such as the point-wise mutual information or aggregates derived from IBM Model 1 scores, instead of a sound probabilistic model. With the exception of (Zhang et al., 2003), where a phrasal alignment is actually built, these scores are mostly used to perform phrase extraction for machine translation. These attempts have been continued in (Xu et al., 2006), and in (Lardilleux et al., 2012; Gong et al., 2013). Note that they start with association scores attached to minimal units, a requirement that is difficult to meet with unsegmented character (or phonemic) strings.

Phrase-to-phrase alignments are also studied in Neubig et al. (2011), where the authors use Bayesian nonparametric techniques on top of inversion transduction grammars (ITG) alignments (Wu, 1997) to extract many-to-many phrasal alignments with *varying levels of granularity*, thereby fixing a well-known issue with phrasal-alignment models,

which typically lack the ability that heuristic approaches have to extract small units embedded within larger units. This work continues in Neubig et al. (2012), where the focus shifts from the alignment of multi-word units (or phrases) to subword units: the resulting alignment of variable length character strings is then used to train a character-based translation model, thereby mitigating the data sparsity issues faced with systems operating at the level of words.

2.6.2.3 Noisy source inputs

Many joint models are approached with machine translation in mind. They also assume clean text transcripts on both side of the bitext. In a language documentation scenario, and particularly if the documented language is unwritten, speech will be recorded and automatically transcribed using ASR, which will lead to imperfect (*noisy*) source inputs. This is likely to harm statistical inference and training; a drastic decrease of performance has been previously documented, indeed, by Jansen et al. (2013) for unsupervised word segmentation using Bayesian models (Goldwater's unigram and bigram models – see Section 2.4.3 – as well as an Adaptor Grammar using collocations – see Section 2.4.5).

Concluding this survey, some recent efforts embracing the language documentation scenario and attempting to make use of ASR noisy outputs should be highlighted. Building on the formalization of the nested Pitman-Yor language model of Mochihashi et al. (2009) as a composition of weighted finite state transducers (WFSTs) proposed by Neubig et al. (2010) (see Section 2.4.4), Adams et al. (2016a) first devise a way to learn a translation model from word lattices. Their model is extended in (Adams et al., 2016b) to additionally learn a lexicon directly from phoneme (instead of word) lattices. Three WFSTs are composed: the first corresponds to the acoustic model (the actual phoneme lattices); the second represents a lexicon transducing phoneme substrings into (pseudo) words; and the third accounting for the lexical translation model. The lexicon and the translation model are successfully evaluated by their ability to decrease phoneme error rate (PER).

2.7 Conclusion and open questions

In this chapter, we have defined two important tasks motivated by a language documentation scenario – word segmentation and word alignment – and the standard metrics to evaluate automatic processing performing those tasks. We subsequently reviewed various works related to these two tasks, in a monolingual and in a bilingual setting. To conclude, we summarize now the lessons we learnt, and list a choice of open questions that will motivate our own work.

What did we learn? Regarding word segmentation, a general evolution during the last two decades exhibits the progressive abandon of methods based on local statistics and heuristic search, in favor of more principled strategies, mostly based on the minimum description length principle or various types of nonparametric Bayesian modeling. In the latter approach, this allows to take into account prior linguistic knowledge, such as the (Zipfian) nature of word distributions, the dependency of words to their context, or other hierarchical processes at work in natural languages. The Adaptor Grammar

framework represents a culminating achievement in that respect. The increased expressivity of the formalisms used in word segmentation has led to substantial performance improvement and this can be leveraged for our language documentation goal.

Automatic word alignment has been an extremely active research field since the early 1990s, especially because it constitutes the cornerstone of phrase-based statistical machine translation, a technique used by the strongest translation systems until the recent advent of neural machine translation. Most research on alignment operates on larger corpora than is available in a low-resource scenario, and is evaluated in terms of translation performance rather than linguistic soundness of the alignments produced. We also observed that alignment models, to be effective, rely on the assumption that aligned units share a similar granularity, and that a one-to-one mapping is an achievable goal. For many language pairs, this assumption is violated, thus requiring the modeling of a segmentation process in addition to the alignment procedure. The shift to a neural paradigm has diverted most efforts in the machine translation community towards other directions of research, although some concepts related to automatic word alignment are still relevant in NMT as we will see in Chapter 6.

Word segmentation and alignment are mutually dependent: the translation of an unsegmented sentence can provide hints to better identify its segmentation (this is indeed the approach taken by field linguists), while alignment requires comparable units, which might necessitate the segmentation of the source or target side of the bitext. We showed that the most promising line of research is the one trying to jointly model both processes, in order to capture this dependency. If the inspiration for the design of such models has often also been machine translation, more recent work is addressing this challenge with language documentation in mind, and is evaluated in a low-resource scenario. We also observed in our survey that, while the body of work on unsupervised word segmentation is vast – and so is the body of work on automatic alignment – much less work has explored joint modeling. In fact, benefits are not as obvious as those observed with the introduction of Bayesian techniques in word segmentation for instance, and these models lead to much more involved inference and training schemes; more work needs to be done to demonstrate their applicability for language documentation. Another insight gained from our study is that bilingual approaches to word segmentation could be most relevant when working with noisy inputs (e.g. automatic transcripts from speech), which are known to harm drastically the robustness of monolingual Bayesian methods.

Open questions Many challenges posed by computational language documentation are left unaddressed. In this final section, we list various open questions that motivate our own work.

- Most techniques in word segmentation have been applied to Indo-European languages with a fusional morphology, and much less on languages with an agglutinative morphology. How will these techniques perform in a realistic low-resource setup on Bantu languages? And how much data do we need to obtain reliable results? (Chapter 3)

- How can we leverage expert knowledge when it is available? And what kind of information is useful to the linguist for the purpose of documenting a new language? (Chapter 4)
- It has been shown that prosodic cues can help word segmentation (Ludusan et al., 2015), but for tonal languages, how can tone be modeled and used to inform segmentation? (Chapter 5)
- The effectiveness of alignment techniques on very small corpora is seldom documented. Can we really take advantage from our bilingual data in order to better segment a language to be documented? (Chapter 6)
- Can we still obtain reliable results with noisy transcripts obtained from speech recordings? (We address this question in (Godard et al., 2018c) and (Ondel et al., 2018), but this work is not reported in this thesis.)

Chapter 3

Preliminary Word Segmentation Experiments

Contents

In Chapter 1, we introduced an approach to collect data for endangered and unwritten languages, and in Chapter 2, we studied various unsupervised learning techniques for word segmentation and alignment which can be used to process these data. In this chapter, we describe two realistic corpora for endangered language documentation that we use throughout this thesis, and experiment with several word segmentation systems, in a monolingual or a bilingual setting. Segmentation results are contrasted with those obtained on English, using an additional corpus. We also explore the impact of the data size and representation, and establish strong baselines for the word segmentation task, while showing the difficulty to take advantage of the bilingual supervision. An early version of this work, with slightly different choices for the experimental setup, has appeared in (Godard et al., 2016). Additionally, the Mboshi corpus has been publicly released and documented in (Godard et al., 2018a).

3.1 Introduction

As we saw in Chapter 2, many models and systems have been proposed to handle word segmentation; mostly in a monolingual setting, but also using the supervision provided by a translation of the unwritten language (UL) utterances into a well-resourced language (WL). In our language documentation scenario (Chapter 1), speech in the UL is collected and translated in the WL. While it is possible to process speech frames directly in an end-to-end NMT translation system, this proves to be quite hard with corpora of a few thousand sentences (Duong et al., 2016; Bansal et al., 2017, 2018a; Scharenborg et al., 2018a); in (Duong et al., 2016) for instance, such an NMT model is outperformed by Moses (Koehn et al., 2007) on a translation task.

3.1.1 A favorable scenario

Following the BULB project's methodology, we assume a phonemic transcription of the recorded speech. In a fully realistic scenario, this transcription would be automatically produced using ASR techniques – more specifically unsupervised acoustic units discovery (AUD) techniques – which would likely lead to highly noisy data. We have explored this scenario in (Godard et al., 2018a; Ondel et al., 2018; Godard et al., 2018d) but we restrict the work presented in this thesis to a more favorable situation: instead of noisy phonemic transcripts, we use graphemic transcriptions of the "unwritten" languages. Indeed, while Mboshi and Myene, the UL languages considered in our experiments, can be considered as rarely written, linguists have nonetheless defined for each language a non-standard graphemic form considered to be close to the language phonology.

If this departs from our goal to support realistic language documentation, it seems a necessary step to establish a better understanding of the impact of various factors on word segmentation: language modeling assumptions, size of the available data, variation across languages, variation in the representation of the UL (e.g. encoding tonal information or not), and use of a WL translation supervision (also varying granularity, e.g. words vs. lemma), etc. Most of these effects would likely be obfuscated by the presence of noisy inputs.

3.1.2 Challenges for low-resource languages

Besides, many challenges specific to the low-resource scenario are somewhat orthogonal to the question of the quality of the AUD we are able to discover. First, a vast majority of the word segmentation techniques presented in Chapter 2 are applied to Indo-European languages with a fusional morphology, and much less to languages with an agglutinative morphology. Additionally, when a low-resource scenario is envisioned, it is often *simulated* using smaller quantities of data from a WL, which is prone to bias results, as realistic language documentation corpora are likely to differ, in content and statistics, from popular WL corpora.[1] Lastly, and as was discussed in Section 2.6, when word segmentation is performed with machine translation in mind, the goal can fun-

[1] For instance the Fisher and CALLHOME Spanish–English Speech Translation Corpus (Post et al., 2013).

damentally diverge from what is pursued in language documentation and preservation, where the linguistic soundness of the units discovered is of prime importance.

In this chapter we seek to better understand how several unsupervised word segmentation techniques perform on real UL corpora, how the size of those corpora matter in segmentation performance, and the practicality of using the bilingual part of the data at our disposal. We also question the representation of the target[2] UL data (including tonal information or not), and experiment with various levels of granularity of the data on the source[2] WL side (words, lemmas, morphs, and POS). We first provide a description of the three corpora used in our experiments (Section 3.2). We then detail our experimental setup and discuss our results (Section 3.3), before summing up key findings (Section 3.4) to provide ground for ensuing chapters of this thesis.

3.2 Three corpora

As we just mentioned, a lot of research on word segmentation for the low-resource scenario has been emulating this situation using truncated corpora from various WL. One key aspect of the work presented in this thesis is to conduct experiments on real endangered languages, following a plausible language documentation scenario. We consider a bilingual French-Mboshi corpus, as well as a monolingual Myene corpus, that we complement with a French-English corpus derived from the TedTalk corpus. We first provide some elements of linguistic description for Mboshi and Myene, two Northwestern Bantu Languages, and proceed to describe qualitatively and quantitatively the three corpora.

3.2.1 Elements of linguistic description for Mboshi and Myene

Mboshi (Bantu C25[3]) is a language spoken in Congo-Brazzaville, and Myene (Bantu B10) corresponds to a cluster of six mutually intelligible varieties (Adyumba, Enenga, Galwa, Mpongwe, Nkomi and Orungu) spoken at the coastal areas and around the town of Lambarene in Gabon.[4] Unlike southern Bantu relatives such as Swahili, Sotho or Zulu, Mboshi and Myene are scarcely studied, protected, and resourced. In this section, we briefly describe the main aspects related to the phonetics, phonology, morphology, and tonology of these two languages.

Phonetics and phonology Mboshi and Myene both have a 7 vowel system (i, e, ɛ, a, ɔ, o, u). Mboshi has an opposition between short and long vowels, which does not exist in Myene. Mboshi consonantal system includes the following 25 phonemes: p, t, k, b, d, β, l, r, m, n, ɲ, mb, nd, ndz, ng, mbv, f, s, ɣ, pf, bv, ts, dz, w, j. It has a set of prenasalized consonants (mb, nd, ndz, ng, mbv) which is common in Bantu languages (Embanga Aborobongui, 2013; Kouarata, 2014). Myene includes the following phonemes: p, t, k, b, d, β, l, r, m, n, f, s, g, y, v, ŋ, w, z – many of them

[2] As we explain in Section 2.1.1.2 (Chapter 2), we choose to identify the UL to the *target*, and the WL to the *source*.

[3] In Malcolm Guthrie's classification of the Bantu languages (Guthrie, 1948, 1967).

[4] Our Myene data correspond to the Orungu variant.

with variants of realization. Prenasalized consonants also exist in Myene (Ambouroue, 2007).

Both languages are rarely written, but linguists have defined a non-standard graphemic form to transcribed them, which is close to the language phonology. Affricates and prenasalized plosives are coded using multiple symbols (e.g. two symbols for dz, three for mbv). For Mboshi, long and short vowels are coded respectively as V and as VV. In Myene, the transcription of the corpus not only uses the phoneme set, but also the main variants (ɲ, tʃ, dz) and some marginal sounds found in loanwords.

Both languages display a complex set of phonological rules. The deletion of a vowel before another vowel in particular, common in Bantu languages, occurs at 40% of word junctions in Mboshi (Rialland et al., 2015). This tends to obscure word segmentation and introduces an additional challenge for automatic processing. Note, however, that deleted vowels have been reintroduced by the annotators in our transcriptions, alleviating this particular difficulty in our experiments.

Morphology Both Mboshi and Myene display a mostly agglutinative morphology. Words are composed of roots and affixes, and almost always include at least one prefix, while the presence of several prefixes and one suffix is also very common. The suffix structure mostly consists of a single vowel V (e.g. -a or -i) whereas the prefix structure may be both CV or V (or CVV in Mboshi). The most common syllable structures are V and CV in both languages. CVC also occurs in Myene, and CVV in Mboshi.[5]

There is a long tradition in describing Bantu languages with the help of a rich set of nominal class prefixes (Bleek, 1851). Whereas Bleek's classification proposes 18 classes, the number of classes varies across languages and even within a language depending on the authors. Most recent work on Mboshi, for instance, describes a system using 13-14 classes (Bedrosian, 1996; Embanga Aborobongui, 2013).

For both languages, the structure of the verbs, also common in Bantu languages, is as follows: Subject Marker — Tense/Mood Marker — Root-derivative Extensions — Final Vowel. A verb can be very short or quite long, depending on the markers involved.

Tonology Both Mboshi and Myene are tonal languages. In Mboshi, the prosodic system involves two tones and an intonational organization without downdrift[6] (Rialland and Aborobongui, 2016). The high tone is coded using an acute accent on the vowel while a low tone vowel has no special marker. Word root, prefix and suffix all bear specific tones, which tend to be realized as such in their surface forms.[7] Tonal modifications may also arise from vowel deletion at word boundaries.

According to Ambouroue (2007), Myene has a more complex tonal system: high tone, low tone, descending tone, and two more tones characterizing intermediate levels between high and low tones. Unfortunately, the transcribed data at our disposal in Myene encode tones in an heterogeneous manner, and we decided to strip tonal information in Myene for our experiments.

[5] CCV and CCCV may also arise due to the presence of affricates and prenasalized plosives mentioned in this section.

[6] A prosodic phenomenon, caused by interactions between tones, where pitch progressively decreases during an utterance.

[7] The distinction between high and low tones is phonological; see (Rialland and Aborobongui, 2016).

Mboshi	wáá ngá iwé léekundá ngá sá oyoá lendúma saa m ótéma
French	si je meurs enterrez-moi dans la forêt oyoa avec une guitare sur la poitrine

Figure 3.1: A tokenized and lowercased sentence pair in the French-Mboshi corpus.

A productive combination of tonal contours in words can also take place due to the preceding and appended affixes. These tone combinations play an important grammatical role, particularly in the differentiation of tenses. However, in Mboshi, the tones of the roots are not modified due to conjugations, unlike in many other Bantu languages.

All these characteristics – phonological, morphological, tonal – describe the way words are formed, influence their average length, and are therefore important for the word segmentation task.

3.2.2 Data and representations

Origin and size Corpora for Mboshi and Myene were collected following the language documentation scenario described in Chapter 1 (Section 1.2.2), using a mobile app dedicated to fieldwork language documentation (Blachon et al., 2016). They both comprise of about 5K sentences (a little less for Myene), transcribed by linguists (see Section 3.1.1) and annotated with reference word segmentations.

At the time of writing, though, the French translations of the Myene corpus have yet to be consolidated, and the Myene corpus is only usable in a monolingual setting. The Mboshi corpus, however, benefits from French translations aligned at the sentence level, and can be used in a bilingual setting. Data processing and cleaning performed before releasing this corpus publicly have been described in (Godard et al., 2018a). It is worth mentioning that, since its release in late 2017-early 2018, this dataset[8] has been used in several studies involving a low-resource scenario (Anastasopoulos and Chiang, 2018a; Scharenborg et al., 2018b; Anastasopoulos and Chiang, 2018b; Bansal et al., 2018b).

Our French-English corpus, an extract of 100K sentences from the TedTalk corpus,[9] provides an additional language contrast, with English treated as a low-resource language. It also allows us to experiment with more data in one of our experiments. Standard pre-processing steps have been performed: tokenization, lowercasing, filtering of sentences longer than 80 words, removal of the punctuation.

Representation and granularity For Mboshi, we consider two representations: one including diacritics denoting tones as described in Section 3.2.1 (`tone`), and the other where diacritics have been removed (`notone`).[10] An example sentence pair from the French-Mboshi corpus with tonal information is displayed in Figure 3.1.

For the two bilingual corpora (French-Mboshi and French-English), we additionally vary the granularity of the units on the French side. We denote the granularity

[8] Available at http://www.islrn.org/resources/747-055-093-447-8/.

[9] https://wit3.fbk.eu/

[10] In the absence of tonal markers for Myene and English, their representations also correspond to `notone`.

provided by the tokenization of French as `word`. Representation `lemma` is the result of a lemmatization performed with the TreeTagger (Schmid, 1994), and representation `morph` corresponds to a morphological segmentation of words into morphs, obtained with the `polyglot` toolkit[11] which provides a Morfessor model (Smit et al., 2014) trained on French. We also experiment with representation `pos` where words have been replaced by their part-of-speech using `wapiti` (Lavergne et al., 2010).

We give elementary statistics for the French-Mboshi corpus in Table 3.1 (only for representation `notone`), for the Myene corpus in Table 3.2, and for the French-English corpus in Table 3.3 (excluding granularity `pos`).

Language	Size id	Granularity	#sent	#tokens	#types	Avg. sent length	Avg. token length
French	0.5K	`morph`	500	6,109	1,085	12.22	2.83
French	0.5K	`lemma`	500	4,131	995	8.26	4.38
French	0.5K	`word`	500	4,131	1,205	8.26	4.18
Mboshi	0.5K	`word`	500	2,902	1,054	5.80	4.29
French	1K	`morph`	1,000	12,707	1,549	12.71	2.85
French	1K	`lemma`	1,000	8,613	1,609	8.61	4.39
French	1K	`word`	1,000	8,613	2,062	8.61	4.21
Mboshi	1K	`word`	1,000	6,171	1,817	6.17	4.17
French	2K	`morph`	2,000	26,174	2,091	13.09	2.86
French	2K	`lemma`	2,000	17,832	2,453	8.92	4.39
French	2K	`word`	2,000	17,832	3,285	8.92	4.20
Mboshi	2K	`word`	2,000	12,576	3,134	6.29	4.22
French	5K	`morph`	5,130	61,276	2,738	11.94	2.86
French	5K	`lemma`	5,130	42,150	3,680	8.22	4.35
French	5K	`word`	5,130	42,150	5,177	8.22	4.15
Mboshi	5K	`word`	5,130	30,556	5,312	5.96	4.19

Table 3.1: Elementary statistics for the French-Mboshi corpus. The average sentence length is an average number of tokens per sentence, and the average token length is an average number of characters per token.

3.3 Experiments and discussion

In this section, we describe the methodology we chose to conduct word segmentation experiments on the three corpora described in Section 3.2, and discuss the results we obtain. Structuring our exploration around three axis of variability, we experiment:

- with 5 systems corresponding to specific language modeling hypotheses; one of these systems makes use of the bilingual data, while the four others operate only on monolingual data (details are in Section 3.3.1);

[11] `http://polyglot-nlp.com`.

Language	Size id	Granularity	#sent	#tokens	#types	Avg. sent length	Avg. token length
Myene	0.5K	word	500	1,912	898	3.82	4.74
Myene	1K	word	1,000	3,899	1,536	3.90	4.70
Myene	2K	word	2,000	7,893	2,489	3.95	4.71
Myene	5K	word	4,579	18,047	4,190	3.94	4.72

Table 3.2: Elementary statistics for the Myene corpus. The average sentence length is an average number of tokens per sentence, and the average token length is an average number of characters per token.

Language	Size id	Granularity	#sent	#tokens	#types	Avg. sent length	Avg. token length
French	0.5K	morph	500	10,018	1,442	20.04	3.06
French	0.5K	lemma	500	6,820	1,318	13.64	4.54
French	0.5K	word	500	6,820	1,733	13.64	4.50
English	0.5K	word	500	6,122	1,522	12.24	4.36
French	1K	morph	1,000	19,570	1,999	19.57	3.02
French	1K	lemma	1,000	13,456	1,964	13.46	4.47
French	1K	word	1,000	13,456	2,705	13.46	4.41
English	1K	word	1,000	12,123	2,297	12.12	4.29
French	2K	morph	2,000	38,641	2,786	19.32	3.01
French	2K	lemma	2,000	26,251	3,110	13.13	4.48
French	2K	word	2,000	26,251	4,414	13.13	4.43
English	2K	word	2,000	23,718	3,678	11.86	4.28
French	5K	morph	5,000	93,935	3,973	18.79	3.02
French	5K	lemma	5,000	64,202	5,227	12.84	4.49
French	5K	word	5,000	64,202	7,930	12.84	4.43
English	5K	word	5,000	57,711	6,537	11.54	4.29
French	10K	morph	10,000	190,515	4,919	19.05	3.02
French	10K	lemma	10,000	129,959	7,721	13.00	4.50
French	10K	word	10,000	129,959	12,269	13.00	4.44
English	10K	word	10,000	116,521	9,973	11.65	4.29
French	50K	morph	50,000	957,363	6,589	19.15	3.02
French	50K	lemma	50,000	654,839	17,049	13.10	4.49
French	50K	word	50,000	654,839	28,938	13.10	4.42
English	50K	word	50,000	588,145	23,622	11.76	4.28
French	100K	morph	100,000	1,917,715	7,043	19.18	3.01
French	100K	lemma	100,000	1,311,566	23,687	13.12	4.48
French	100K	word	100,000	1,311,566	40,530	13.12	4.42
English	100K	word	100,000	1,177,654	33,083	11.78	4.28

Table 3.3: Elementary statistics for the French-English corpus. The average sentence length is an average number of tokens per sentence, and the average token length is an average number of characters per token.

- with increasing data sizes (0.5K, 1K, 2K, and approximately 5K sentences);[12]
- with various representations when applicable, i.e. with representations `notone` and `tone` in Mboshi, and with representations `word`, `lemma`, `morph`, and `pos` in French (for the bilingual method).

We systematically perform two runs for each experiment in order to assess the variability of the results.

3.3.1 Models and parameters

Choosing parameters One key difficulty in comparing methods and systems lies in the choices made for parameters and hyperparameters. In an early version of the work presented in this chapter (Godard et al., 2016), we decided to optimize parameters and hyperparameters on the smallest (0.5K sentences) extract of the French-English corpus using `hyperopt` (Bergstra et al., 2013). The random search algorithm was run several hundreds of times for each method, and the optimal parameters were then frozen to carry out the experiments on the French-Mboshi parallel corpus (and the French-English corpora of larger sizes). The rationale was to avoid tuning parameters on the data we were most interested in studying and documenting, namely the French-Mboshi corpus, while providing a reasonable effort to make systems comparable.

There are important drawbacks to this approach, however. First, tuning the models on English data undercuts the possibility to assess any linguistic contrast in terms of segmentation performance later on. Second, choosing the smallest extract of the French-English corpus is likely to be detrimental to larger extract, as optimal parameters and hyperparameters might differ for these larger corpora, which could introduce another bias to our analysis. To address the first problem, we could imagine holding out a portion of the data in each language for tuning, but this would not address the second issue. Additionally, this would posit the availability of reference segmentations for each new corpus under study, an impractical assumption in a language documentation scenario.

Here, we choose to give priority to reproducibility and practicality for language documentation, and rather opt to set parameters to "reasonable" values for each method. Our main goal, indeed, is to understand which system, in effect, is likely to provide the best results to a linguist attempting to document a new language.

Algorithms We now describe the word segmentation systems used in our experiments. Unspecified parameters are kept to their default settings.

- `dpseg`[13]: this system is Goldwater's implementation of the Dirichlet process-based language models introduced by Goldwater et al. (2006a, 2009), and described in Section 2.4.3 of this thesis. We choose a bigram model, and perform 20,000 Gibbs sampling iterations, with the unigram concentration parameter set to 3000, the bigram concentration parameter set to 300, and the probability to generate a word boundary set to 0.2.

[12] And additional sizes 10K, 50K, and 100K for the bilingual method `pisa` on French-English data.
[13] `http://homepages.inf.ed.ac.uk/sgwater/resources.html`.

- `pgibbs`[14]: an extension of `dpseg` in which the Dirichlet process can optionally be replaced by the more general Pitman-Yor process (PYP) (Neubig, 2014); this implementation notably provides an effective parallelization of the sampling process through blocked sampling. We experiment with a 3-gram model, and perform 2,000 blocked Gibbs sampling iterations[15] with a block size of 40, using distributions sampled from PYPs. The average length parameter (corresponding to a Poisson prior) is set to 5, and the maximum word length to 15. Sampling of the PYP hyperparameters is enabled but parallelization is not used.
- `latticelm`[16]: an implementation of the model proposed by Mochihashi et al. (2009), described in Section 2.4.4 of this thesis, which replaces the base distribution of the PYP language model found in `dpseg` or `pgibbs` (a unigram model of characters with a uniform distribution over characters) by another hierarchical PYP language model at the character level (spelling model). This system can also take (phoneme) lattices as an input,[17] implementing the extension of (Mochihashi et al., 2009) described in (Neubig et al., 2010) and (Heymann et al., 2014). We set both the language model and the spelling model to a 3-gram dependency order, and perform 2,000 Gibbs sampling iterations.
- `pypshmm`: another generalization of `dpseg`, which relies on a nested PYP, similarly to `latticelm`. This model, described in (Löser and Allauzen, 2016), constitutes the basis of an extension we introduce in Chapter 5. We use a 1-gram dependency for the language model, and a 3-gram dependency for the spelling model. The maximum word length is set to 15 and the Gibbs sampler is run for 24 hours.
- `pisa`[18]: this bilingual method corresponds to Model 3P, a model introduced by Stahlberg et al. (2012) which generalizes IBM Model 3 to the case where the target side is an unsegmented character stream (see Section 2.6.2.1). We use the implementation of the authors with 10 iterations for the EM algorithm. This system requires `giza++` to be run in order to initialize the parameters of Model 3P. The first run computed corresponds to default `giza++` parameters, while the second run sets `giza++` parameters to values recommended in (Stahlberg et al., 2012).[19]

Figure 3.2 presents boundary metrics (BP, BR, BF), token metrics (WP, WR, WF), and type metrics (LP, LR, LF), as defined in Section 2.1.2, as well as the average length of the tokens in the segmented output, for the 4 monolingual methods; two runs are plotted for target representation `notone` and all corpora sizes. Figure 3.3 presents the corresponding results for the bilingual method, `pisa`. The second run has been obtained with different parameters for `giza++` in the initialization of Model3P; additionally, various granularities (`word`, `lemma`, `morph`, and `pos`) are considered for the

[14] https://github.com/neubig/pgibbs.
[15] A smaller number of iterations, compared to `dpseg`, is necessary to achieve convergence.
[16] https://github.com/fgnt/LatticeWordSegmentation.
[17] A feature we do not use here.
[18] https://code.google.com/archive/p/pisa/.
[19] Precisely, `deficientdistortionforemptyword` is set to 1, and `emprobforempty` is set to 0.1. `maxfertility`, however, is set to 20 instead of 12.

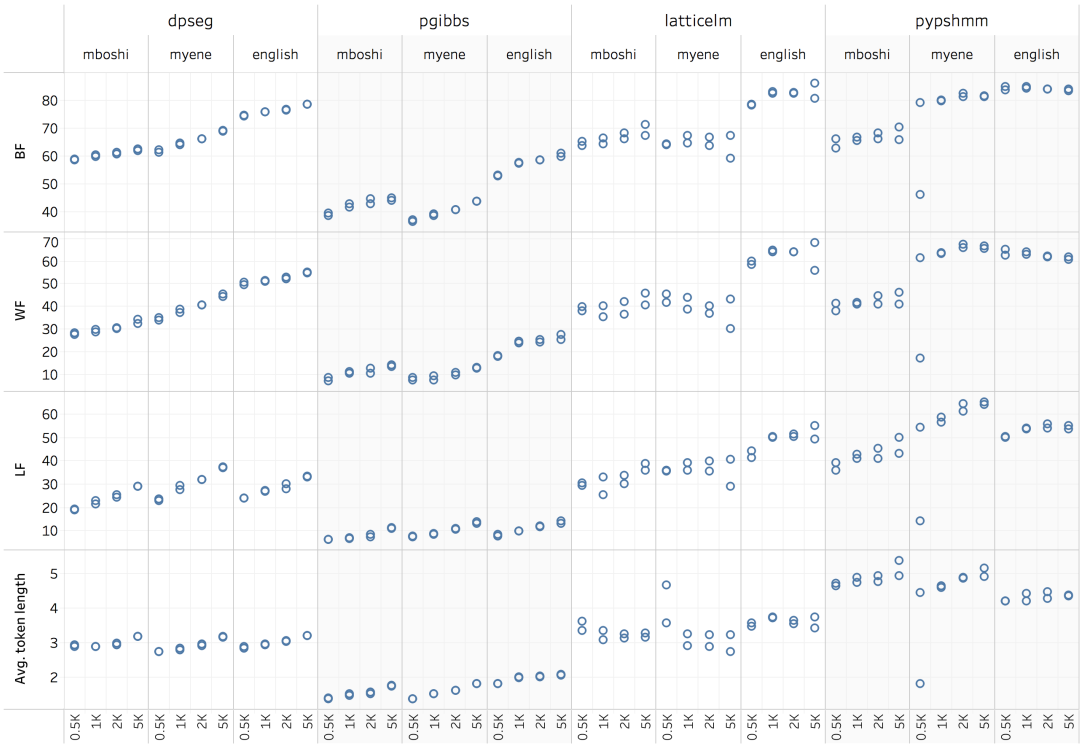


Figure 3.2: Boundary, token, and type F-measure (BF, WF, and LF), and average token length for all monolingual methods on `notone` representation, with configurations described in Section 3.3.1.

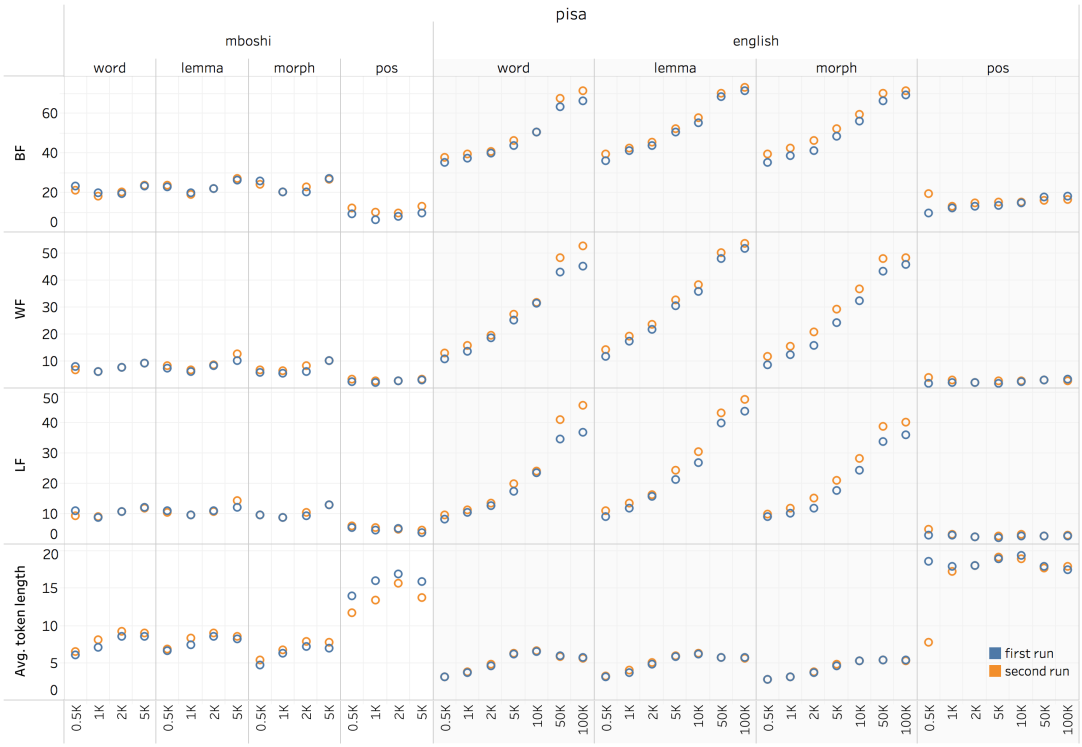


Figure 3.3: Boundary, token, and type F-measure (BF, WF, and LF), and average token length for the bilingual method (`pisa`) on `notone` representation (target), with configuration described in Section 3.3.1.

source (French) in this figure, as well as additional sizes (10K, 50K, 100K) for the French-English data. In order to make results comparable for all sizes and representations, we only compute the metrics on the first 500 sentences of each corpus after stripping tone markers (diacritics) when they are present.[20]

To give the reader an intuition of the way these metrics translate qualitatively into segmentations, Table 3.4 shows the segmentation results obtained by a run of `dpseg` on the French-English 5K corpus. Figure 3.4 displays the corresponding first 10 segmented sentences.

BP	BR	BF	WP	WR	WF	LP	LR	LF	X
75.52	83.92	79.50	51.54	56.77	54.03	47.55	22.01	30.10	5.06

Table 3.4: Word segmentation performance evaluated with boundary metrics (BP, BR, BF), token metrics (WP, WR, WF), type metrics (LP, LR, LF), and sentence exact-match (X) for a run of `dpseg` on the French-English 5K corpus. The corresponding first 10 segmented sentences are shown Figure 3.4. (Results are for the whole corpus.)

1	it canbe avery compl icated thing the ocean
2	and it canbe avery compl icated thing what human health is
3	andI'm goingto start with this one if mo m ma ain't happy ain't nobody happy
4	weknow that right we've experi ence d that
5	that's the the me ofmy talk
6	and we're making the ocean pretty un happy in alotof different ways
7	thisisa shot of Can neryRow in one thousand nine hundred and thirty two
8	Can neryRow atthe time had the bigge st industr ial can n ing oper ation onthe we st co ast
9	we pile de norm ous amount sof pollu tion into the air and into the water
10	they say youknow what you smell

Figure 3.4: A segmentation example for the first 10 sentences of the French-English 5K corpus after a run of `dpseg`.

3.3.2 Discussion

In this section, we analyze our results and discuss the contrasts in performance that we observe between methods, languages, sizes, representations and granularities.

First observations In terms of boundary F-measure, the results achieved by `dpseg`, `latticelm`, and `pypshmm` lie in the same ballpark, with slightly higher results for `latticelm` and `pypshmm` (Figure 3.2, top). On Myene, `pypshmm` performs substantially better. As the three methods involve different Markov dependency orders for the language model (henceforth LM), 1-gram for `pypshmm`, 2-gram for `dpseg`, and 3-gram for `latticelm`, the common observation (Goldwater et al., 2009; Mochihashi et al., 2009; Johnson and Goldwater, 2009, *inter alia*) that higher dependency orders are beneficial

[20] Removing tone markers only affects type metrics.

for word segmentation is not confirmed by these results. More surprisingly in the case of `pgibbs` (with a 3-gram LM), this even leads to much lower results, about 20 points below `dpseg` and its 2-gram LM. Token and type metrics (WF and LF in Figure 3.2) confirm the poor performance of `pgibbs`, while showing a clearer benefit in using `latticelm` and `pypshmm`. These two methods introduce more structured models for the lexicon, with their 3-gram dependency for the spelling model (henceforth SM), which serves as a base distribution for the hierarchical LM. Our results indicate therefore a benefit to introduce a context dependency at the character level, but fail to exhibit the same at the word level. Overall, however, the two runs performed indicate a greater stability for `dpseg` and `pgibbs`, while `latticelm` and `pypshmm` exhibit higher variances, especially on 0.5K corpora.

The bilingual method `pisa`, benefitting from the weak supervision provided by the French translation, produces disappointing results, as shown in Figure 3.3. Boundary F-measures are consistently 40 points lower than those obtained with `dpseg`. Token and type F-measures are also much lower. Varying units granularities on the French side does not result in significant performance change (columns `word`, `lemma`, and `morph` in Figure 3.3), with the exception of `pos` granularity leading to a degradation, particularly steep for the French-English corpus, for all metrics; conversely, the average token length becomes much higher than the true average token length (4.28 in the 100K French-English corpus, see Table 3.3). The second run represented in Figure 3.3, with `giza++` parameters set to those recommended in (Stahlberg et al., 2012), improves segmentation results, especially on the larger extracts of the French-English corpus, and when French granularity is kept to `word`. However, even on the 100K extract, the token F-measure (52.68% in the second run) is still lower than the token F-measure obtained on the 5K extract with `dpseg` (55.09%, averaging values of the two runs in Figure 3.2). If the poor results of the method are not that surprising for small corpora of up to 5K sentences, since the initialization with `giza++` is likely to produce bad alignments with so little data, they are less expected on the 100K extract of the French-English corpus. In (Stahlberg et al., 2012), the authors compare successfully Model3P to Adaptor Grammars, a monolingual Bayesian framework (see Section 2.4.5) that we will study in Chapter 4. The most likely explanation is that the 123K sentences extract of the English-Spanish Basic Travel Expression Corpus (BTEC) used in that work is not comparable to our 100K sentences TedTalk extract. BTEC is indeed a highly redundant dataset with simple and short utterances.[21]

More analysis To gain more understanding of these results, we look at the contrast between precision and recall for the boundary metric in Figure 3.5. For all monolingual methods but `pypshmm`, the precision is lower than the recall, indicating a tendency to oversegment the data. This was indeed already visible in Figure 3.2, where the average token lengths produced by these methods are below 4 characters, while the true average token length is above 4 characters for all corpora (see `word` representation

[21] Detailed statistics of the particular BTEC dataset used in (Stahlberg et al., 2012) do not appear in the article, but the authors indicate a 12K vocabulary size for English, to compare (in Table 3.3) to a 33K vocabulary size for English in our 100K corpus. Besides, on the French-English BTEC at our disposal, we find an average English sentence length of 7.66 tokens (to compare to more than 11 tokens per sentence in English for all corpora in Table 3.3).

in Table 3.1, 3.2, and 3.3). This effect is particularly severe with `pgibbs`: the average length of found tokens stands below 2 characters (Figure 3.2), and boundary precision (BF in Figure 3.5) is the lowest of all methods. `pypshmm`, on the other end, produces segmentations with higher boundary precision than recall, which reflects a tendency to undersegment; recall is especially low in Mboshi with this method, with an average length for found tokens close to 5 characters, above the true average length (4.19 in the 5K French-Mboshi corpus). This explains `pypshmm`'s degraded performance on Mboshi, visible in Figure 3.2.

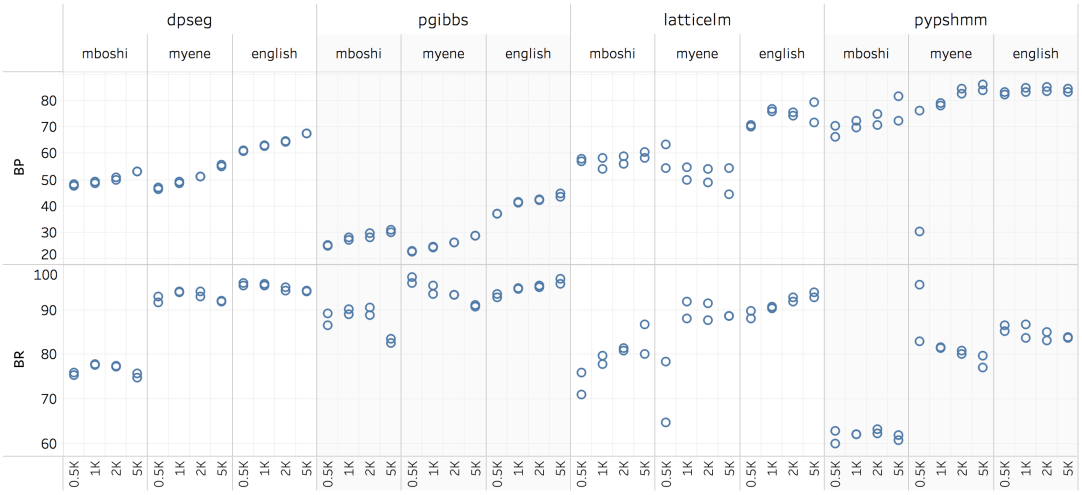


Figure 3.5: Boundary precision and recall (BP and BR) for all monolingual methods with source representations `notone` for all languages.

As, in theory, `pgibbs` implements a model that subsumes the DP-based bigram language model implemented by `dpseg`, it should be possible to achieve at least comparable results. In practice, `pgibbs` has a slightly different parametrization, and it is not straightforward to define a configuration that would be identical to that used for `dpseg`. Using `hyperopt` as we did in (Godard et al., 2016),[22] we are in fact able to find hyperparameters leading to competitive results, even though, as already mentioned in Section 3.3.1, this is not a practical approach for computational language documentation. These results, shown in Figure 3.6, are on par with `dpseg`, except for English data where the performance is still lower.

We also experiment with a 1-gram LM configuration for `dpseg` to further explore the surprising lack of benefit of using LMs with higher Markov dependency orders. In Figure 3.7, we observe in fact stronger results for the 1-gram configuration on our data, especially for the F-measure on types (LF). The superiority of a 2-gram LM, documented in (Goldwater et al., 2009) using child-directed speech data, is not apparent with our Mboshi and Myene data. This 1-gram LM configuration is also surprisingly stronger on English data. If we contrast these last results with a 2-gram LM configuration with parameters and hyperparameters optimized using `hyperopt`, the gap is almost

[22] That is, optimizing parameters and hyperparameters on the 0.5K extract of the French-English corpus.

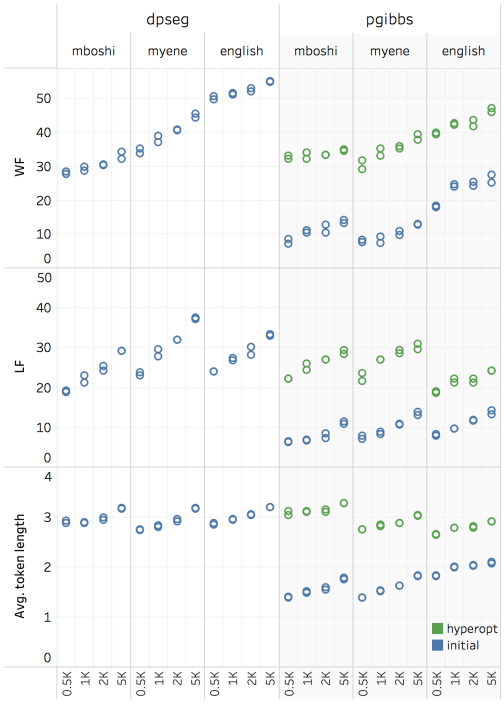


Figure 3.6: Token and type F-measure (WF and LF) and average token length obtained with `pgibbs`' hyperparameters optimized on the 0.5K extract of the French-English corpus using `hyperopt`. The initial results (Figure 3.2) for both `pgibbs` and `dpseg` are reproduced.

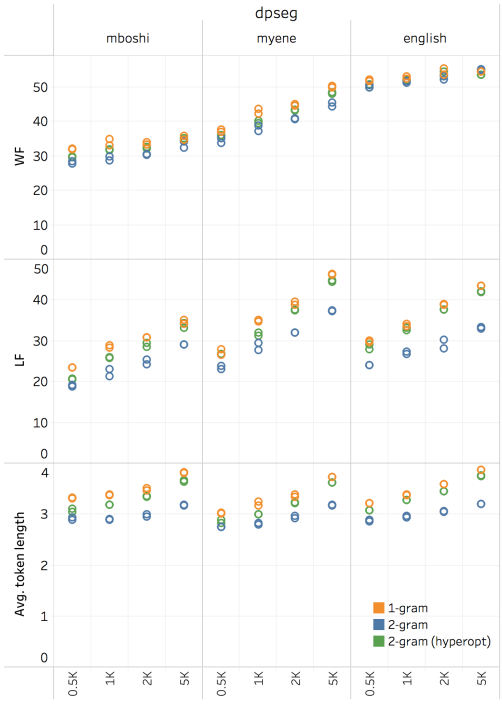


Figure 3.7: Token and type F-measure (WF and LF) and average token length for `dpseg` using a 1-gram LM, and a 2-gram LM with optimized parameters (`hyperopt`). To ease comparison, the initial results for `dpseg` (2-gram LM) are reproduced.

closed but the 1-gram results remain slightly stronger (Figure 3.7). As we will see in Chapter 4 (Section 4.4.1), the dependencies modeled between words are tightly coupled with the dependencies modeled between subword units or characters. As the strong results obtained with `latticelm` (with a 3-gram LM and a 3-gram SM) suggest, a higher LM dependency order is only detrimental when the SM dependency order is kept to 1 (both `dpseg` and `pgibbs` make use of a unigram character model).

Other contrasts As we already mentioned, the granularity of the source side in the bilingual approach does not have a strong influence on segmentation results. If we examine now the impact of the target representation in Mboshi (`notone` vs `tone`) for all methods (Figure 3.8), we cannot conclude that the presence of tonal information marked with diacritic symbols is helpful or detrimental to word segmentation. That said, representation `tone` leads to a larger character (and phoneme) inventory,[23] as well

[23] In `tone`, the 7 vowels present in Mboshi can now appear with an acute accent denoting a high tone. The number of character types thus becomes 31 (instead of 24 for `notone`).

as to a greater number of types in the lexicon.[24] This is likely to make the segmentation task harder at constant data size, and the fact that token and type metrics (WF and LF) end up being similar for `notone` and `tone` might indicate that tones could still be useful for segmentation. We will investigate this question in Chapter 5.

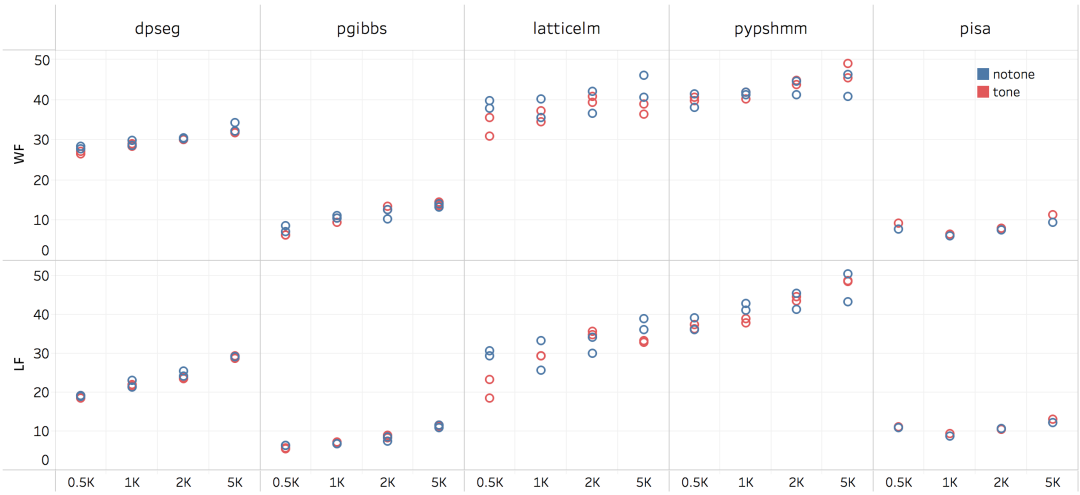


Figure 3.8: Token and type F-measure (WF and LF) for all methods in Mboshi using representations `notone` and `tone`.

Except for `pisa` on smaller corpora (Figure 3.3), and `latticelm` on Myene data (Figure 3.2), all methods show a clear improvement of their performance with more data. This can be seen in Figure 3.8, as well as in Figure 3.2, 3.3, 3.6, and 3.7.

Performance discrepancies across languages can be either explained by the construction of the corpora and their statistical properties, or by intrinsic linguistic properties. In all the results presented in this section, we observe better performance on English than on Mboshi.[25] A similar observation, in fact, motivated the study conducted by Fourtassi et al. (2013), where the English language was in fact shown to have favorable properties with respect to segmentation. The authors quantified these properties in terms of segmentation ambiguity, introducing the Normalized Segmentation Entropy (NSE). For a particular sentence of length N (in characters or phonemes) in the corpus, and given a reference lexicon, P_k denotes the probability[26] of each possible segmentation of this sentence, and the NSE is computed using Shannon's entropy (Shannon, 1948), with a normalization by the number of possible boundaries in the sentence:

$$NSE = \frac{-\sum_k P_k \log_2(P_k)}{N-1}.$$

[24] On the Mboshi side of the French-Mboshi 5K, `tone` representation yields 6,613 types (instead of 5,312 for `notone`, almost a 25% increase in the lexicon size).

[25] This might seem counter-intuitive, since, in our data, sentences are longer in English than in Mboshi (see Table 3.3 and 3.1), and more frequent sentence boundaries (which are also word boundaries) are likely to provide a useful supervision.

[26] Under a unigram model, whose lexical probabilities are defined by the (normalized) frequencies in the corpus.

In their experiments on English and Japanese (with adult and child directed speech data), the authors found a much lower NSE for English, showing an intrinsically lower segmentation ambiguity for this language when compared to Japanese. The same is likely to be true when comparing English to Mboshi.

Most of the time, results on Myene are also higher than those obtained on Mboshi, but this is more likely to come from the statistics of these corpora, as we will explain in Chapter 4. In both cases, construction of the corpora or linguistic properties, this highlights the need to collect and process actual under-resourced languages in order to get a fair approximation of the performance of natural language processing techniques.

3.4 Conclusion

In this chapter, we presented preliminary word segmentation experiments in Mboshi and Myene, two low-resource African languages from the Bantu family; the Mboshi corpus was released publicly to be used and studied by other researchers. Experiments were also contrasted with French-English data. Several monolingual and bilingual segmentation methods were assessed, with varying representation and size for the data.

One lesson learnt is that, with no more than 5K sentences, more complex and expressive models might not be able to perform better than simpler models. For instance, using a language model with a higher Markov dependency order does not necessarily yield better performance, and sometimes even leads to weaker results. Among the methods tested, `dpseg` produces stable and strong results despite a tendency to oversegment, while `pypshmm` obtains the best segmentations, especially on type metrics, but at the cost of a relative higher variance in the results. These two methods can therefore be considered as strong "off-the-shelf" tools, as they do not require hyperparameter tuning to be effective. This is crucial in a language documentation scenario, where the availability of segmented references cannot be assumed for a new language being studied. `latticelm` also constitutes a very competitive tool, especially when symbolic transcriptions will have to be replaced by streams or lattices of phonemes produced by an ASR system; we explore this scenario in (Ondel et al., 2018).

Another lesson learnt is that taking advantage of a bilingual supervision, in the form of translations into a well-resourced language, so as to obtain better segmentations, is not a straightforward goal with small corpora sizes. The bilingual method assessed in our experiments, `pisa`, in fact yields the least accurate segmentations in terms of boundary, token, and type metrics. In all fairness, this methods also provides an *alignment* information that we are not able to evaluate at the time of writing. This alignment step, though, is of utmost importance for linguists, as it associates unknown discovered units to known units. In fact, working with a translation is at the core of the practice of a field-linguist studying a new language.

A third important observation concerns the variability of the results when using Bayesian model with sampling methods, such as Gibbs sampling, and particular attention to this issue is needed to avoid drawing conclusions hastily.

Finally, our experiments demonstrate the need to use realistic low-resource corpora, instead of simulating the low-resource scenario with small quantities of WL data, in order to gain a more accurate understanding of the performance of unsupervised seg-

mentation methods. Some results discussed in this chapter were indeed counter-intuitive for us, and this encourages more work on such data.

These experiments are drawing us to find auxiliary sources of information capable of improving our results. In Chapter 4, we question whether expert linguistic knowledge can be such a source. We also explore the role of tones for word segmentation in Chapter 5, as the question was left unanswered. Lastly, and as a third auxiliary source, we seek in Chapter 6 to leverage more efficiently the bilingual supervision provided by translations.

Chapter 4

Adaptor Grammars and Expert Knowledge

Contents

After having established strong baselines for the fully unsupervised word segmentation task in Chapter 3, we focus, in this chapter, on incorporating expert linguistic knowledge to improve our results. Conversely, we investigate the possibility to gain new linguistic insights using word segmentation accuracy to validate linguistic hypotheses. We show that the Adaptor Grammar framework, introduced in Chapter 2, is a suitable tool for these two goals. Parts of the work presented in this chapter appears in (Godard et al., 2018b).

4.1 Introduction

As noted in Chapter 1, computational language documentation, as a field, is emerging from the realization that a tighter and more efficient collaboration between linguists and computer scientists is urgently needed if we are to meet the challenges of a large-scale language extinction. However, the two research communities involved in this endeavor often struggle to cooperate efficiently. Their knowledge backgrounds differ, and the definition of why a problem is interesting (or not) may vary depending on the two communities. To make matters worse, the theoretical and experimental platforms used by the researchers from both fields do not intersect much, which hinders concrete opportunities for a fruitful dialogue. Consequently, for lack of investing enough energy working on the same problems with the same tools and towards the same goals, we might not achieve the efficiency that is needed, as time is running out for many languages. This view constitutes the underlying motivation of the work presented in this chapter, which results from a close collaboration with linguist Annie Rialland, and many conversations with Martine Adda-Decker. Annie was instrumental in the process of encoding linguistic knowledge in the formalism of Adaptor Grammars (Johnson et al., 2007b), and in discussing the potential benefits of our approach for a linguist.

4.1.1 Using expert knowledge

We pursue two main goals along these lines. Our first goal is to use *expert knowledge*, when available, in order to improve upon the baselines established in Chapter 3 for the unsupervised word discovery task. In the context of a transdisciplinary project like BULB (see Section 1.2.2), it is possible to engage linguists in formalizing their linguistic knowledge regarding Mboshi and Myene (see Section 3.2 for the description of both corpora), in the hope that it will compensate for the small amount of available data. We would like, for instance, to take advantage of morphological and phonotactic constraints in the two Bantu languages, which display very similar structures. Additionally, a list of prefixes in Mboshi, and some additional knowledge regarding its consonantal system, are also at our disposal.

Such expert knowledge can readily be integrated in grammar rules using the framework of Adaptor Grammars (AGs, see Section 2.4.5). An interesting property of this framework is, indeed, its compatibility with two strategies usually thought to be mutually exclusive: rule-based approaches, still in wide use inside the linguistics community, and statistical learning, prevalent in natural language processing circles.

This approach is practical for many low-resourced languages, as most are not mere *terra incognita*. Linguists can link most under-studied languages to a language family, and provide minimal information, such as, e.g., a phoneme inventory. Moreover, resources such as the World Atlas of Language Structures (WALS)[1] gather phonological, lexical, and grammatical properties of a large number of languages from various descriptive studies, and this knowledge can thus be found more easily. We do not know, however, to which degree incorporating expert knowledge can improve word segmentation accuracy in a low-resource scenario.

[1] https://wals.info/

4.1.2 Testing hypotheses

Our second goal is to study ways to help linguists explore language data when little expert knowledge is available. Our proposal is to complement the grammatical description activity with task-oriented search procedures, that will speed up the exploration of competing hypotheses. The intuition is that better grammars should not only truthfully match the empirical data, but also improve the quality of automatic analysis processes.

Such an automatic process (e.g., the word segmentation task here), should thus be viewed as an extrinsic *validation procedure*, rather than a goal in and of itself. This process might also yield *new linguistic insights* regarding the language(s) under focus. In this chapter, we also study the practicality of this approach.

4.1.3 Related work

In Section 2.4.5, we already mentioned several extensions to the Adaptor Grammar framework. While AGs are essentially viewed as an unsupervised grammatical inference tool, several authors have also tried to better inform grammar inference with external knowledge sources. This is the case of Sirts and Goldwater (2013), who study a semi-supervised learning scheme combining annotated data (parse trees) with raw sentences. The linguistic knowledge considered in (Johnson et al., 2014) aims to better model function words in a language acquisition setting: explicitly representing the occurrence of these short (typically monosyllabic) tokens in front of content-bearing words was shown to improve the resulting word segmentations. The work of Eskander et al. (2016) considers the use of additional dictionaries, storing partial lists of prefixes or suffixes collected either on the Internet, or discovered during a first round of training. We study similar complementary information, which are collected in close collaboration with linguistic experts.

The main contribution of this chapter is a methodology for systematically exploring a subpart of the space of possible grammars, refining grammar rules at four levels of description, from the most generic to the most language specific (see Section 4.3). This results in a comparison of 162 alternative accounts of the grammar for two languages. Our results, analyzed in Section 4.4 show that enriching grammar rules with language specific knowledge has a consistent positive impact in performance for the word segmentation task. They validate our hypotheses that i) improved grammatical descriptions actually correlate with better automatic analyses; ii) Adaptor Grammars provide a framework around which linguists and computer scientists can effectively collaborate, with tangible results for both communities.

4.2 Word segmentation using Adaptor Grammars

AGs have been used to infer the structure of unsegmented sequences of symbols, offering a plausible modeling of language acquisition (Johnson, 2008b; Johnson and Goldwater, 2009); they have also been used for the unsupervised discovery of word structure, applied to the Sesotho language by Johnson (2008a). One notable outcome of the latter study was to demonstrate the effectiveness of having an explicit hierarchical model of the

internal structure of words; an observation that was one of our primary motivations for using AGs in our language documentation work. In this series of studies, AGs are shown to generalize models of unsupervised word segmentations such as the Bayesian nonparametric model of Goldwater (2006), delivering hierarchical (rather than flat) decompositions for words or sentences.

For the purpose of word segmentation, we first assume a linguistic grammar G, which parses sequences of letters (or phones) into "Words", which themselves recursively decompose into smaller units such as "Morphs", "Syllables", etc. To induce word segmentation from parse trees, we will consider that each span covered by the non-terminal symbol "Word" defines a linguistic word. It is important to note, however, that the parsing is unsupervised,[2] and that this non-terminal symbol might correspond in practice to linguistic units that are larger or smaller than a word. Figure 4.2 illustrates this on two example parses. Likewise, when examining the output of the training process, we are in a position to collect sets of word types (or morph types, syllable types, etc.) and will do so based only on the identity of the root symbol, i.e. without any certainty regarding the linguistic status of the collected sequences.

4.3 Grammars

We now present the methodology used to design a large set of grammars in order to perform contrastive experiments. We gradually vary the amount of expert knowledge taken into account, and ensure modularity so as to better isolate the contributions of each linguistic hypothesis.

4.3.1 Structuring grammar sets

Our starting point is the set of grammars used in (Johnson and Goldwater, 2009) and (Eskander et al., 2016) which we progressively specialize through an iterative refinement process involving a linguist. As we wish to evaluate specific linguistic hypotheses, the initial space of interesting grammars has been generalized in a modular, systematic, and hierarchical way as follows. We distinguish four sections in each grammar: sentence, word, syllable, character. For each section, we test multiple hypotheses, gradually incorporating more linguistic structure. Every hypothesis inside a given section can be combined with every hypothesis of any other section,[3] thereby allowing us to explore a large quantity of grammars and to analyze the contribution of each particular hypothesis. Recursive rules like "Words $\rightarrow$ Word Words; Words $\rightarrow$ Word" are abbreviated as "Words $\rightarrow$ Word +". Adapted non-terminals are underlined.

4.3.2 The full grammar landscape

All the grammar sections (sentence, word, syllable, character) experimented in this chapter are detailed in Figure 4.1. We now describe the way each section was designed.

[2] The design of the grammar constitutes a supervision, but the sampling procedure leading to the production of parses is fully unsupervised.

[3] Note that if a non-terminal is absent from a hypothesis (e.g. Syllable in a word level hypothesis), the corresponding non-terminal in the subsequent hypotheses (e.g. at the syllable level) will be ignored.

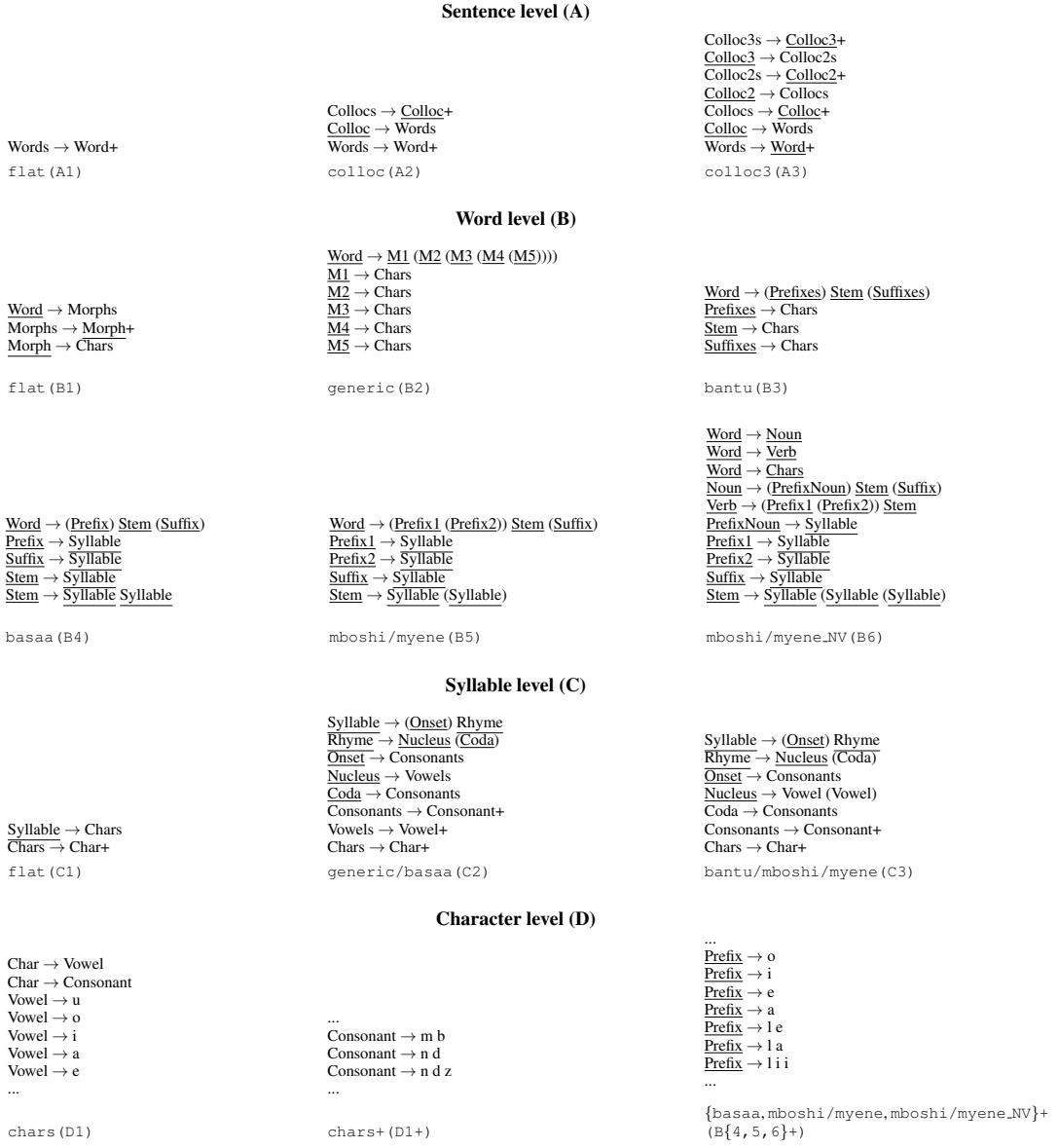
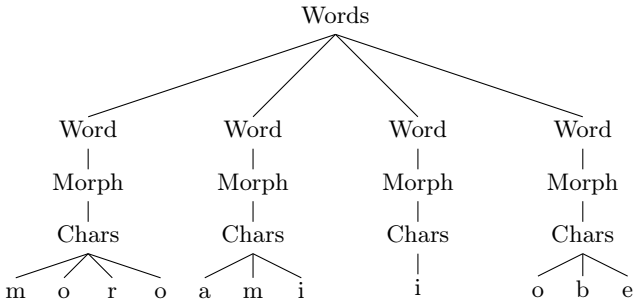
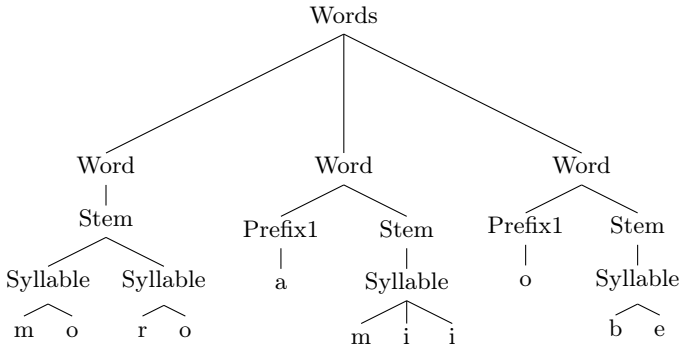


Figure 4.1: Grammar rules for all the hypotheses presented in Section 4.3.



(a) A3B1C1D1, *"moro ami i obe"* (incorrect word segmentation)



(b) A3B5C2D1+, *"moro amii obe"* (correct word segmentation)

Figure 4.2: Examples of parses – some non-terminals have been omitted for readability – obtained with two grammars, and the corresponding word segmentations for Mboshi sentence *"Moro a-mii o-be"*. (CL1.man 3SG-swallow.PST CL14-bad; since *Moro* is an irregular noun, the prefix and the stem are difficult to separate, which is signaled by a dot, following the Leipzig glossing rules.)

- sentence level: we model 3 different hierarchies of words. We first introduce the `flat` variety with two rules generating right-branching parse trees. `colloc` adds a single level of word collocation, aimed to capture recurrent local word associations (such as frequent bigrams); `colloc3` displays a deeper hierarchical structure with three levels of collocations. Exploring more realistic syntactic structures is left for future work.
- word level: here we propose 6 competing hypotheses. `flat` is similar to previous sentence variety but at the word level instead of the sentence level. `generic` corresponds to a more structured version of `flat`, as the specification of a sequence of 5 adapted morphemes allows, in principle, the Adaptor Grammar to learn some morphotactics. `bantu` defines a generic morphology for Bantu languages (also suitable for other language families). `basaa` implements the morphology of Basaa (A43), a Bantu language described in (Hamlaoui and Makasso, 2015). `mboshi/myene` corresponds to a somewhat crude morphology of Mboshi, also applicable to Myene. Last `mboshi/myene_NV` refines `mboshi/myene` with a specification of the morphology of nouns and verbs. Additionally, for `basaa`, `mboshi/myene`

and mboshi/myene_NV which introduce a notion of prefix, we also test a variant (called respectively basaa+, mboshi/myene+ and mboshi/myene_NV+) containing an explicit list of prefixes in Mboshi.

- syllable level: we contrast 3 hypotheses : flat is similar to previous sentence and word varieties but at the syllable level, defining the syllable as a mere sequence of characters. generic/basaa is a generic set of rules modeling phonotactics applicable to a wide scope of languages (including Basaa mentioned in the preceding level). bantu/mboshi/myene displays a set of rules more specific to Mboshi and Myene: the nucleus of a syllable contains at most two vowels, and the presence of a coda is discouraged.[4] Note that the difference between generic/basaa and bantu/mboshi/myene remains small.
- character level: rules in the chars set simply rewrite the characters (terminals) observed in our data. chars+ adds rules to capture the digraphs or trigraphs occurring in Mboshi (see details in Section 3.2.1).

4.4 Experiments and discussion

We now experiment along the methodology presented in Section 4.3 with two monolingual corpora: the first corpus is the Mboshi part of the French-Mboshi 5K (see Table 3.1); the second corpus is the Myene 5K (see Table 3.2). Both corpora are described in Section 3.2, and for each of them, we consider the representation notone (no tone markers, see Section 3.2.2). We report word segmentation performance using precision, recall, and F-measure on tokens (WP, WR, WF), and types (LP, LR, LF). We also report the exact-match (X) metric which calculates the proportion of correctly segmented utterances.[5] In all the figures, and in this section, we use the following compact names for grammatical hypotheses at each level:

- A1 (flat), A2 (colloc), A3 (colloc3),
- B1 (flat), B2 (generic), B3 (bantu), B4 (basaa), B5 (mboshi/myene), B6 (mboshi/myene_NV), with additional "+" variants for B4, B5, and B6 when a list of prefixes is provided, for instance B6+ (mboshi/myene_NV+),
- C1 (flat), C2 (generic/basaa), C3 (bantu/mboshi/myene),
- D1 (chars), D1+ (chars+).

For each language, we evaluate our 162 grammar configurations using Mark Johnson's code,[6] collecting parses after 2,000 sampling steps.[7] We adapt all non-recursive non-terminals and use a Dirichlet prior to estimate the rule probabilities. We place a

[4] In theory, we should not include a coda in this last hypothesis, but loanwords and proper names in our data made the Adaptor Grammar fail to parse without a coda. To decrease the impact of this rule, we chose not to adapt the corresponding non-terminal, in contrast to generic/basaa.

[5] The exact-match metric includes single-word utterances.

[6] http://web.science.mq.edu.au/~mjohnson/Software.htm

[7] The large number of experiments we are dealing with did not allow us to average over several runs. Stable results were obtained on a subset of grammars. Two particular configurations in Mboshi (A3-B6-C3-D1+ and A1-B6-C1-D1) did not reach 2,000 iterations within the maximum wall clock time allowed by the cluster used for these experiments (2 weeks), and are left out of the discussion.

uniform Beta prior on the discount parameter of the Pitman-Yor process, and a vague Gamma prior on the concentration parameter. Figure 4.3 presents token metrics and type metrics, as well as the sentence exact-match, for both corpora on all grammars.

4.4.1 Word segmentation results

We first analyse the impact, on word segmentation performance, of the choices made at each grammatical level, and subsequently observe several contrasts between both corpora, before comparing our results to the baselines established in Chapter 3.

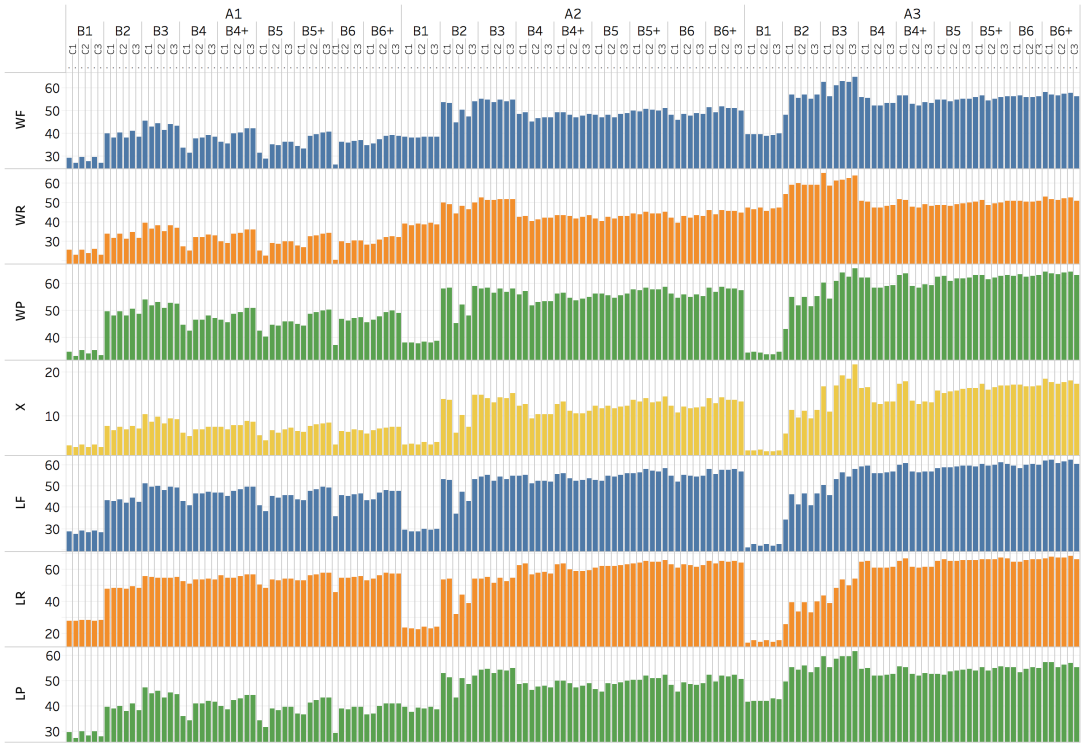
Impact of sentence level variants We can see in Figure 4.3 that A2 and A3 hypotheses globally yield better results than A1 in both languages. For Mboshi, the benefit of A3 vs. A2 appears especially on token metrics (WP, WR, WF), but this contrast is less clear on Myene. For both languages, however, our results confirm that modeling collocation-like word groups at the sentence level is important, an observation already abundantly documented for English in (Goldwater, 2006; Goldwater et al., 2009; Johnson and Goldwater, 2009). Experiments in Sesotho (Johnson, 2008a) showed a lesser impact of modeling collocations, but these word dependencies seem nevertheless related to a universal linguistic property.

One word of caution, though: when hypothesis A3 is coupled with hypothesis B1, this leads to the worse results for token and type F-measures (WF and LF) across all configurations in Myene, and for type F-measure (LF) in Mboshi. This might explain the surprising results obtained in Chapter 3 (Section 3.3.2) where higher order language models performed poorly due to oversegmentation, which only happened when character spelling models did not take dependencies between characters into account (1-gram SM). This modeling assumption could be compared, here, to the conjunction of hypotheses A3 and B1; such hierarchical imbalance can lead to a situation where the non-terminal "Word" in fact corresponds to either words or morphs.[8]

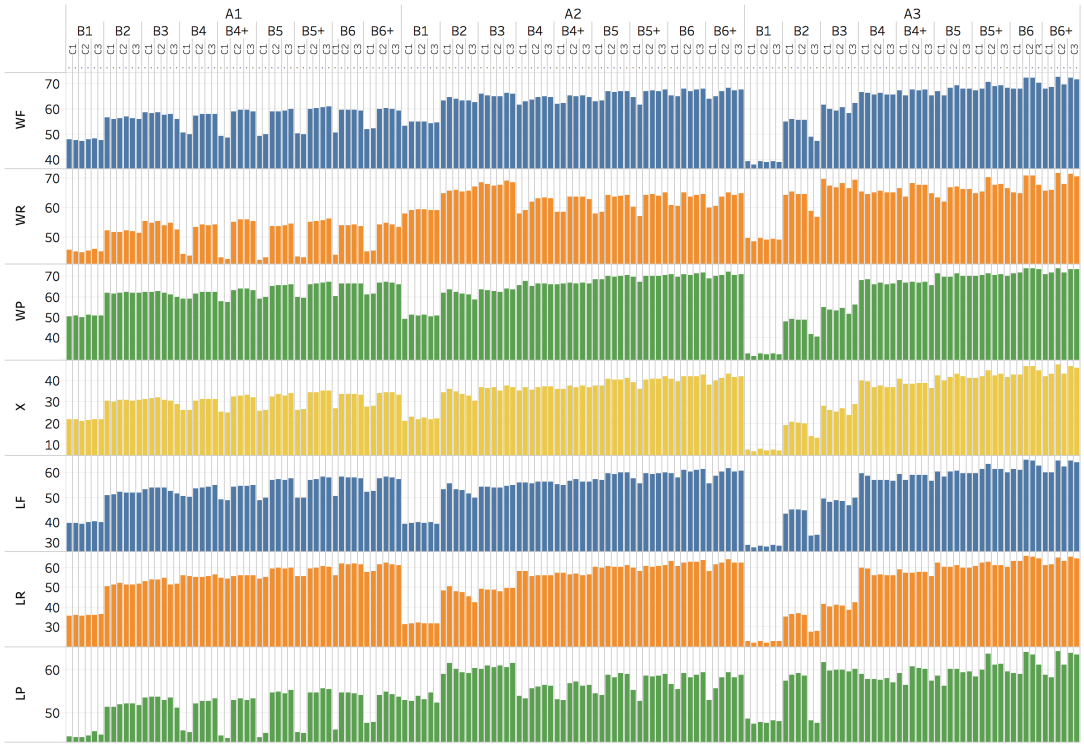
Impact of word level variants If we now focus solely on the A3 hypothesis for Myene in Figure 4.3, we observe a general trend upwards for all metrics. The benefit of gradually using more language-specific grammars, from B1 to B6+, is clear. While this trend is also observed for Mboshi, the less specific B3 hypothesis yields the strongest results on token metrics (WP, WR, WF). Precision on types (LP) with B3 is also the strongest, but B6+ achieves better performance on type recall and F-measure (LR and LF). The contrast between B1 and B2 for all metrics on both languages (keeping a focus on A3, but this can also be seen for A1 and A2) highlights the benefit of modeling some morphotactics inside the word-level hypotheses, which seems to correspond to another universal linguistic property (the dependency between morphemes inside a word).

Impact of syllable level variants It is difficult to see a clear trend for the impact of syllable-level variants in Figure 4.3. Importantly, the syllable level will only be effective when combined with word level variants B4, B5 and B6 (and their "+" versions)

[8]See also (Fourtassi et al., 2013) for a discussion about the choice of the non-terminal level used for the evaluation.



(a) Mboshi corpus



(b) Myene corpus

Figure 4.3: Word segmentation performance evaluated with token metrics (WP, WR, WF), type metrics (LP, LR, LF), and sentence exact-match (X) for Mboshi (top) and Myene (bottom). All grammars are broken down by A, B, C, and D levels (D1 shown before D1+).

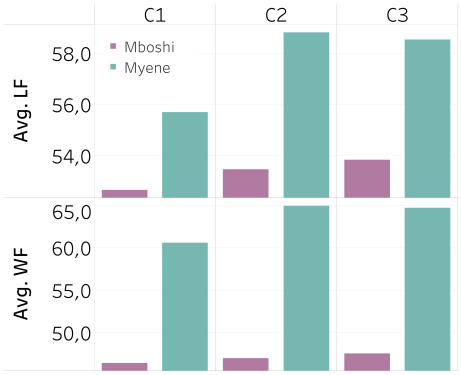


Figure 4.4: Impact of C variants on Mboshi and Myene. Token F-measure (WF) and type F-measure (LF) are averaged over hypotheses B4, B4+, B5, B5+, B6, and B6+.

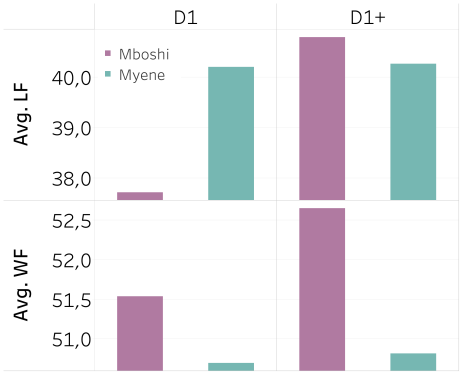


Figure 4.5: Impact of D variants on Mboshi and Myene. Token F-measure (WF) and type F-measure (LF) are averaged over hypotheses with A level set to A3 and B level set to B1, B2, or B3.

which model the concept of syllable: when combined with B1, B2 or B3, each C level hypothesis will default to its "Chars → Char +" rule. Figure 4.4 illustrates the impact of C1, C2, and C3 by averaging type and token F-measures (LF and WF) over all grammar sections with a syllable non-terminal (B4, B4+, B5, B5+, B6, and B6+). The benefit of C2/C3 vs. C1 appears more clearly, especially on type F-measures and for Myene.[9] Nevertheless, the impact of the syllable level, and the capacity to incorporate phonotactics in our models, seems of less significance for word segmentation than choices made at the word and sentence levels.

Impact of character level variants In Figure 4.3, it is also hard to see if there is any benefit in using D1+ over D1, i.e. adding digraphs or trigraphs to the consonant inventory. Averaging over all hypotheses at the A, B, and C levels does not show any clearer impact. It is likely that refined models at the syllable level (C) compensate for a less accurate consonant inventory through the adaptation of their non-terminals, and do learn some phonotactics. This would explain the weak contribution of D1+. To test this hypothesis, we set the sentence level to A3 (the best compromise for Mboshi and Myene) and the word level to B1, B2, or B3 (levels without a Syllable non-terminal, in order to cancel the effect of the syllable level C). The token and type F-measures averaged over the considered hypotheses are shown Figure 4.5. We do observe a benefit in using the D1+ character variant in Mboshi, but not in Myene. This is not surprising, as the digraph and trigraph rules added by the D1+ variant are specific to Mboshi and do not cover the inventory for Myene.

Stronger results in Myene Segmentation performance is globally superior in Myene. This can probably be explained by corpora statistics (see Tables 3.1 and 3.2 in Chapter 3), as the average number of words per sentence is 3.94 in Myene, and 5.96 in Mboshi.

[9]The differences between C2 and C3, two very similar hypotheses, are hardly significant.

Since sentence boundaries are also word boundaries, the proportion of already known word boundaries is higher in Myene, which makes word segmentation easier. This effect is further amplified by the much larger number of single-word utterances in Myene (364) compared to the number found in Mboshi (11). Single-word utterances are, in effect, already properly segmented and are likely to provide a very useful supervision for the model in Myene.

Figure 4.3 also reveals an interesting contrast: token results are higher than type results in Myene, while the converse is true in Mboshi. The token/type ratio (5.75 tokens for one type in Mboshi, and 4.30 in Myene) indicates a higher lexical diversity in Myene, which might explain weaker results on types. Strong results on types for Mboshi, on the other hand, suggest that AGs have the capacity to generalize well on low-frequency events, a property of particular interest in the low-resource scenario.

Comparison to our baselines Overall, our best performing grammars in terms of token F-measure are A3-B3-C3-D1+ for Mboshi (64.78%) and A3-B6+-C2-D1 for Myene (72.62%). As for type F-measure, the best performing grammars are A3-B6+-C3-D1 for Mboshi (62.25%) and A3-B6-C2-D1 for Myene (64.96%).

These token F-measures are about 30 points higher than those obtained with `dpseg`, the Dirichlet process-based bigram word segmentation system of Goldwater et al. (2006a, 2009),[10] which yields 33.26% token F-score on Mboshi and 44.92% on Myene[11] (see Section 3.3 in Chapter 3). With `pypshmm` (Löser and Allauzen, 2016), the best token F-measure obtained on Mboshi during preliminary experiments (43.60%) is still 20 points below our results, but the discrepancy is smaller on Myene (64.72%, to compare to 72.62%). Similar gaps are observed on type F-measures for both `dpseg` and `pypshmm`.

In addition, if we consider the F-measure computed on word boundaries (not reported here), the performance reaches 82.34% for Mboshi (A3-B3-C3-D1+) and 86.33% for Myene (A3-B6+-C2-D1). The corresponding results (see "BF" in Figure 3.2) with `dpseg` were respectively 62.25% (Mboshi) and 69.09% (Myene), while `pypshmm` improved these results to 68.14% (Mboshi) and 81.51% (Myene). Overall, our linguistically-informed AGs enable a quite substantial jump in word segmentation performance when compared to the strong baselines established in Chapter 3.

4.4.2 How can this help a linguist?

Our second goal is to understand more precisely how such experiments can be useful for linguists, beyond the benefit of having access to better automatic word segmentation tools for their data.

Phonological status of complex consonants In the analysis of the results (Section 4.4.1 above) we showed the benefit of integrating digraphs or trigraphs in the consonants inventory for Mboshi. This result is of special interest for linguists, since it is in line with the most recent phonological analyses of Mboshi (Embanga Aborobongui, 2013; Kouarata, 2014; Amboulou, 1998) which agree in recognizing complex consonants

[10] `https://homepages.inf.ed.ac.uk/sgwater/resources.html`.
[11] Results for `dpseg` or `pypshmm` are averaged over two runs.

reference	a i e o le la laa lii lee loo baa bii boo maa mo moo mi mii yee
discovered	a o i e la waa laa mo yee le ya ngai mi ɯ sa ma baa lo lii me

Figure 4.6: An example of the 20 most frequently found prefixes in Mboshi, using grammar A3-B4-C3-D1 (no supervision). True discovered prefixes, present in the reference, are underlined.

(represented by digraphs or trigraphs) in the phonological inventory of this language. The analysis of complex consonants, in particular prenasalized consonants, has generated many debates in Bantu linguistics (Odden, 2015; Herbert, 1986; Downing, 2005); the present experiments provide more substance to support the integration of complex consonants in the phonological inventory of Mboshi.

Learning prefixes without supervision Since parses are produced to segment sentences into words, it is possible to extract the most frequent prefixes or suffixes (for B variants introducing such a concept). The precision on the 20 most frequently found prefixes for grammars without prefix-supervision (B3, B4, B5 and B6)[12] reaches 58.21% in Mboshi, and 61.21% in Myene. An example of discovered prefixes is shown in Figure 4.6. The capacity of AGs to learn true prefixes without supervision can thus help linguists in the process of documenting a new language.

On the supervised variants (B4+, B5+, and B6+) including rules corresponding to prefixes in Mboshi (e.g. "Prefix → a"), the average precision achieved in Mboshi is 61.11%, and 63.07% in Myene: the benefit of the supervision is limited. Yet, it holds for all B variants with a concept of prefix, as shown in Figure 4.7.[13] Token metrics for

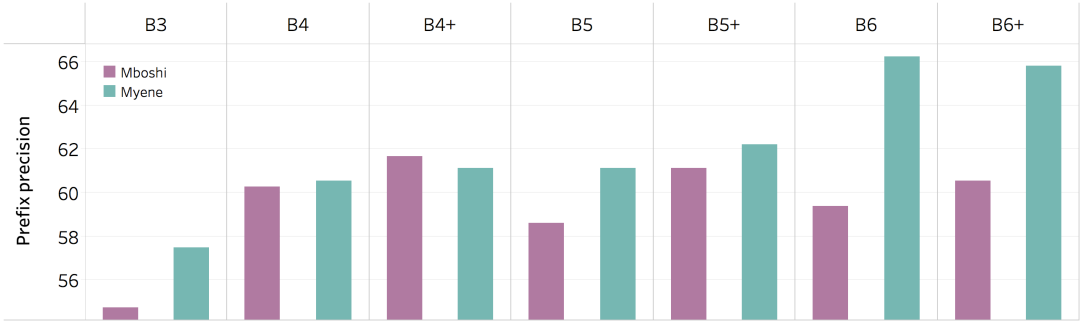


Figure 4.7: Precision on the 20 most frequently found prefixes in Mboshi and Myene for hypotheses B3, B4, B4+, B5, B5+, B6, and B6+ (precision is averaged over all A, C, and D variants).

Mboshi with the supervised variants also indicate a benefit for word segmentation (WF in Figure 4.3, top). It is important to note that the supervision provided in B4+, B5+,

[12] We include B3 variant, interpreting its non-terminal Prefixes as a prefix.

[13] The exception on B6+ for Myene can be explained by the two grammar configurations we left out (see footnote 7). Once reintegrated in Myene, B6+ has a higher precision than B6.

and B6+ does not guarantee that the Adaptor Grammars will recover all the prefixes. For one thing, sentences can be parsed ignoring these additional rules. But we also do not know whether all Mboshi prefixes are present or not in the data.

Hierarchy and complexity Recursivity is an important property of language, even considered a specificity of human language (Hauser et al., 2002; Everett, 2005), and can be distributed differently across languages (Evans and Levinson, 2009). It is tightly linked to the presence of hierarchical structures in linguistic data. The notion of linguistic complexity,[14] on the other hand, relates to recursivity but also involves other parameters, such as word or morpheme category and constituency. Additionally, linguistic complexity is sometimes understood in terms of the difficulty experienced by a language learner (Pallotti, 2015). To compare various levels of hierarchy and complexity, across or within languages, using AGs with a systematic grammar landscape and various word segmentation metrics, as we do here, opens the possibly to support quantitatively certain linguistic hypotheses relative to these notions. More specifically, A variants correspond to three different levels of hierarchy. At level B, variants B1, B2, B3 do not correspond to hypotheses of increasing hierarchical depth, but to hypotheses with different levels of complexity (e.g. distinguishing between morph's positions); conversely, B4, B5, and B6 introduce a new hierarchical level (Syllable), as well as an increasing morphological complexity. Lastly, C2 and C3 display the same hierarchical levels with comparable complexity, while C1 is simpler and flatter.

To understand how using AGs with this approach can help support quantitatively a linguistic hypothesis, we can look at word segmentation as a weak form of supervision for parsing. When a word segmentation is incorrect, the corresponding parse is also deemed to be incorrect. Conversely, if the word segmentation is correct, the parse might (or might not) be correct above and below. That is, the word segmentation task provides a way to rule out a fraction of the erroneous parses. The same reasoning can hold at other levels of description: identifying correct prefixes indicates a local adequacy of the parses and possibly above and below, while failing to accurately segment prefixes reveals an incorrect parse, and thus a flaw in the grammatical hypothesis. For that reason, we think that the methodology presented in this chapter can help to contrast grammatical descriptions: if it cannot formally validate a particular hypothesis, it can identify hypotheses yielding a larger quantity of possibly accurate parses.

It is interesting to visualize, in Figure 4.8, the regularities appearing for the various combinations of level A and B. The best performing grammars in terms of token F-measure[15] correspond to the true average token length – this could not be otherwise. But we can also clearly see how the combinations between levels A and B determine the average length of a word, and how augmenting the hierarchy at level A, all things otherwise equal, decreases this length, while augmenting the hierarchy at level B increases it. Using word segmentation as a weak supervision for parsing, i.e. ruling out grammars yielding on average a higher number of inaccurate parses, this suggest an iterative approach in which grammar descriptions could be successively refined (e.g. "around" A3B3 for Mboshi, and A3B6+ for Myene in our experiments).

[14] A harder notion to define (Pallotti, 2015).
[15] The same curves are observed for type F-measures.

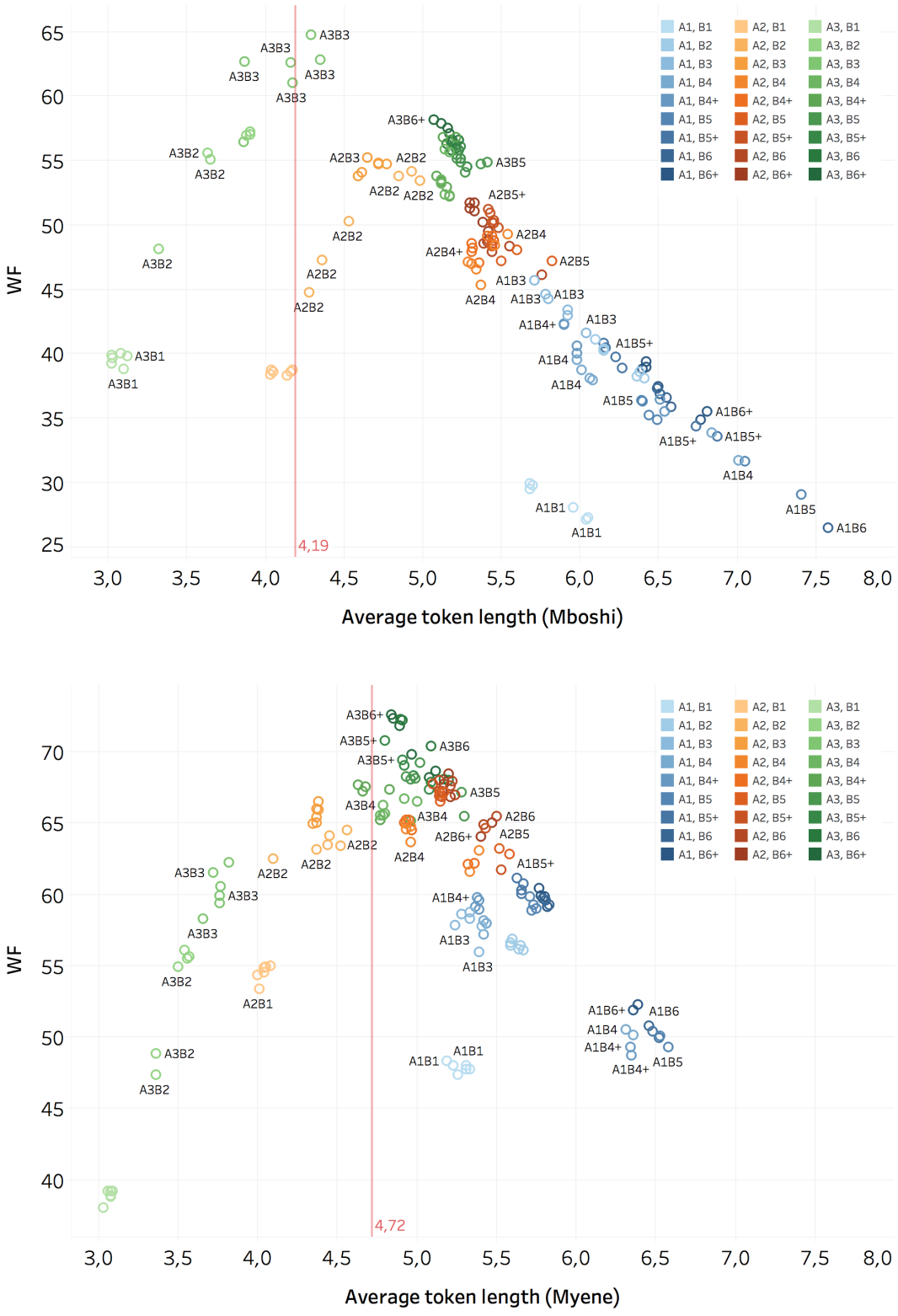


Figure 4.8: Token F-measure plotted against average token length for all the grammars. Colors and labels indicate hypotheses in levels A and B (only some labels are represented for readability). Vertical red lines indicate the true token average length for both corpora.

4.5 Conclusion

In this chapter, we were pursuing two main goals: i) improve upon strong baselines for the unsupervised discovery of words in two very low-resource Bantu languages; ii) explore the Adaptor Grammar framework as an analysis and prediction tool for linguists studying a new language.

Systematic experiments with 162 grammar configurations for each language have shown that using AGs for word segmentation is a way to test linguistic hypotheses during a language documentation process. Conversely, we have also shown that specializing a generic grammar with language specific knowledge greatly improves word segmentation performance.

In addition, the best word segmentation results using this approach are way higher than our Bayesian baselines, and the conjugated positive effects of non-terminal adaptation and structure supervision by an expert are manifest. This proves the necessity to further foster collaborations between linguists and computer scientists, in order to speed up the documentation process and to meet the challenges of language documentation and preservation. The development of intuitive user interfaces, on one hand, and a particular attention to online learning approaches, on the other hand, could help make these collaborations more productive.

The main limit of the results presented in this chapter lie in the absence of bilingual grounding for the units we discover. A linguist could use discovered words or affixes to, then, identify regularities with respect to the translations. However, this process should be automated, and discovered units aligned to known units in the well-resourced language. This constitutes an important motivation for the work presented in Chapter 6.

Chapter 5

Towards Tonal Models

Contents

In Chapter 3, we established strong baselines for the unsupervised word segmentation task in a low-resource scenario, and in Chapter 4 we greatly improved our results using expert knowledge. In this chapter, we investigate a question raised during preliminary experiments from Chapter 3: are tonal features useful for unsupervised word discovery? Considering Mboshi, we first show that tone annotation improves the performance of *supervised learning* when using a simplified representation of the data. To leverage this information in an unsupervised setting, we then propose two probabilistic models based on a hierarchical Pitman-Yor process that incorporates tonal representations in its backoff structure. We compare these models with the tone-agnostic baseline already used in Chapter 3, and analyze if tone helps unsupervised segmentation on our small dataset. The work presented in this chapter is the result of a collaboration with Kevin Löser and has appeared, in part, in (Godard et al., 2018c). The code base used for the new models experimented here has been written by Kevin Löser (Löser and Allauzen, 2016). We collaborated on the design and implementation of the new methods, and I contributed the preliminary study and the experiments.

5.1 Introduction

We observed, in Chapter 3, that the segmentation results obtained in Mboshi with representation `tone` (including tonal markers) were similar to those obtained with representation `notone` (in which markers are removed). We nevertheless argued in Section 3.3.2 that more investigation was needed. In representation `tone`, the characters' inventory and the size of the lexicon are larger, and this is likely to make parameter estimation for the segmentation model a harder task if no more data is provided. Since we did not observed any decrease in performance with representation `tone`, we posited that this could indicate a usefulness of tonal information for segmentation. In this chapter, we attempt to untangle these two contradictory effects in using the tonal representation: on one hand the representation is richer than its non-tonal counterpart, and contains information that might correlate with word boundaries; but on the other hand, it causes the inventory of symbols to increase, and thus the number of types in the lexicon, making the estimation of the model's parameters harder than it is already, since we have very small quantities of data. In other words, we seek to answer the following question: does tone matter for unsupervised word discovery in low-resource conditions?

Many of the nonparametric Bayesian models we described and experimented with in the previous chapters were originally designed as computational models of language acquisition. They have mostly been applied to non-tonal languages, but Johnson and Demuth (2010) investigated the use of Adaptor Grammars for unsupervised word segmentation of Mandarin Chinese. Tones were shown to have a small impact on segmentation accuracy and were reported to yield a small improvement for simple grammars, but no improvement with more complex ones. In an earlier work, Johnson (2008a) presented results on Sesotho, a tonal Bantu language, yet without specifically discussing the role of tones. Also relevant is the work of Ludusan et al. (2015), who studied the role of prosodic information in word segmentation. The approach was tested on English and Japanese, and for both languages it was shown that prosodic boundaries were helping word segmentation.

In the present chapter, we study tonal information in Mboshi, where the distribution of tones obey morphological, lexical, and syntactical rules (see Section 3.2.1). In order to assess the presence of a usable tonal signal in our data, we first design a *supervised* experiment to segment various representations of the Mboshi data (Section 5.2). We show that tones help disambiguate word boundaries in the supervised setting, suggesting that tonal information could also contribute to the *unsupervised* discovery of words. We then present, in Section 5.3, two new hierarchical nonparametric Bayesian models, which use generalized backoff schemes to integrate tonal representations. Based on the experiments reported in Section 5.4, we conclude that taking advantage of tonal regularities is hard with very small corpora, but that the proposed new models have the potential to capture such information.

5.2 A preliminary study: supervised word segmentation

If tones can help predict word boundaries in a tonal language, it should be possible to identify useful tonal features in a supervised segmentation experiment. In this spirit,

we train decision tree classifiers with various features corresponding to alternative representations of the Mboshi data, and compare segmentation predictions.

5.2.1 Data and representations

We consider the Mboshi part of the French-Mboshi 5K corpus described in Chapter 3 (Section 3.2). In our transcriptions, Mboshi's tonal system is simply represented using diacritics on vowels: the presence of an acute accent marks a high tone, and its absence, a low tone (see also Section 3.2.2). Our approach consists in varying the representation of the input text, and comparing the full transcription with diacritics (`tone`) to:

- the transcription `notone` where diacritics are removed;
- the transcription `xV` where vowels are replaced by the symbol '`V`';
- the transcription `xLH` where high-toned vowels are replaced by the symbol '`H`' and low-toned vowels are replaced by '`L`';
- the transcriptions `CV` (resp. `CLH`) where consonants in `xV` (resp. `xLH`) are replaced by a generic symbol, '`C`'.

We expect the systematic comparison of tonal representations (`tone`, `xLH`, and `CLH`) with their non tonal counterpart (`notone`, `xV`, `CV`) to shed some light on the usefulness of this information. Additionally, we consider two synthetic tonal representations built from the `notone` transcription:

- `regular` representation marks each final vowel of each word with a high tone (other vowels are thus marked with a low tone, which is the default in `notone`);
- `random` representation marks each vowel of each word with a high tone with probability 0.5 (and otherwise keeps the low tone default).

These two synthetic datasets correspond to edge cases, where tonal information is either irrelevant to segmentation (`random`), or perfectly correlated to the presence of word boundaries (`regular`). Type statistics (words and symbols) and representation examples are given in Table 5.1. In particular, we observe that the Mboshi 5K corpus with representation `tone` contains about 20% more word types than when represented with `notone`; a similar increase in the number of symbols (the 7 vowels found in Mboshi have a high-tone variant in `tone`) is also observed.

5.2.2 Disambiguating word boundaries with decision trees

For each representation of the text, we train a decision tree classifier[1] to predict a binary decision corresponding to the presence or absence of a word boundary after each character.[2] This prediction is based on features encoding a fixed-length window of characters centered around the decision point. In our Mboshi corpus, we hold out 10% of the sentences for testing purposes, and train the classifier on the rest of the

[1] We use `scikit-learn`'s implementation (`http://scikit-learn.org/stable/modules/tree.html`).

[2] In effect, this means ignoring the word boundary before the beginning of the first word of each sentence; we also ignore the word (and sentence) boundary at the end of each sentence's last word.

representation name	#word types	#symbol types	transcription
notone	5,312	24	wa ayɛɛ la midi
tone	6,613	31	wa áyɛɛ la midí
xV	2,349	18	wV VyVV lV mVdV
xLH	4,081	19	wL HyLL lL mLdH
CV	220	2	CV VCVV CV CVCV
CLH	831	3	CL HCLL CL CLCH
regular	5,312	31	wá ayɛɛ́ lá midí
random	10,229	31	wa áyɛ́ɛ la mídí

Table 5.1: Type statistics and representation examples for the Mboshi 5K corpus. For all representations, the number of tokens found in the 5,130 sentences of the corpus is 30,556; the average sentence length (number of tokens per sentence) is 5.96, and the average token length (number of characters per token) is 4.19.

representation name	BP	BR	BF
notone	91.13	91.20	91.16
tone	91.32	90.48	90.90
xV	86.25	86.04	86.15
xLH	87.44	87.47	87.45
CV	64.89	48.89	55.77
CLH	71.11	61.10	65.73
regular	99.21	99.21	99.21
random	84.31	84.14	84.22

Table 5.2: Precision, Recall and F-measure on word boundaries in various text representations of the Mboshi 5K corpus with a decision tree classifier (5-words half-window width).

data. After experimenting with character windows of varying size, with padding at the beginning and end of the sentence, we determined that a half-window size[3] of 5 characters produced the best results for most representations, and allowed for a fair comparison. In Table 5.2, we report the corresponding boundary precision, recall and F-measure (BP, BR, and BF, as defined in Section 2.1.2).

For the pseudo-orthographic (notone and tone) text representations, it seems that the tonal information is of little value to disambiguate the frontiers, and even harms performance. Adopting a coarser representation for vowels in xV and xLH, we observe a more favorable situation for the tonal representation (xLH), even though the difference is small. However, as we simplify the text representation even further, with all consonants denoted with symbol C, and despite a drop in all boundary metrics, the contrast between a representation ignoring tones (CV) and one that captures them (CLH) becomes much

[3]Defining the number of characters before and after the character after which a boundary is considered.

clearer, with representation `CLH` yielding a boundary F-measure (BF) 10 points higher. This strongly supports the idea that a tonal signal can be used to improve segmentation.

Comparing synthetic representations `regular` and `random` to representation `notone` is also informative. While `regular` leads to almost perfect segmentations and confirms the capacity of the classifier to leverage regular tonal patterns, results obtained with `random` representation show that augmenting the character inventory without any information gain makes the segmentation task indeed harder: a drop of about 7 points in boundary F-measure is observed. The absence of an equivalent drop in performance for representation `tone`, with a character inventory identical to its `random` counterpart, indicates a tonal regularity in our data that is useful in these segmentation experiments, supporting the hypothesis made in Section 3.3.2.

5.3 Nonparametric segmentation models with tone information

These results motivate the design of new models which could capture the tonal signal directly on pseudo-orthographic representations and in the absence of any supervision. Our baseline is the model introduced by Löser and Allauzen (2016), whose implementation, `pypshmm`, was used in the experiments presented in Chapter 3, and obtained the strongest segmentation results. In this section, we describe a tonal extension of `pypshmm`'s *spelling model*, aiming to leverage tonal information to improve segmentation.

5.3.1 Language model

As discussed in Section 2.4.1, Pitman-Yor processes (PYPs) are a class of stochastic processes used to model sparse probability distributions with a countably infinite support. These distributions follow a power law, and are therefore especially suited to model distributions arising in linguistic data. We remind the reader that a Pitman-Yor process $\mathrm{PYP}(\alpha, \beta, G_0)$ is defined by some base distribution G_0, and two *concentration* ($\alpha \in]-\beta, \infty[$) and *discount* ($\beta \in [0, 1[$) hyperparameters; this process generates sparse versions of G_0, whose degree of sparsity is controlled by α and β. In the Chinese restaurant metaphor, assuming the presence of $i-1$ customers seated at $K(\mathbf{z}_{-i})$ tables according to a particular seating arrangement $\mathbf{z}_{-i}$, where each table is labelled with one of the previously generated words $\mathbf{w}_{-i}$, the conditional probability to generate word w at step i is given by

$$P(w_i = w \mid \mathbf{w}_{-i}, \mathbf{z}_{-i}) = \frac{n_w(\mathbf{w}_{-i}) - K_w(\mathbf{z}_{-i}, \mathbf{w}_{-i}) \cdot \beta + (K(\mathbf{z}_{-i}) \cdot \beta + \alpha) \cdot G_0(w)}{i - 1 + \alpha}, \tag{5.1}$$

where $n_w(\mathbf{w}_{-i})$ is the number of times word w has been previously generated, and $K_w(\mathbf{z}_{-i}, \mathbf{w}_{-i})$ the number of tables labelled w in the restaurant (Equation (2.12) corresponds to the equivalent conditional distribution with a Dirichlet process).

We model a sentence as a concatenation of words drawn from a distribution P_{LM}, itself drawn from a $\mathrm{PYP}(\alpha, \beta, P_{\mathrm{SM}})$. P_{SM}, the spelling model, defines a distribution over word forms (additional details in Section 5.3.2). We tokenize a corpus $s_1, \ldots, s_n$

of n sentences by Gibbs-sampling every segmentation s_i conditioned on all other segmentations. Following Mochihashi et al. (2009), we use a forward filtering-backward sampling algorithm to sample segmentations; as this method only approximates the posterior distribution, a Metropolis-Hastings correction step is also performed. P_{LM} is never explicitly represented,[4] but sampling $w \sim P_{\text{LM}}$ can be done by maintaining a table assignment associated to P_{LM}, such that for every token t there is a customer seated at a table labelled with the type of t. We also place agnostic priors on the PYP hyperparameters,[5] and resample these hyperparameters as well as the table assignments every 200 iterations.

5.3.2 A spelling model with tones

The spelling model defines a distribution over character strings that reflects how word forms should look like. Standard candidate spelling models are n-gram models of character sequences $w = u_1, \ldots, u_K$[6] defined by:

$$P_{SM}(w) = P(\text{stop} \,|\, u_{K-n+1}) \cdot \prod_{k=1}^{K} P(u_k \,|\, u_{k-n+1} \ldots u_{k-1}) \,. \tag{5.2}$$

In experiments presented in Chapter 3, introducing such n-gram dependencies to the spelling models of various methods proved to be essential to improve our results.[7] As we want to learn regularities inside tonal patterns, we need to extract tone features from the surface form of the characters. This is in fact reminiscent of the factored language models introduced by Bilmes and Kirchhoff (2003), where words (in their case) are represented as bundles of features – for instance morphological classes, roots, etc. In these models, it is not only possible to back off to a shorter word history, but also, say, to a part-of-speech history. The notion of a *backoff graph* is also introduced in this work, where parallel backoff is allowed. We do not explore parallel backoff in here, but our approach relies on designing backoff schemes making use of tonal features.

Contexts and backoff In order to also integrate tone information in the spelling model P_{SM}, we define the set of *contexts* $\mathcal{K}$ as the set of sequences $\tau_1, \ldots, \tau_j, u_{j+1}, \ldots, u_k$, with $k \in \{0, \ldots, n-1\}$ and $j \in \{0, \ldots, k\}$, where τ_i are *tone* symbols from the set $\{\texttt{H}, \texttt{L}, \texttt{C}\}$ (high tone, low tone, consonant), and u_i are regular characters. A *context* is thus a sequence of length at most $n-1$ comprising a prefix of tones and a suffix of characters.[8]

For every non-empty context $\kappa \in \mathcal{K}$, we further assume that the conditional distribution $P(u \,|\, \kappa)$ recursively arises from a PYP with base distribution $P(u \,|\, \varphi(\kappa))$, where φ is a *backoff function*. The base distribution of the unigram distribution (κ is empty) is the uniform distribution. In this setting, base distributions of PYPs themselves arise

[4] It has an infinite support.

[5] $\alpha \sim \exp(1)$ and $\beta \sim Beta(1, 5)$.

[6] where we add initial padding symbols as needed.

[7] while increasing the Markov dependency order for the language models surprisingly did not help learn better segmentations.

[8] Prefix or suffix can be empty.

from PYPs, giving rise to a *hierarchical* PYP (Teh, 2006) whose structure is defined by the backoff function φ.

Spelling model variants We first design a backoff scheme φ_{MULTI} where characters are first replaced by tones (rightwards), then dropped (rightwards), as follows:

$$\begin{aligned} u_1, \ldots, u_{n-1} &\xrightarrow{\varphi_{\text{MULTI}}} \tau_1, u_2, \ldots, u_{n-1} \xrightarrow{\varphi_{\text{MULTI}}} \tau_1, \tau_2, u_3, \ldots u_{n-1} \\ &\xrightarrow{\varphi_{\text{MULTI}}} \ldots \xrightarrow{\varphi_{\text{MULTI}}} \tau_1, \ldots, \tau_{n-1}, \xrightarrow{\varphi_{\text{MULTI}}} \tau_2, \ldots, \tau_{n-1} \\ &\xrightarrow{\varphi_{\text{MULTI}}} \tau_3, \ldots, \tau_{n-1} \xrightarrow{\varphi_{\text{MULTI}}} \tau_{n-1}, \xrightarrow{\varphi_{\text{MULTI}}} \emptyset\,. \end{aligned} \tag{5.3}$$

On a Mboshi example, backoff will thus unfold as follows (with $n = 4$): `ámid` $\xrightarrow{\varphi_{\text{MULTI}}}$ `Hmid` $\xrightarrow{\varphi_{\text{MULTI}}}$ `HCid` $\xrightarrow{\varphi_{\text{MULTI}}}$ `HCLd` $\xrightarrow{\varphi_{\text{MULTI}}}$ `HCLC` $\xrightarrow{\varphi_{\text{MULTI}}}$ `CLC` $\xrightarrow{\varphi_{\text{MULTI}}}$ `LC` $\xrightarrow{\varphi_{\text{MULTI}}}$ `C` $\xrightarrow{\varphi_{\text{MULTI}}} \emptyset$. This model, referred to as MULTI in our experiments, can in theory learn that a tonal pattern such as `HCLC` is more likely to occur before a word boundary than, say, the pattern `HCHC`.

Additionally, we evaluate an alternative backoff scheme also sensitive to tone patterns, φ_{LAST}, where only one single tone is remembered:

$$\begin{aligned} u_1 \ldots u_{n-1} &\xrightarrow{\varphi_{\text{LAST}}} \tau_1 u_2 \ldots u_{n-1} \xrightarrow{\varphi_{\text{LAST}}} \tau_2 u_3 \ldots u_{n-1} \\ &\xrightarrow{\varphi_{\text{LAST}}} \ldots \xrightarrow{\varphi_{\text{LAST}}} \tau_{n-2} u_{n-1} \xrightarrow{\varphi_{\text{LAST}}} \tau_{n-1} \xrightarrow{\varphi_{\text{LAST}}} \emptyset\,. \end{aligned} \tag{5.4}$$

This is illustrated on the same Mboshi example: `ámid` $\xrightarrow{\varphi_{\text{LAST}}}$ `Hmid` $\xrightarrow{\varphi_{\text{LAST}}}$ `Cid` $\xrightarrow{\varphi_{\text{LAST}}}$ `Ld` $\xrightarrow{\varphi_{\text{LAST}}}$ `C` $\xrightarrow{\varphi_{\text{LAST}}} \emptyset$. This model is referred to as LAST in our experiments. The rationale is to contrast MULTI with a less expressive, but also theoretically easier to estimate, tonal model. Instead of backing off to more complex tonal patterns, we want to see if gathering statistics corresponding to left contexts sharing a single tone (e.g. `ámid`, `émid`, and `ómid`) can help predict word boundaries.

Finally, we compare our tone models MULTI and LAST to the baseline PYP n-gram spelling model — referred to later on in our experiments as BASE (and simply as `pypshmm` in Section 3.3 of Chapter 3) — that is unable to distinguish between high and low tones. In this baseline, the backoff scheme φ_{BASE} is simply defined by:

$$\begin{aligned} u_1 \ldots u_{n-1} &\xrightarrow{\varphi_{\text{BASE}}} u_2 \ldots u_{n-1} \xrightarrow{\varphi_{\text{BASE}}} u_3 \ldots u_{n-1} \\ &\xrightarrow{\varphi_{\text{BASE}}} \ldots \xrightarrow{\varphi_{\text{BASE}}} u_{n-2} u_{n-1} \xrightarrow{\varphi_{\text{BASE}}} u_{n-1} \xrightarrow{\varphi_{\text{BASE}}} \emptyset\,. \end{aligned} \tag{5.5}$$

It corresponds on the previously used example to: `ámid` $\xrightarrow{\varphi_{\text{BASE}}}$ `mid` $\xrightarrow{\varphi_{\text{BASE}}}$ `id` $\xrightarrow{\varphi_{\text{BASE}}}$ `d` $\xrightarrow{\varphi_{\text{BASE}}} \emptyset$.

5.4 Experiments and discussion

We first conduct unsupervised segmentation experiments with `dpseg` and `pypshmm` (BASE) varying data representations, and then study the performance of the proposed tonal models (MULTI and LAST). In all experiments, 4 runs are performed and all `pypshmm` spelling models variants use a 3-gram spelling model.

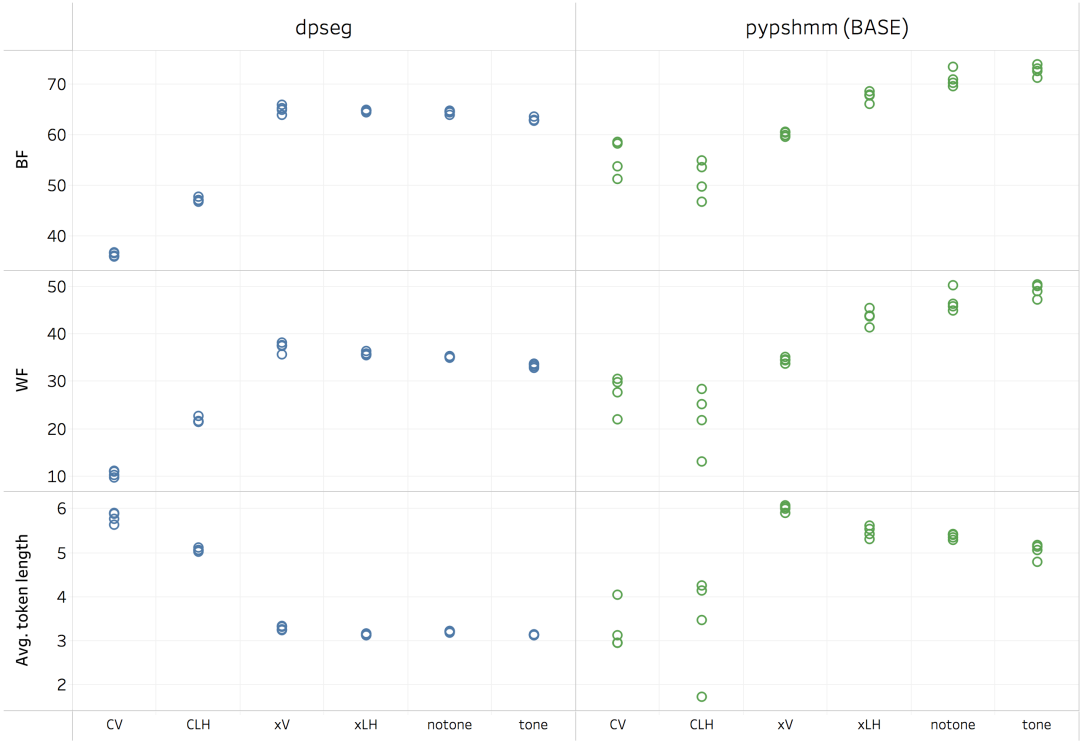


Figure 5.1: Boundary F-measure, token F-measure (BF and WF), and average token length for `dpseg` and `pypshmm` (BASE) on the Mboshi 5K corpus for representations `CV`, `CLH`, `xV`, `xLH`, `notone`, and `tone` (4 runs).

5.4.1 Representations

In Figure 5.1, we report boundary and token F-measures (BF and WF) obtained using `dpseg` (with a bigram LM) and `pypshmm` (BASE variety) on the Mboshi 5K corpus for representations `CV`, `CLH`, `xV`, `xLH`, `notone`, and `tone`. We omit type F-measures, as these would not be comparable between representations. Note, also, that we evaluate the entire corpus, as we did in Chapter 4 (but not in Chapter 3 where we compared various corpus sizes, and evaluated on the first 500 sentences). Lastly, we also strip tonal markers for the evaluation. Both `dpseg` and `pypshmm` are configured as described in Section 3.3.1.

As expected, the most impoverished representations `CV` and `CLH` yield poor segmentation results, both with `dpseg` and `pypshmm`. `dpseg` seems to capture the tonal signal better, with a clearer contrast between `CV` and `CLH`, while `pypshmm` exhibits a lot of variability in the results.

Representations `xV` and `xLH`, however, are much more competitive, and even on par with `notone` and `tone` in the case of `dpseg`, which is consistent with the results observed in the supervised experiment (Table 5.2). This time, the contrast between non-tonal `xV` and tonal `xLH` is clearer with `pypshmm`, favoring the tonal representation. Notwithstanding the replacement of 14 symbols (Mboshi has a 7 vowels inventory, each likely to carry a low or high tone) by only 2 symbols encoding the tonal information in

xLH, the segmentation performance is still very high. This is of particular interest for future experiments on true speech input, where coarse grain pseudo-phones or pseudo-syllable units could be extracted.

It seems harder to conclude with the results obtained with representations `notone` and `tone` (see also Figure 3.8 in Chapter 3), even though `pypshmm` seems to produce slightly better segmentations with `tone`. As observed before, `pypshmm` outperforms the bigram version of `dpseg`, but at the cost of less stable results.

5.4.2 Tonal modeling

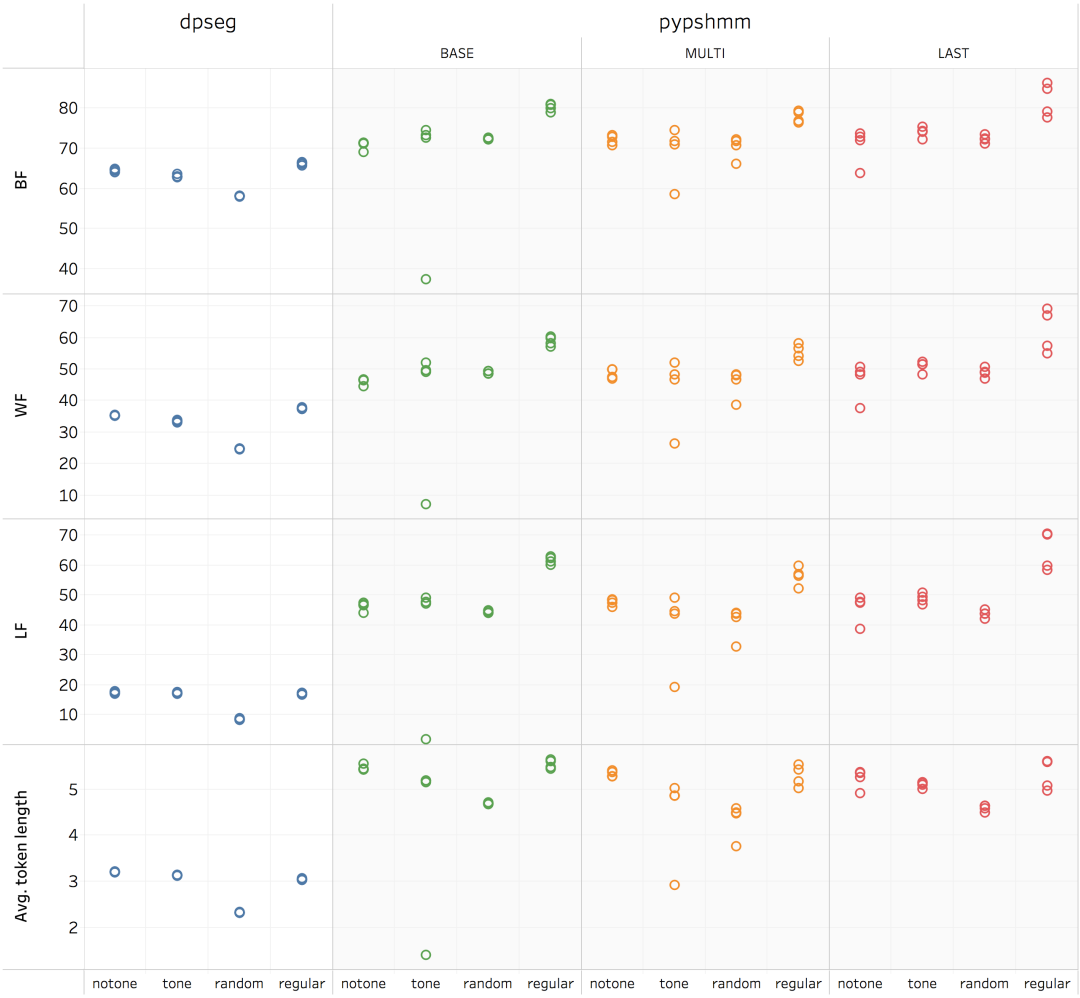


Figure 5.2: Boundary, token, type F-measure (BF, WF and LF), and average token length, on Mboshi 5K for `dpseg` and `pypshmm` BASE, and its tonal extensions MULTI and LAST. We compare representations `notone`, `tone`, `random`, and `regular` (4 runs).

We now turn, in Figure 5.2, to `pypshmm`'s tonal extensions proposed in Section 5.3. Models MULTI and LAST are compared to `pypshmm` baseline BASE (and to `dpseg`) on

representations `notone`, `tone`, `random`, and `regular` described in Section 5.2.1. We complement boundary and token F-measure (BF and WF) with type F-measure (LF), and evaluate as in Section 5.4.1. This time, `pypshmm` is trained for twice longer (48 hours instead of 24 hours) to try to alleviate the high variance observed in previous experiments on representations, yet without much success: the variance in the 4 runs for representation `tone` with BASE and MULTI, for instance, is particularly high.

If model MULTI seems hardly convincing, compared to BASE, and even increases variance in the results, probably due to a higher number of parameters at constant data size, model LAST seems to be able to capture some tonal information, especially on the synthetic `regular` representation. The token F-measure (WF) averaged over the 4 runs (62.16%) is indeed more than 3 points higher than the one obtained with BASE (58.82%). On the more realistic representation `tone`, however, the benefit of using model LAST is harder to see (and computing averaged F-measures would be misleading since BASE model produced a very distant outlier in the 4 runs).

A more detailed analysis of our segmentation results (in particular, contrasting precision and recall), and a qualitative examination by an expert linguist of the segmented utterances, failed to demonstrated a benefit of using models MULTI and LAST over model BASE on representation `tone`. All this seems to indicate that, without a very strong correlation between the tonal signal and the presence of word boundaries (which we artificially created with representation `regular`), the increased difficulty in estimating the model's parameters (even with model LAST) overweighs the potential usefulness of modeling tones in our setup.

For all models, including `dpseg`, representation `random` proves to be harder to segment, as we expected and already observed in the preliminary supervised experiment (Table 5.2). More surprisingly, `dpseg` fails to exploit tonal regularities in `regular`, and unlike BASE, MULTI, and LAST, does not produce better segmentations with this representation than with `tone`. This is likely due to `dpseg` having a 1-gram SM, preventing it to learn positional patterns inside a word and in particular the high-tone marker for a word-end in representation `regular`.

5.5 Conclusion

In a preliminary study, we showed that when learning a segmentation classifier on a simplified representation of a Mboshi corpus where all characters were collapsed to two 'vowel' and 'consonant' categories, supplying that classifier with tones provided an increase in performance and even led to a decent segmentation accuracy despite the considerable simplification of the data. This proved that segmentation can benefit from sensitivity to tonal cues, and we tried to leverage the latter in an unsupervised setting by introducing hierarchical n-gram spelling models that incorporate tone-conditional distributions in their hierarchy. These models were compared to a baseline Pitman-Yor n-gram spelling model and to a Dirichlet process-based bigram language model for several data representations.

One of our tonal models, LAST, seemed to have the capacity to capture tonal patterns on synthetic data, and to improve upon the baseline spelling model. Yet, with real tonal data, the newly proposed models were not able to take advantage of the tonal signal.

Beyond the limited supply (5K sentences) of data, this might be because the models we proposed cannot learn tonal regularities at the grammatical level, and are limited by design to learn lexically-based tonal regularities. Yet, tones in Mboshi, as in most Bantu languages, play as much a grammatical role as a lexical one. It would be interesting to test our models on languages with a mostly lexical tonology, such as Mandarin, Thai, or Vietnamese (Hyman, 2016).

In the absence of better performing tonal models, it seems wiser to ignore tonal information in a low-resource unsupervised word discovery task, thus reducing the inventory of characters: if we showed that tonal information has a potential to help segmentation, exploiting this signal without supervision is likely to require a larger quantity of data.

Chapter 6

Word Segmentation with Attention

Contents

So far, we have explored word discovery in the low-resource setting with various methods, and we have tried to improve our results making use of expert knowledge (Chapter 4) or modeling tonal cues (Chapter 5). The former strategy proved to be very successful, but the latter was less conclusive. In this chapter, we aim to take advantage of the supervision provided by the translation of the unwritten language (UL) in a well-resourced language (WL). This is also a condition to move towards automatic segmentation and alignment. During preliminary experiments (Chapter 3) we could not achieve improvements using a bilingual method, so we investigate a completely different approach here, based on artificial neural networks used in neural machine translation (NMT). Considering an encoder-decoder architecture with an attention mechanism, we use attention matrices as alignment matrices to induce word segmentation for the UL. This neural segmentation method introduced by (Zanon Boito et al., 2017) and (Godard et al., 2018d) was developed in collaboration with Marcely Zanon Boito, partly during the JSALT 2017 workshop at Carnegie Mellon University. We further investigate and extend it in this chapter. We achieve significant improvements in the precision of automatically discovered words, and obtain much better segmentation results than the

previously investigated bilingual method (`pisa`), without reaching the level of performance of the best monolingual Bayesian algorithms.

6.1 Introduction

In the low-resource scenario which motivates the work presented in this thesis, our intuition is that a translation of the UL into a WL will help in the word discovery task. Working with bilingual data is also key not only to discover words, but to align these words to known words in the WL. That is indeed the way linguists approach the documentation of a new language, creating interlinear glosses with the help of bilingual speakers. And that is also why such bilingual data is collected in the BULB project, according to the methodology described in Chapter 1. At the end of Chapter 3, however, we observed that taking advantage of the bilingual supervision at our disposal was not straightforward, and that the bilingual method that we experimented with - `pisa` (Stahlberg et al., 2012) - actually led to poorer segmentation results (see Section 3.3.2).

How can we make use of bilingual data to improve word discovery in a low-resource scenario? Does the increased complexity of a bilingual model undermine the – intuitively useful – signal such bilingual supervision should provide? Automatic word alignment on very small corpora is indeed known to be a difficult task (Pourdamghani et al., 2018). In this chapter, we want to leverage advances in machine translation, achieved in recent years by training artificial neural networks, and to understand if these models can produce useful segmentations results in a low-resource setting.

As we pointed out in Section 2.1.1.2, word alignments can be seen as binary matrices. A popular architecture used in NMT, the encoder-decoder with attention (see Section 6.2), is of particular interest to us as it involves attention matrices. Such matrices, as we will see in Section 6.3, can be seen as (soft) alignment matrices. Following the "align to segment" approach depicted in Figure 2.2b of Chapter 2, we first perform word discovery using attention matrices, expanding on the research started in (Zanon Boito et al., 2017) and (Godard et al., 2018d), and inspired by the work of Cohn et al. (2016) and Duong et al. (2016). The "align to segment" approach, however, is hindered by the fact that the content per fine-grained unit (a phoneme or a character here) is limited, making such units difficult to align to words. In the spirit of the joint models described in Section 2.6, we extend the baseline neural segmentation method by training our models with additional signal meant to improve segmentation. This signal takes the form of a word-length bias in the attention mechanism, or of an auxiliary loss constraining attention matrices.

In this chapter, we describe the encoder-decoder architecture with attention (Section 6.2), and our baseline word segmentation approach, together with the two proposed extensions (Section 6.3). We conduct and discuss experiments on the Mboshi 5K corpus in Section 6.4, and conclude in Section 6.5.

6.2 Encoder-decoder with attention

Statistical machine translation (Koehn, 2010), and the phrase-based approach in particular, was delivering state-of-the-art results in machine translation until 2015 (Koehn,

2017). Techniques involving artificial neural networks had then already been applied successfully for re-ranking or re-ordering (Le et al., 2012; Devlin et al., 2014), but pure[1] neural machine translation systems only started to produce reasonable results after the introduction of sequence-to-sequence models with RNN encoder and decoder (Sutskever et al., 2014; Cho et al., 2014a,b). These models performed poorly when translating long sentences though, a problem subsequently addressed by the introduction of an attention mechanism by Bahdanau et al. (2014), which allowed NMT to achieve state-of-the-art translation performance. More recently, other architectures – such as the Transformer model of Vaswani et al. (2017) or the convolutional model of Gehring et al. (2017) – have further improved these performance. In our context, however, Bahdanau's attention mechanism is of particular interest as it produces matrices that can be interpreted as soft alignment matrices.[2] In this section, we succinctly describe the standard encoder-decoder architecture used in sequence-to-sequence modeling, and the attention mechanism that we will then use to induce word segmentations.

6.2.1 RNN encoder-decoder

Sequence-to-sequence models are designed to transform a variable-length input, or *source*, sequence into a variable-length *target* output sequence. The source sequence is typically a sentence, i.e. a sequence of words $w_1, \dots, w_J$, and the target sequence is also a sequence of words $\omega_1, \dots, \omega_I$.[3] In our context, though, the target sequence will be a sequence of characters or phonemes.

In the RNN Encoder-Decoder architecture introduced by Sutskever et al. (2014) and Cho et al. (2014b), an *encoder* consisting of a recurrent neural network (RNN) – with one or several layers – reads a source sequence of word embeddings $e(w_1), \dots, e(w_J)$ representing the input sentence, and produces a dense representation c of this sentence. Vector c is then fed to an RNN *decoder* producing the output translation $\omega_1, \dots, \omega_I$ sequentially, much like an RNN language model (Mikolov et al., 2010).

Encoder Denoting the encoder's hidden states for each step of the input sequence as h_j, such hidden states are computed as:

$$h_j = \phi(e(w_j), h_{j-1})\,. \tag{6.1}$$

In most cases, ϕ corresponds to a long short-term memory (LSTM, see Hochreiter and Schmidhuber, 1997) unit or a gated recurrent unit (GRU, see Cho et al., 2014b), and h_J is used as the fixed-length context vector c initializing the RNN decoder. Figure 6.1 depicts such an encoder.

Decoder On the target side, the decoder predicts each word ω_i, given the context vector c (in the simplest case, h_J, the last hidden state of the encoder) and the previously

[1] i.e. not relying on any phrase-based SMT component.

[2] The Transformer architecture can also produce soft alignments, but several heads or layers will lead to different attention matrices, which might be hard to interpret or use, a problem recently addressed by Alkhouli et al. (2018).

[3] Note that we index the source with j and the target with i here, departing from our own conventions, in order to maintain more consistency with Bahdanau's notations.

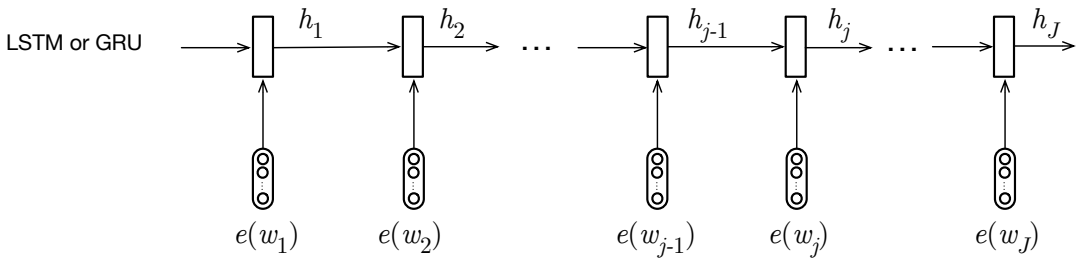


Figure 6.1: An RNN encoder with LSTM or GRU cells. $e(w_j)$ corresponds to the embedding of word w_j, and h_j is the hidden state of the encoder at step j.

predicted words, using the probability distribution over the output vocabulary V_T:

$$\begin{cases} P(\omega \,|\, \omega_1, \ldots, \omega_{i-1}, c) = g(\omega_{i-1}, s_i, c) \\ \omega_i = \text{argmax}_{\omega_k} P(\omega = \omega_k \,|\, \omega_1, \ldots, \omega_{i-1}, c)\,, \end{cases} \tag{6.2}$$

where s_i is the hidden state of the decoder RNN, and g a nonlinear function (e.g. a multi-layer perceptron with a softmax layer, see Figure 6.2) computed by the output layer of the decoder. The hidden state s_i of the decoder is, in turn, updated according to:

$$s_i = f(s_{i-1}, \omega_{i-1}, c)\,, \tag{6.3}$$

where f corresponds to the function computed by an LSTM or GRU cell.

Training The encoder and the decoder are trained jointly to maximize the probability of the true target translation $\mathbf{\Omega} = \Omega_1, \ldots, \Omega_I$ given the source sentence $\mathbf{w} = w_1, \ldots, w_J$ under the model. At each decoding step i, and for each sentence pair, a probability distribution $P(\omega \,|\, \omega_1, \ldots, \omega_{i-1}, c)$ over the output vocabulary V_T is computed (see Equation (6.2)). Denoting

$$p_{\Omega_i} = P(\omega = \Omega_i \,|\, \omega_1, \ldots, \omega_{i-1}, c)\,, \tag{6.4}$$

the negative log-likelihood loss (NLL, also known as the cross-entropy loss) for this sentence pair is defined by:

$$\mathcal{L}_{\text{NLL}}(\mathbf{\Omega} \,|\, \mathbf{w}) = -\sum_{i=1}^{I} \log(p_{\Omega_i})\,, \tag{6.5}$$

and minimizing this loss function is equivalent to maximizing $P(\mathbf{\Omega} \,|\, \mathbf{w})$ with respect to the parameters of the model.

As reference target words are available during training, Ω_i (and the corresponding word embedding) can be used instead of ω_i in Equations (6.2) and (6.3) (see also Figure 6.2). This training technique is known as *teacher forcing*[4] (Williams and Zipser, 1989).

[4] As we will see in Section 6.3.1, teacher forcing will also be used at "test" time in our scenario, since we are not training these models for a translation task, but a word segmentation task.

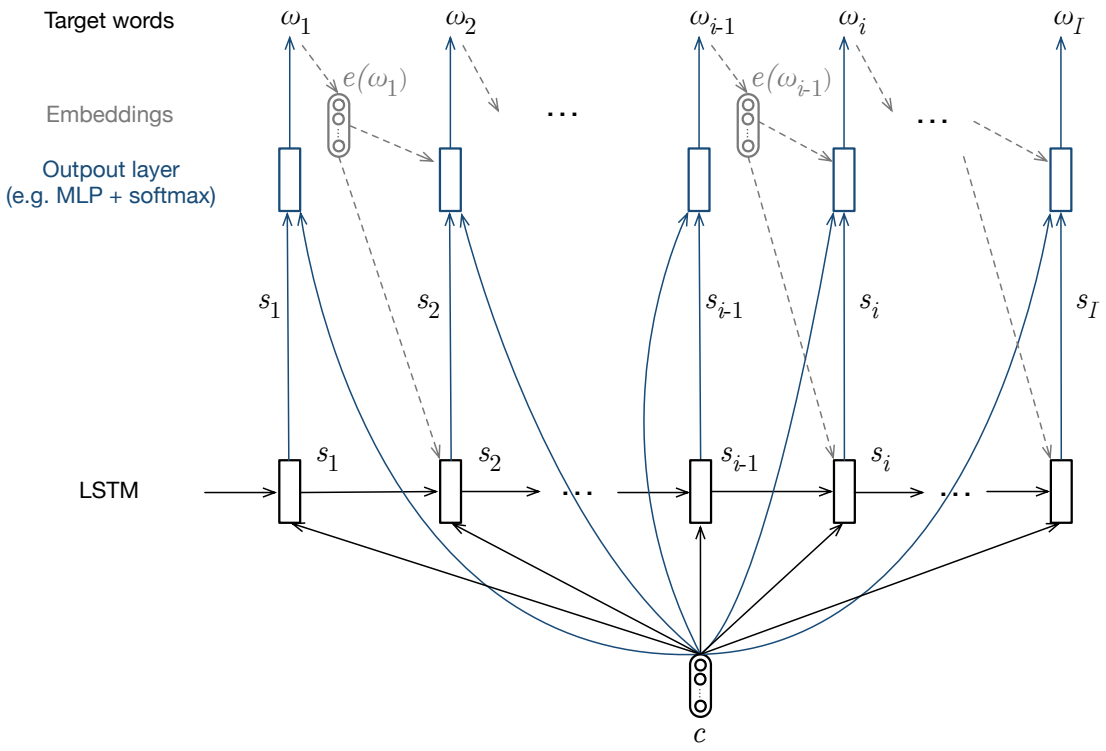


Figure 6.2: A single-layer RNN decoder with LSTM units, and an output layer computing a probability distribution over the output vocabulary. $e(\omega_i)$ corresponds to the embedding of target word ω_i, and s_i is the hidden state of the decoder at step i. The context vector c is, in the simplest case, the last hidden state of the encoder.

6.2.2 The attention mechanism

Bahdanau's attention Encoding a sentence of variable length in a fixed-length vector can lead to poor translation results with long sentences (Cho et al., 2014a).[5] To address this problem, Bahdanau et al. (2014) introduce a mechanism allowing the decoder to *attend* at specific parts of the input sentence's representation, with the hope that the decoder will predict each target word using only relevant parts of the source.

This is achieved by computing a distinct context vector (a position-dependent aggregated representation of the source) for each time step of the decoding, updating s_i and predicting a new target word ω_i according to:

$$\begin{cases} s_i = f(s_{i-1}, \omega_{i-1}, c_i) \\ P(\omega \,|\, \omega_1, \ldots, \omega_{i-1}, c_i) = g(\omega_{i-1}, s_i, c_i) \\ \omega_i = \text{argmax}_{\omega_k} \, P(\omega = \omega_k \,|\, \omega_1, \ldots, \omega_{i-1}, c_i) \,. \end{cases} \tag{6.6}$$

In the first equation, f corresponds to an LSTM or GRU cell, as in Equation (6.3). Each context vector is defined as a weighted sum of the encoder's hidden states h_j:

[5] Sutskever et al. (2014) partly mitigate this issue by reversing source sentences.

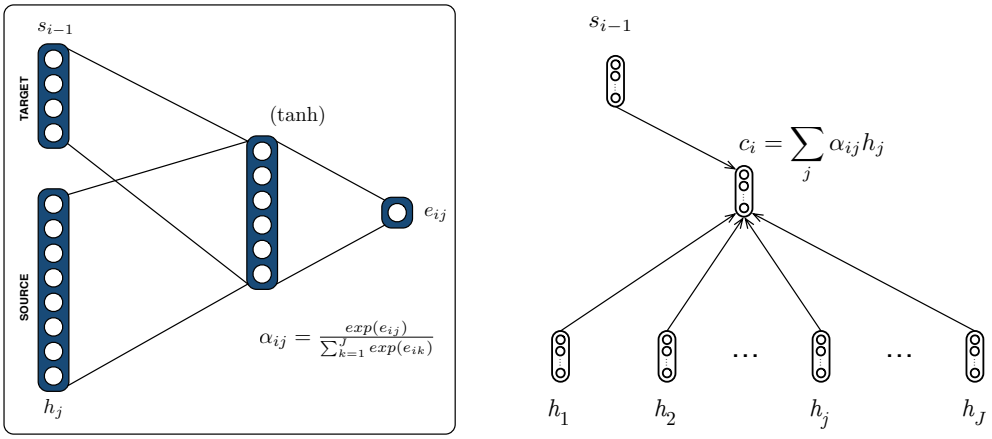


Figure 6.3: Bahdanau's attention mechanism.

$$c_i = \sum_{j=1}^{J} \alpha_{ij} h_j \,, \tag{6.7}$$

where weights α_{ij} are produced by an *alignment model* consisting in a multi-layer perceptron (MLP) followed by a softmax layer (Figure 6.3). If we denote by a the function computed by the MLP, then

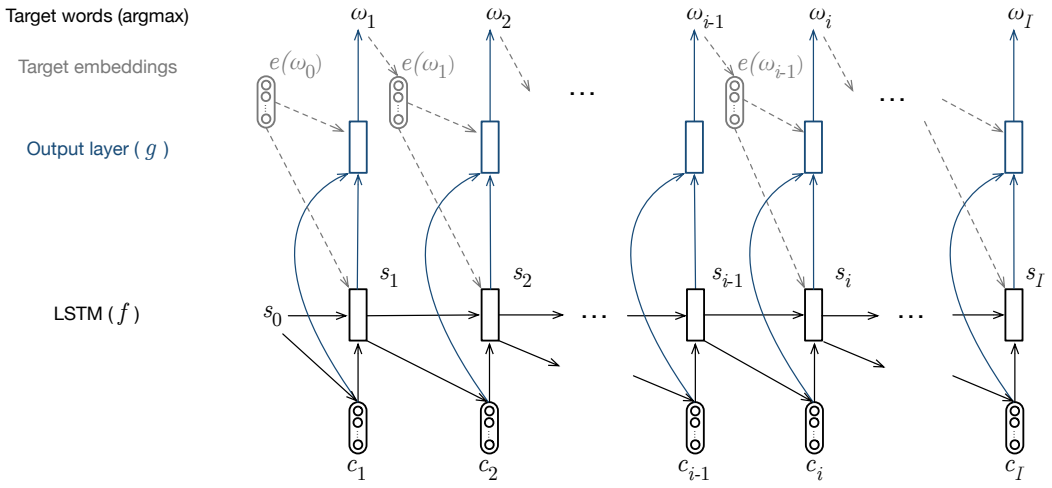
$$\begin{cases} e_{ij} & = a(s_{i-1}, h_j) \\ \alpha_{ij} & = \frac{exp(e_{ij})}{\sum_{k=1}^{J} exp(e_{ik})} \,, \end{cases} \tag{6.8}$$

where e_{ij} is known as the *energy* associated to α_{ij}. Lines in the attention matrix $A = (\alpha_{ij})$ sum to 1, and weights α_{ij} can be interpreted as the probability that target word ω_i is aligned to source word w_j. (Bahdanau et al., 2014) indeed investigated qualitatively such soft-alignments and concluded that their model can correctly align target words to relevant source words. Our segmentation method (Section 6.3) relies on the assumption that the same holds when aligning characters or phonemes on the target side to source words. Additionally, the investigation of (Ghader and Monz, 2017) supports, to a certain degree, the agreement between attention and word alignment.[6] Lastly, the context vector c_i can be interpreted as a summary of the useful source information at decoding step i.

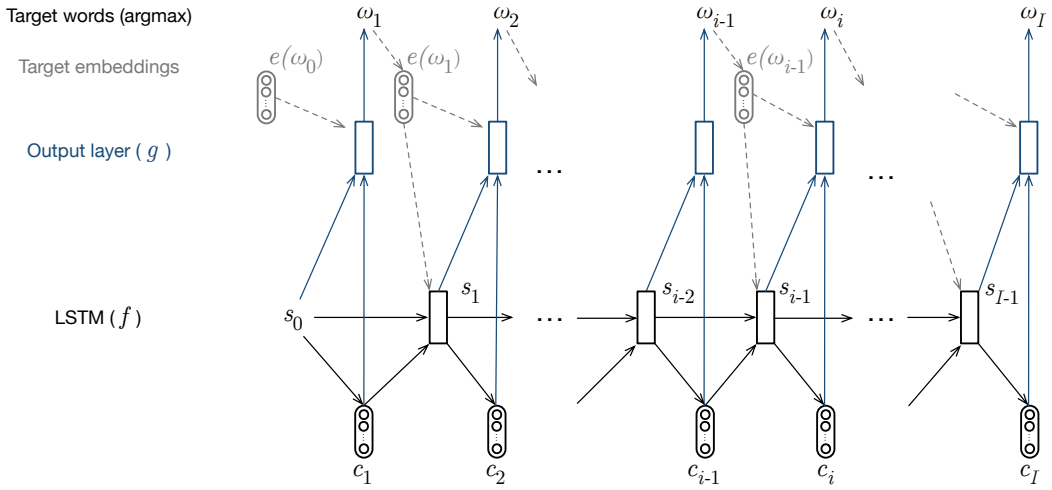
Update or generate first? According to Equation (6.6), each decoding step involves first an update of the hidden state s_i of the decoder's RNN, and then a prediction for the current target word ω_i (Figure 6.4a).

However, Peter et al. (2017) and other researchers interpret the decoding equations in (Bahdanau et al., 2014) in the reversed order: generating the current target word

[6] In particular, the authors show that this agreement can substantially vary depending on target POS tags and attention mechanisms. For instance, target-noun-related attention better correlates with alignments than is the case for verbs. Koehn and Knowles (2017) also discuss the relationship between attention and word alignment.



(a) Update first: at each time step i, RNN state s_i is updated with function f, and current target word ω_i is generated using output function g.



(b) Generate first: at each time step i, current target word ω_i is predicted with output function g, and RNN state s_i is updated using this prediction.

Figure 6.4: Update or generate first in the RNN decoder.

first, then updating the current RNN state. This leads to the following equations (see also Figure 6.4b):

$$\begin{cases} P(\omega_i \,|\, \omega_1, \ldots, \omega_{i-1}, c_i) = g(\omega_{i-1}, s_{i-1}, c_i) \\ s_i = f(s_{i-1}, \omega_i, c_i)\,. \end{cases} \tag{6.9}$$

Note this is not a mere indexation choice: during training (or forced decoding), the context vector c_i is computed with ground-truth target word Ω_{i-2} (through s_{i-1} when updating first). But when we generate first, c_i is instead computed with Ω_{i-1}. As it seems sensible to choose Ω_{i-1} rather than Ω_{i-2} to predict ω_i, this could explain why generating first yield better results in our experiments (see Section 6.4.1 for other implementation choices). Updating first might conversely explain the one-position mismatch between attention and word alignment observed by Koehn and Knowles (2017).

Other attention mechanisms The attention mechanism introduced by Bahdanau et al. (2014) has been further explored by many researchers. Luong et al. (2015), for instance, compare a *global* to a *local* approach for attention, and examine several architectures to compute alignment weights α_{ij}. Yang et al. (2016) additionally propose a recurrent version of the attention mechanism, where a "dynamic memory" keeps track of the attention received by each source word, and demonstrate better translation results. We also briefly describe in Section 7.2.1 a more general formulation of the attention mechanism introduced by Kim et al. (2017), where structural dependencies between source units can be modeled.

With the goal of improving alignment quality, Mi et al. (2016) calculate a distance between attentions and word alignments learnt with the reparameterization of IBM Model 2 from Dyer et al. (2013) (see Section 2.5.2); this distance is then added to the cost function during training. To improve alignments also, Cohn et al. (2016) introduce several refinements to the attention mechanism, in the form of structural biases common in word-based alignment models reviewed in Section 2.5. In this work, the attention model is enriched with features able to control positional bias, fertility, or symmetry in the alignments, which leads to better translations for some language pairs, under low-resource conditions. More work seeking to improve alignment and translation quality can be found in (Tu et al., 2016; Liu et al., 2016; Sankaran et al., 2016; Feng et al., 2016; Kuang et al., 2017; Alkhouli and Ney, 2017).

6.3 Attention-based word segmentation

Our goal is to discover words in an unsegmented stream of target characters. In this section, we describe a baseline segmentation method inspired by the "align to segment" strategy alluded to before (Section 2.6). We then propose two extensions providing the model with a signal relevant to the segmentation process, so as to move towards a joint learning of segmentation and alignment.

6.3.1 Align to segment

Baseline methodology As we just saw, an attention matrix $A = (\alpha_{ij})$ can be interpreted as a soft alignment matrix between target and source units. Since $\sum_{j=1}^{J} \alpha_{ij} = 1 \ \forall i$, and $0 \leq \alpha_{ij} \leq 1 \ \forall i, j$, each line of the matrix can be seen as a probability distribution where α_{ij} is the probability for target word ω_i to be aligned to source word w_j.

In our context, where words need to be discovered in the UL, we consider instead a sequence of characters (or phonemes) π_i on the target side. In the spirit of the "align to segment" approach described in Section 2.1.1 (Figure 2.2b) and following (Zanon Boito et al., 2017) and (Godard et al., 2018d),[7] we perform word segmentation on the target side with the following steps:

1. We train an RNN encoder-decoder model with attention (see Section 6.2.2) on a corpus where each sentence pair is composed of a source sequence of words and a target sequence of characters. Training is performed with teacher forcing (see Section 6.2.1).
2. After training, we force decode the entire corpus and we extract one attention matrix for each sentence pair. Forced decoding means that we do not make use of partial hypotheses ω_i produced by the trained model, but instead reinject ground-truth words Ω_i at each decoding step i. This is because we are not interested in producing new target sentences, as would be the case in a standard translation setting.[8]
3. We segment target sequences via a simple post-processing of the extracted attention matrices. For each target unit π_i of the UL, we identify the source word w_{a_i} to which it is most likely to be aligned. That is, $\forall i, a_i = \text{argmax}_j \, \alpha_{ij}$. Given these alignment links a_i from target units to source words, we deduce a word segmentation on the target side: when two consecutive target units are not aligned with the same source word, we introduce a word boundary in the target.

We now review two existing variants to this method.

Reversed architecture It is possible to devise an equivalent segmentation scheme by training from UL units to WL words, as experimented in (Zanon Boito et al., 2017). However, this implicitly defines alignment probabilities for each target word over source characters (or phonemes), instead of alignment probabilities for each target character (or phoneme) over source words. As the columns in attention matrices are not guaranteed to sum to 1, the post-processing step becomes less principled. Zanon Boito et al. (2017) also reported poorer segmentation performance when training this way, from UL characters to WL words.

[7] The former of these two works was applied to graphemes, and the latter to automatically discovered pseudo-phones.

[8] For the same reason, we do not use a development set during training, and train our models on the entire data, see Section 6.4.1.

Smoothing Considering a (simulated) low-resource setting, and building on Cohn et al. (2016)'s work, Duong et al. (2016) propose to smooth attentional alignments, either by post-processing attention matrices, or by flattening the softmax function in the attention model (see Equation (6.8)) with a temperature parameter T:

$$\alpha_{ij} = \frac{exp(e_{ij}/T)}{\sum_{k=1}^{J} exp(e_{ik})/T} \,. \tag{6.10}$$

This makes sense as the authors examine attentional alignments obtained while training from UL phonemes to WL words. But when translating from WL words to UL characters, this seems less useful: smoothing will encourage a character to align to many words.[9] This technique is further explored by Lin et al. (2018), who make the temperature parameter trainable and specific to each decoding step, so that the model can learn how to control the softness or sharpness of attention distributions, depending on the current word being decoded.

6.3.2 Extensions: towards joint alignment and segmentation

One important limitation in this segmentation approach, and its variants, lies in the absence of signal relative to segmentation during training. We want to move towards jointly learning alignment and segmentation, so we introduce two extensions aiming at guiding the training of our models with constraints derived from the particular segmentation heuristic we use.

Word-length bias When we sketched the minimum description length (MDL) principle in Section 2.2.3, we introduced the idea that the optimal code length of a word is inversely proportional to its probability. This information theoretic way of explaining efficiency in language communication has been empirically supported in natural languages, where frequent words tend to be short. This is known as Zipf's "Law of Abbreviation", a language universal. More recently, Piantadosi et al. (2011) have shown across ten different languages that *average information content*[10] is actually a better predictor than frequency for word length. Experimental psychology also correlates word length and conceptual complexity (see Lewis and Frank, 2016). Kanwal et al. (2017) have also experimentally shown that speakers optimise the mapping between form and meaning under the concurrent pressures of accuracy and efficiency.

The universal correlation between meaning or frequency on the one hand, and word length on the other hand, leads us to make the assumption that the length of aligned source and target words should also correlate. Being in a relationship of mutual translation, we can expect them to have comparable frequencies and meaning, hence comparable lengths. For us, this means that the longer a source word is, the more target units should be aligned to it. We attempt to implement this idea in the attention mechanism as a word-length bias consisting in changing the computation of the context vector from

[9] A temperature below 1 would conversely sharpen the alignment distribution. We did not observe significant changes in segmentation performance varying the temperature parameter.

[10] The authors define this measure as $-\sum_c P(C = c \,|\, W = w) \log P(W = w \,|\, C = c)$, where W and C are random variables corresponding to a word and its "context".

Equation (6.7) to:

$$c_i = \sum_j \psi(|w_j|)\, \alpha_{ij}\, h_j \tag{6.11}$$

where ψ is a monotonically increasing function of the length $|w_j|$ of word w_j. This way, we will encourage target units to attend more to longer word. Given the segmentation method that we use during post-processing, this is a necessary condition for producing target segments whose lengths are correlated to source lengths.

In practice, we just choose ψ to be the identity function, and we renormalize so as to preserve the property of lines summing to 1 in the new attention matrices. Consequently, the context vectors c_i are now computed with attention weights $\tilde{\alpha}_{ij}$:

$$\begin{cases} \tilde{\alpha}_{ij} & = \frac{|w_j|}{\sum_j |w_j|\, \alpha_{ij}}\, \alpha_{ij} \\ c_i & = \sum_j \tilde{\alpha}_{ij}\, h_j\,. \end{cases} \tag{6.12}$$

We finally segment the target from attention matrices $A = (\tilde{\alpha}_{ij})$, with the method described in Section 6.3.1.

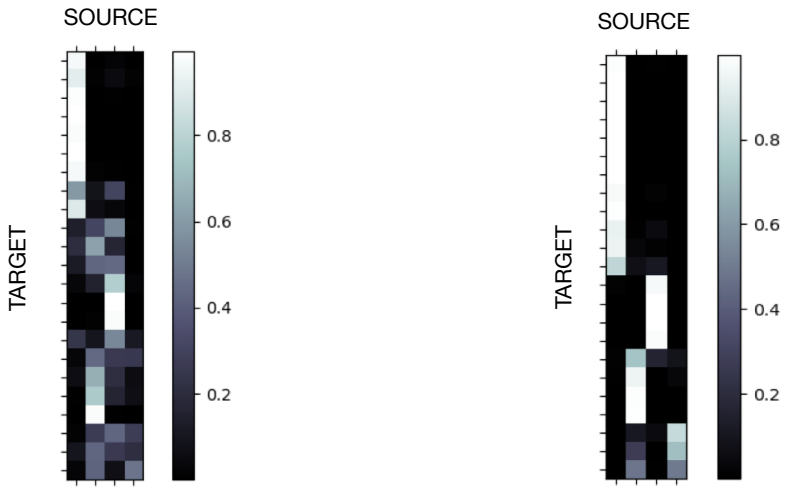
Auxiliary loss Another way to inject segmentation-awareness inside our training procedure is to try to control the number of target words that will be produced by the post-processing of attention matrices. Even if typological discrepancies (see Section 2.1.3) between source and target languages will typically lead to a different average number of words on the source and target sides, we assume that guiding the target segmentation so as to produce a number of words close to the one found on the source side could lead to conceptually sound segmentations.[11]

We attempt to do this by complementing our main loss function $\mathcal{L}_{\text{NLL}}$ (see Section 6.2.1) with an auxiliary loss $\mathcal{L}_{\text{AUX}}$ defined as:

$$\mathcal{L}_{\text{AUX}}(\mathbf{\Omega} \,|\, \mathbf{w}) = |I - J - \sum_{i=1}^{I-1} \alpha_{i,*}^{\top} \alpha_{i+1,*}| \tag{6.13}$$

The rationale behind this auxiliary loss can be better understood recalling our segmentation heuristic: for each target unit π_i, we identify the source word w_{a_i} to which it is most likely to be aligned ($\forall i, a_i = \text{argmax}_j\, \alpha_{ij}$, see Section 6.3.1); a boundary is then inserted on the target side when two consecutive target units are not aligned to the same source word. Intuitively, the dot product between consecutive lines in the attention matrix will be closer to 1 if the consecutive target units are aligned to the same source word, and closer to 0 if they are not. The sum term aims at quantifying the number of target units that will *not* be followed by a word boundary after segmentation. Therefore $I - \sum_{i=1}^{I-1} \alpha_{i,*}^{\top} \alpha_{i+1,*}$ should quantify the number of word boundaries that will be produced on the target side. And the auxiliary loss should guide the model towards learning attention matrices resulting in target segmentations having the same number of words as the source.

[11] Ideally, we would only correlate (instead of matching) to the number of words found on the source side, as both languages might be typologically distant. In order to address this, we explore a light supervision scheme in our experiments (see method AUX+RATIO in Section 6.4).



(a) without auxiliary loss

(b) with auxiliary loss

Figure 6.5: Effect of the proposed $\mathcal{L}_{\text{NLL}}$ auxiliary loss on an example attention matrix for a sentence pair. Lines are indexed by target characters (or phonemes) and columns, by source words.

Figure 6.5 illustrates the effect of our auxiliary loss on an example sentence pair. Without auxiliary loss, the segmentation will yield, in this particular case, 8 target segments (Figure 6.5a), while the attention learnt with auxiliary loss will yield 5 target segments (Figure 6.5b); the source sentence, on the other hand, has 4 tokens.[12]

6.4 Experiments and discussion

In this section, we describe implementation details for our baseline segmentation system and for the extensions proposed in Section 6.3.2, before presenting data and results.

6.4.1 Implementation details

Many iterations led us to fix what we consider a reasonable baseline system. Some preliminary experiments (not reported here) where conducted with the LIG-CRIStAL NMT system,[13] while others used the XNMT toolkit (Neubig et al., 2018),[14] but we ultimately re-implemented Bahdanau's encoder-decoder with attention in PyTorch (Paszke et al., 2017)[15] in order to integrate the extensions described in Section 6.3.2. The last version of our code, which handles mini-batches efficiently, heavily borrows from Joost Basting's code.[16] Source sentences include an end-of-sentence (EOS) special symbol

[12] We count here the end-of-sentence token corresponding to the last column in the attention matrices.

[13] https://github.com/eske/seq2seq.

[14] https://github.com/neulab/xnmt.

[15] https://pytorch.org/. We use version 0.4.1.

[16] https://github.com/bastings/annotated_encoder_decoder.

(corresponding to w_J in our notation) and target sentences include both a beginning-of-sentence (BOS) and an EOS symbol (respectively ω_0 or Ω_0, and ω_J or Ω_J in our notation). Padding of source and target sentences in mini-batches is required, as well as masking in the attention matrices and during loss computation.

Overall, we are trying to keep the architecture a simple as possible. NMT models require the training of many parameters, and with our very limited data, we know that this is a challenging task.[17] Our implementation follows (Bahdanau et al., 2014) very closely, with some minor changes.

Encoder As in (Bahdanau et al., 2014), we use a single-layer bidirectional RNN (Schuster and Paliwal, 1997) with GRU cells: GRU-based RNNs have been shown to perform similarly to LSTM-based RNNs (Chung et al., 2014), while computationally more efficient. We set a dimension of 64 for the hidden states of the forward and backward RNNs, the same as the one chosen in (Zanon Boito et al., 2017; Godard et al., 2018d) which seems a good compromise with our data size for the segmentation task. Word embeddings also have a dimension of 64. In Equation (6.1), h_j corresponds to the concatenation of the forward and backward states for each step j of the source sequence (h_j's dimension is 128).[18]

Attention The alignment MLP model computes function a from Equation (6.8) as $a(s_{i-1}, h_j) = v_a^\top \tanh(W_a s_{i-1} + U_a h_j)$ – see Appendix A.1.2 in (Bahdanau et al., 2014) – where v_a, W_a, and U_a are weight matrices. Note that there is no bias in the MLP, unlike what was done in (Zanon Boito et al., 2017).

For the computation of weights $\tilde{\alpha_{ij}}$ in the word-length bias extension (Equation (6.12)), we arbitrarily attribute a length of 1 to the end-of-sentence symbol on the source side.

Decoder The decoder's RNN is initialized via an encoder *bridge*, i.e. using the last backward state of the encoder ($\overleftarrow{h_1}$ in Bahdanau's notation) and a non-linear function (tanh) for state s_0. Some authors concatenate the last forward state $\overrightarrow{h_J}$ to $\overleftarrow{h_1}$, but we did not observe any significant difference doing so. We use a single-layer GRU RNN with hidden size 64 (as in the encoder), and output embeddings of dimension 64.

In preliminary experiments, we observed better segmentation results adopting the "generate first" approach during decoding (see Section 6.2.2). During training and forced decoding, the hidden state s_i is therefore updated using ground-truth embeddings $e(\Omega_i)$. Ω_0 is the BOS symbol.

Our implementation of the output layer (g in Equation (6.9)) consists of a MLP with a softmax layer: a linear projection (with weight biases, and to a dimension of 128) of Ω_{i-1}, s_{i-1}, and c_i, is projected after a non-linearity (tanh) to the output vocabulary dimension, and a softmax is computed.

Training We train our models for 800 epochs on the whole corpus with the Adam algorithm (the learning rate is set to 0.001). Parameters are updated after each

[17] We indeed posit that the disappointing results obtained with `pisa` stem essentially from the difficulty to estimate the parameters of that model.

[18] Initial states for both directions are kept to PyTorch's default initialization to 0.

mini-batch of 64 sentence pairs.[19] We do not use any split during training, as controlling convergence with a development set and early stopping led to weaker segmentation results.[20]

A dropout layer (Srivastava et al., 2014) is applied to both source and target embedding layers, with a rate of 0.5, as in (Godard et al., 2018d). We also tried to add a dropout layer after the encoder and decoder RNNs,[21] but this harmed our segmentation results.

The weights in all linear layers are initialized with Glorot's normalized method (Equation (16) in Glorot and Bengio, 2010), and bias vectors are initialized to 0. Embeddings are initialized with the normal distribution $\mathcal{N}(0, 0.1)$.[22] Except for the particular situation of the bridge between the encoder and the decoder, the initialization of RNN's weights is kept to PyTorch's defaults.

During training, we minimize the NLL loss $\mathcal{L}_{\text{NLL}}$ (see Section 6.2.1), adding optionally the auxiliary loss $\mathcal{L}_{\text{AUX}}$ (Section 6.3.2). When the auxiliary loss is used, we schedule it to be integrated progressively so as to avoid degenerate solutions[23] with coefficient $\lambda_{\text{AUX}}(k)$ at epoch k defined by:

$$\lambda_{\text{AUX}}(k) = \frac{\max(k - W)}{K} \tag{6.14}$$

where K is the total number of epochs and W a *wait* parameter. The total minimized loss at epoch k is then given by $\mathcal{L}_{\text{NLL}} + \lambda_{\text{AUX}} \cdot \mathcal{L}_{\text{AUX}}$. In practice, and after trying values ranging from 100 to 700, we set the wait parameter W to 200 in our experiments. We approximate the absolute value in Equation (6.13) by $|x| \triangleq \sqrt{x^2 + 0.001}$, in order to make the auxiliary loss function differentiable.

6.4.2 Data and evaluation

In the experiments presented in this chapter, we use the bilingual French-Mboshi 5K. The description of the corpus can be found in Section 3.2, and statistics are in Table 3.1. On the Mboshi side, we consider the representation `notone` (no tone). On the French side, and in addition to granularities `word` (the standard tokenization), `lemma`, `morph`, and `pos` (Section 3.2.2), we consider two new representations: `poslen`, where the POS tag of each French word is suffixed by the corresponding word length (e.g. "`ADJ_5`"), and `length`, where each word is replaced by a token `WORD` suffixed by the word length (e.g. "`WORD_4`").

[19] Mini-batches are created anew through shuffling and length-sorting at each epoch.

[20] This might be because we do not need to generalize, and to prevent overfit, unlike with a standard translation task subsequently evaluated on a test set. Note that without dropout, however, we do not learn any useful alignments in the attention matrices. Thus, dropout seems to provide enough regularization for our task.

[21] Only for the "output" state, not the state values used inside the recursion.

[22] This seemed to slightly improve segmentation results when compared to Glorot's normalized method

[23] When the NLL loss has not "shaped" yet the attention matrices into soft alignments, the auxiliary loss can lead to trivial optimization solutions, in which a single column in the attention matrices has a certain number of weights set to 1 (to reach the proper value in the sum term from Equation (6.13)), while all other weights in the matrices are zeroed. The model is subsequently unable to escape this solution.

We denote the baseline segmentation system as BASE, the word-length bias extension as BIAS, and the auxiliary loss extensions as AUX. We also report results for a variant of AUX we call AUX+RATIO, in which the auxiliary loss is computed with a factor corresponding to the true ratio $r_{\text{MB/FR}}$ between the number of words in Mboshi and in French averaged over the first 100 sentences[24] of the corpus. In this variant, the auxiliary loss (Equation (6.13)) is computed as $|I - r_{\text{MB/FR}} \cdot J - \sum_{i=1}^{I-1} \alpha_{i,*}^{\top} \alpha_{i+1,*}|$.

As before in this thesis, we evaluate word segmentation with precision, recall, and F-measure on boundaries (BF, BR, BP), tokens (WF, WR, WP), and types (LF, LR, LP) – more details can be found in Section 2.1.2. In Figure 6.6, we present segmentation results with different granularities on the source (French) side obtained with the baseline method (BASE). Two runs have been performed for each representation.

In Figure 6.7, boundary, token, and type metrics are complemented with the sentence exact-match (X),[25] and we report segmentation results for all methods with the `word` representation on the French side. This time we perform 10 runs in order to measure variability in our results.

6.4.3 Discussion

We first examine the results obtained with our neural segmentation baseline BASE. We then analyse the impact of the word-length bias and the auxiliary loss described in Section 6.3.2.

Baseline results Results in Figure 6.6 are comparable[26] to the Mboshi results shown in Figure 3.2 and Figure 3.3 from Chapter 3. Results from Chapter 4 and Chapter 5 are also comparable, but not considered in this chapter as they rely on additional information (expert knowledge or tones). The first observation we can make is that our BASE neural segmentation approach does not perform as well as the best monolingual Bayesian methods we studied. Comparing average values over 2 runs, boundary F-measure (BF) on `word` is about 7 points higher with `dpseg` (62.24%, to compare to 55.62% with BASE), and more than 10 points higher with `pypshmm`. Similar observations can be made for token metrics. For type metrics, our system is on par with `dpseg` (a result already reported in (Zanon Boito et al., 2017)), but `pypshmm` performs vastly better (46.69% vs. 30.05% in BF).

So why insist, and try to improve this neural segmentation method? The main reason is that discovering words alone is seldom useful for language documentation. Linguists need to align discovered units to known words in order to access their semantic content, build a bilingual dictionary, further investigate linguistic phenomena, etc. Only a bilingual approach can automatically provide such alignments. The bilingual method tested in Chapter 3, `pisa`, however, led to very low results in boundary, token, and type F-measure (Figure 3.3) when compared to monolingual methods. We improve `pisa`'s

[24] This provides a plausible supervision in the CLD scenario, while relaxing the assumption that the number of words should be the same on target and source sides.

[25] Proportion of correctly segmented utterances.

[26] In Chapter 3, in order to compare corpora of different sizes, we evaluate only the first 500 segmented utterances of each corpus, instead of the 5K utterances in this chapter. In Chapters 4 and 5, we also evaluate the whole Mboshi 5K corpus, as already noted in Section 5.4.1.

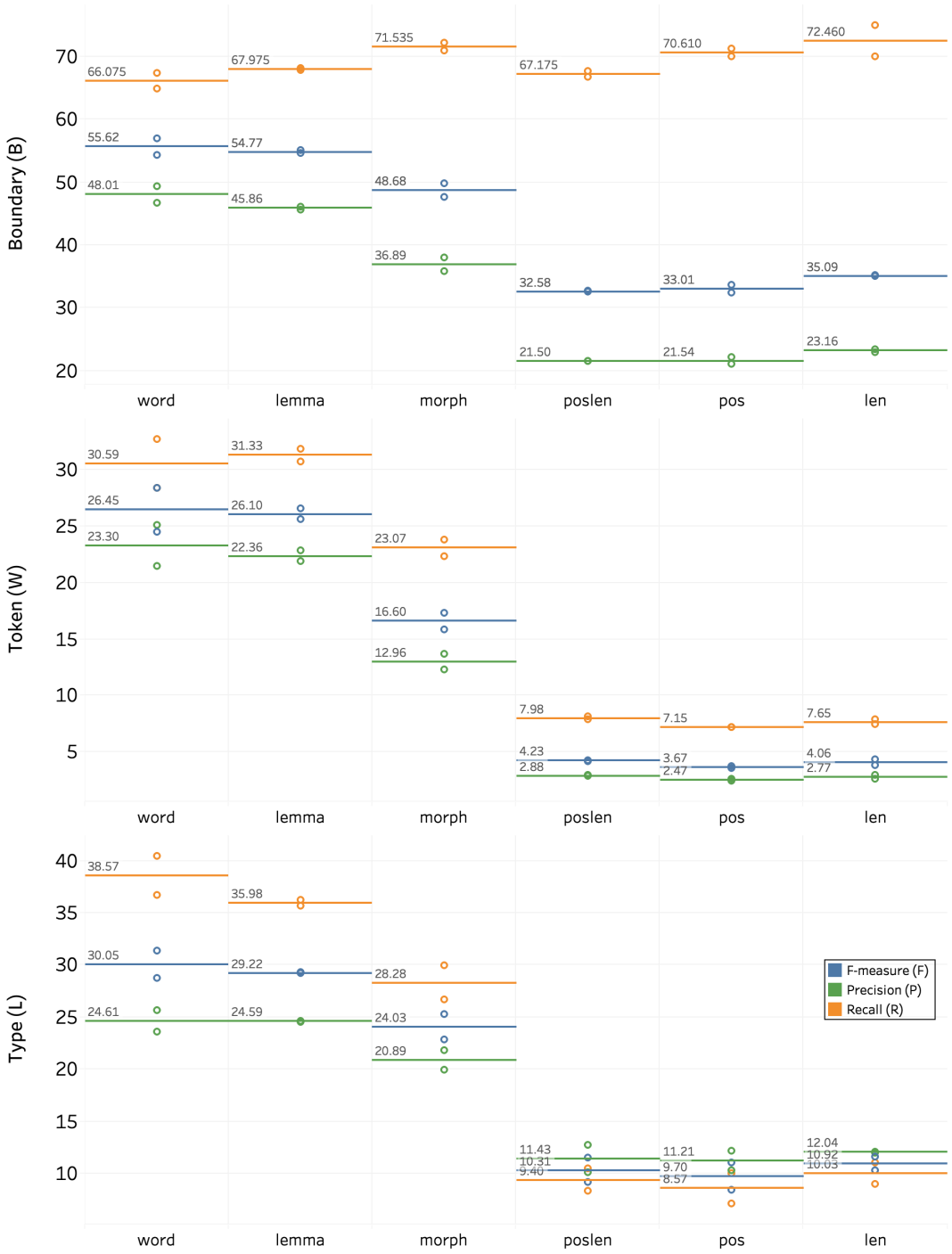


Figure 6.6: Boundary, token, and type metrics (F-measure, precision, recall) with BASE method on the Mboshi 5K corpus for French representations `length`, `pos`, `poslen`, `morph`, `lemma`, and `word`. Horizontal colored lines correspond to values averaged over the 2 runs.

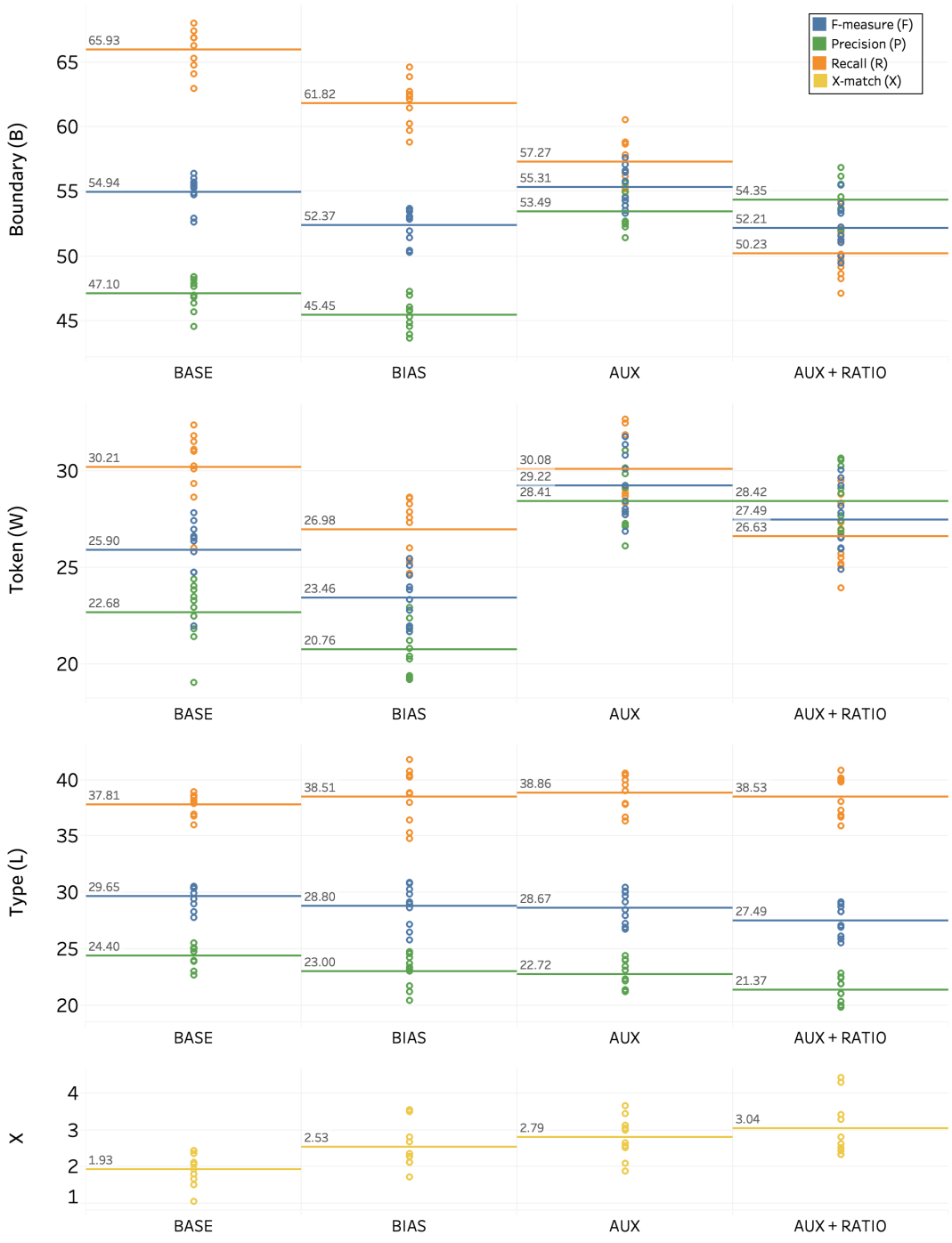


Figure 6.7: Boundary, token, and type metrics (F-measure, precision, recall), and sentence exact-match (X) with methods BASE, BIAS, AUX, and AUX+RATIO, on the Mboshi 5K corpus for French representation `word`. Horizontal colored lines correspond to values averaged over the 10 runs.

results by a margin of 20% to 30%, and can hope to be able to retrieve meaningful alignments.[27]

In terms of source-side representation, our results in Figure 6.6 exhibit a steady degradation of performance when moving towards coarser representations. This is more pronounced with BASE for `lemma` and `morph` than it was with `pisa`. Additionally, the new representations `length` and `poslen`, incorporating an information on the length of source word, yield results in the same ballpark as those obtained with `pos` representation, which were already the lowest with `pisa`.

Effects of the word-length bias Providing word-length information to the model in representations `length` and `poslen` follows the same rationale motivating the addition of a word-length bias in the attention mechanism (see Section 6.3.2). In Figure 6.7, corresponding boundary, token, and type results are reported in column BIAS. As we observed a non-negligible variability in the results obtained with different runs, we now report on 10 runs in this figure, and provide average values. Unfortunately, this second attempt to integrate word-length information from the source side proves to be unsuccessful, and results obtained with BIAS are lower than those obtained with BASE, except for the sentence exact-match metric (X).

To assess whether or not the introduction of our word-length bias encourages, in effect, target units to "attend more" to longer source word in BIAS, we compute the correlation between source words' lengths and the quantity of attention these words receive (that is, for each source position, we sum attention column-wise: $\sum_i \tilde{\alpha}_{ij}$). Results for BIAS, and there counterpart for BASE, AUX, and AUX+RATIO computed with $\sum_i \alpha_{ij}$, are in Table 6.1. We do observe an increased correlation between word lengths and attention when using method BIAS, but this correlation seems already high without any biasing mechanism (BASE, or AUX and AUX+RATIO which we discuss next). Consequently, these results do not contradict our hypothesis that long words are likely to be aligned to long words, and short words with short words. But forcing this correlation appear to be counter-productive.

method	correlation (avg. over 10 runs)
BASE	0.681
BIAS	**0.729**
AUX	0.665
AUX+RATIO	0.662

Table 6.1: Correlation between word lengths and attention (p-value for Pearson coefficient is 0 for each run).

Effects of the auxiliary loss We now turn to the results obtained when training our networks with an auxiliary loss. On (averaged) boundary F-measures (BF) presented in

[27] Evaluating alignments on our data would require a manual reference, which is being built at the time of writing. In this thesis, on the other hand, we tend to avoid using synthetic data, as our experiments have shown that it can be misleading for real documentation scenarios (Section 3.4).

Figure 6.7, AUX performs similarly to BASE, but with a much higher precision, as well as a degraded recall, indicating that the new method does not oversegment as much as BASE. More insight can be gained looking at various statistics on the automatically segmented data in Figure 6.8. The average token and sentence lengths, indeed, for AUX, are closer to the ground-truth values on our corpus. The global number of tokens produced are also brought closer to their references. On token metrics, a similar effect is observed, but the trade-off between a lower recall and an increased precision is more favorable overall, and leads to a significantly increased F-measure (29.22% vs. 25.90%).

If these results are encouraging for documentation purposes, where precision is arguably a more valuable metric than recall to guide the design of tools providing an aid to linguists, and hint them as robustly as possible towards structures present in the data, they are likely sensitive to the particular language pair we study here. The goal we had in mind designing the auxiliary loss was to obtain target segmentations with a similar number of units than the one found in the source. Albeit conceptually acceptable, in general, if we want to segment and align semantically equivalent units on both sides in a one-to-one mapping, the particular evaluation we compute at the word level will be sensitive to the typological similarity (or dissimilarity for that matter) inside the language pair. As Mboshi is more agglutinative than French (5.96 words per sentence on average in the Mboshi 5K, vs. 8.22 for French), we questioned whether injecting some supervision regarding this would improve the performance. In that spirit, method AUX+RATIO makes use of the true sentence length ratio found on the first 100 sentences in the corpus (which could be manually annotated in a realistic language documentation scenario). Results in Figure 6.7 are globally deteriorated, except for the boundary precision (BP) and the sentence exact-match (X), but interestingly, precision becomes stronger than recall for both boundary and token metrics, indicating undersegmentation. This is confirmed in Figure 6.8 by an average token length for the first time above the ground-truth (and equivalently, an average sentence length below the true value).

A puzzling observation, when using these auxiliary losses (AUX or AUX+RATIO), is that the average sentence length decreases to get closer to the true value (5.96). If this leads to an increased precision, it appears counter-intuitive as the auxiliary loss should penalize segmentations producing less words than those found in French (8.22 word per sentence on average). That said, training curves show monotonically decreasing auxiliary losses, and the comparison between average token lengths and average sentence lengths obtained with AUX and AUX+RATIO (Figure 6.8) is conform to the intuition: with the reference average word ratio ($r_{\text{MB/FR}} = 0.789$ on the first 100 sentences of the corpus, see Section 6.4.2) injected in the auxiliary loss, Mboshi tokens become longer, and sentences shorter. Note also that the proposed auxiliary loss only comes into play after a partial convergence of the model according to the main (negative log-likelihood) loss. At this point during training, the impact of the auxiliary loss is likely to be quite constrained by the structure already learnt by the neural network.

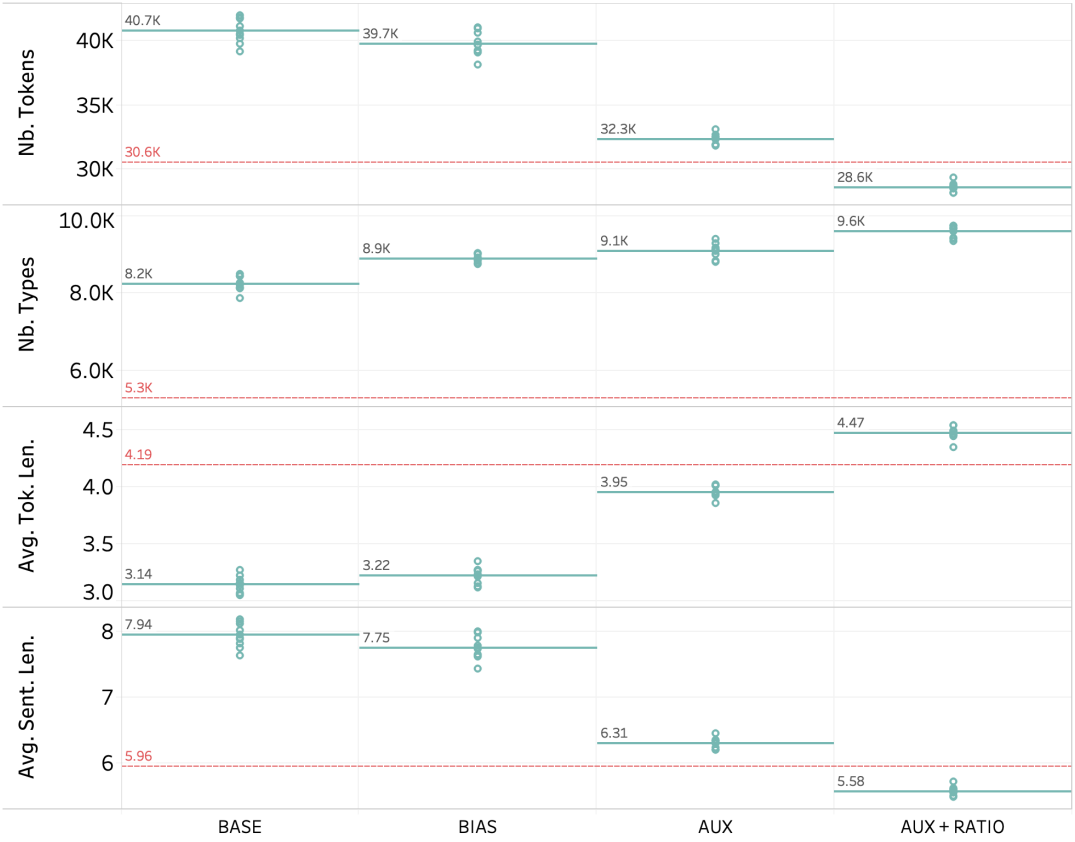


Figure 6.8: Statistics on segmentations produced by methods BASE, BIAS, AUX, and AUX+RATIO, on the Mboshi 5K corpus for French representation `word`: number of tokens, types, average token length (in characters), average sentence lengths (in tokens). Solid (teal-colored) lines correspond to average values (10 runs). Dashed (red) lines indicate the ground-truth values in the Mboshi 5K corpus.

6.5 Conclusion

In this chapter, we described a neural segmentation method in the spirit of the "align to segment" approach, and proposed some extensions to move towards joint segmentation and alignment with this method. This involved the introduction of a word-length bias in the attention mechanism, and the design of an auxiliary loss aiming to integrate segmentation information in the objective function.

Aligning phonemes to words without supervision and under very limited data conditions is a very challenging task. For one thing, units on the target side (characters or phonemes) bear less information than units on the source side (words). Additionally, the attention matrices, treated here as soft-alignment matrices, are not guaranteed to correspond to meaningful alignments.

If we did not beat the best monolingual (Bayesian) methods on the segmentation task, we did improve previous results, in particular for the precision metric, with the

addition of an auxiliary loss. We also explored an alternative extension in the form of a modification to the attention mechanism biasing attention towards longer source words. This method was effective, and produced an increased correlation between (source) word length and attention received, but was not achieving interesting segmentation results.

The real benefit of the work described here should be assessed with alignment metrics. This is left for future work, as the reference alignments are not yet available on our data, although currently being created by linguists. Nevertheless, and unlike monolingual methods, the approach presented here is capable of automatically producing alignments, as was the case with the only bilingual method we were able to experiment with so far (`pisa`). The results we report with our neural segmentation method, however, are much higher than those obtained with `pisa`. Another incentive for our method was provided by experiments conducted in (Godard et al., 2018d) on automatically discovered phone units from speech. Under realistic noisy conditions, the neural approach appeared much more robust than `dpseg`, one of the best performing Bayesian method.

Lastly, and given the benefit of using expert knowledge demonstrated by our experiments with Adaptor Grammars in Chapter 4, integrating elements of supervision, for instance a list of frequent words or the presence of certain word boundaries on the target side, seems a promising avenue for future work. Structured attention (Kim et al., 2017), briefly discussed in Section 7.2.1, could be a way to introduce weak supervision in our segmentation method. More expressive attention models, as proposed by Cohn et al. (2016), should also be tested. For lack of taking advantage of plausible linguistic constraints, the danger with very limited data is to see the benefit of bilingual supervision be outweighed by an increased complexity in the learning process.

Chapter 7

Conclusion

Contents

In this closing chapter, we recall the motivations of our work and summarize our contributions. We identify the main directions for future work and conclude with some perspectives in computational language documentation.

7.1 Summary

In **Chapter 1**, we presented some facts showing that the world's language diversity is under considerable pressure, and that traditional documentary linguistics alone cannot face the challenge of documenting all endangered languages: half of about the 7,000 known languages in the world might die out by the end of this century. Language documentation is a time-consuming effort, and the magnitude of language endangerment would require a much larger number of field linguists than is currently available to address the issue. Therefore, a new field, computational language documentation (CLD), is emerging and aims at helping linguists in their documentation and preservation work, by providing them with automatic processing tools. The work presented in this thesis embraces this goal, and focuses on the particular problem of unsupervised word discovery, i.e. to segment into words an unsegmented stream of phonemes or characters. In the BULB project's data collection scenario (see Section 1.2.2), such a stream of phonemes has been first automatically transcribed from speech data, and a translation into French has also been collected and transcribed.

After formally defining the word segmentation task, as well as the related word alignment task, we showed in **Chapter 2** how the two tasks can play together in a 'segment, then align' or an 'align to segment' approach. In this chapter, we then surveyed the relevant literature and the useful concepts or methods for our work. In particular, we described nonparametric Bayesian models for word segmentation, as well as joint models for unsupervised segmentation and alignment. This led us to make several observations:

- In the past two decades, researchers have moved from word segmentation methods based on local statistics and heuristic search towards more principled models and techniques. Nonparametric Bayesian modeling, and the Adaptor Grammar framework in particular, have led to substantial improvements in word segmentation, and allow the integration of prior knowledge that could be used in the context of CLD.
- Automatic word alignment methods heavily depend on the assumptions made for word granularities on the source or target side. Additionally, the shift to a neural paradigm in machine translation (MT) during recent years has somewhat hindered research in generative alignment models.
- The most promising line of research for word segmentation is the one trying to model jointly segmentation and alignment. Until recently, this research was essentially motivated by MT, but several contributions have lately shown the promises of this approach for CLD. Joint modeling could for instance allow for more robust segmentation when working on automatic (noisy) transcripts from speech.

We concluded Chapter 2 with a list of research questions (see Section 2.7) motivating our own work.

7.1.1 Findings

The first set of questions were centered around the practicality of using word discovery methods in very-low-resource conditions, and especially on real corpora collected in a language documentation scenario. To answer these questions, we introduced in **Chapter 3** two new corpora from low-resource Bantu languages, Mboshi and Myene, and conducted preliminary word segmentation experiments with several existing monolingual or bilingual methods. With an additional French-English corpus, we also assessed the impact of data size and representation. We learnt that with very limited data conditions (about 5,000 sentences), more expressive models can be less effective than simpler ones, especially if tuning is not a practical option. For instance, a more complex parameter estimation for the bilingual method we experimented with seemed responsible for performance degradation. Hence, making use of bilingual supervision appeared less straightforward than we could have expected. More generally, the work presented in Chapter 3 made a strong case for conducting CLD research with realistic language documentation corpora, as these are typically different in nature than simulated low-resource corpora, and can lead to counter-intuitive results. This is a condition to build effective tools for CLD.

We also initially questioned to which extent expert knowledge could be leveraged to improve unsupervised word discovery, and what kind of useful information could be provided to the linguist in a documentation process. In **Chapter 4**, and in collaboration with linguists, we showed that word segmentation can be greatly improved using Adaptor Grammars and specializing a generic grammar with language specific knowledge. We also showed that systematic experiments with a carefully designed grammar landscape can provide useful linguistic insights. On the whole, this work proved that tight collaborations between linguists and computer scientists can be beneficial for the documentation process, and should be encouraged.

In another question, we pondered the worth of tonal information for word segmentation, and **Chapter 5** specifically investigated that question. We were able to show that tonal information is indeed useful in a supervised setting, which inspired us to design a tonal extension to the strongest nonparametric Bayesian model we experimented with in Chapter 3. The capacity of the new model to capture some tonal signal on synthetic data was demonstrated, but we unfortunately failed to exhibit a benefit of using the tonal model on real data. If the limited data supply is likely to be a strong factor (our tonal model requires the estimation of more parameters than its baseline counterpart), we also believe the lack of capacity of our model to capture grammatical, and not only lexical, structures is to blame. In the absence of better performing models, we made a recommendation to strip tonal information for word segmentation in a CLD scenario.

Lastly, and as bilingual supervision seemed uneasy to take advantage of (Chapter 3), we questioned whether we could devise new ways to use bilingual data for word segmentation. In **Chapter 6**, we proposed and extended a neural word segmentation method. We showed that one of our extensions involving an auxiliary loss significantly improved the precision of the segmentations. Results, still, were not on par with the strongest monolingual Bayesian methods, but we argued that learning alignments between the source and the target was crucial. For lack of manual references, the evaluation of such alignments were left for future work, but we showed vast improvements in segmentation when compared to the only other bilingual method at our disposal, making our method promising. Other work (Godard et al., 2018d) also suggested that this method is more robust to noisy conditions than Bayesian methods.

7.1.2 Synthesis of the main results for Mboshi

We present in Table 7.1 a bird's-eye view of the word discovery results from preceding chapters on the Mboshi 5K corpus (Section 3.2.2).

- `dpseg` is the Dirichlet process-based bigram language model of Goldwater et al. (2006a) (see Section 2.4.3);
- "AGs" corresponds to the results obtained with grammar A3-B3-C3-D1+ (see Section 4.3.2) in Chapter 4;
- `pypshmm` is the model described in Chapter 5, with BASE the baseline version, and MULTI the tonal variant (stars indicate results obtained on representation `tone` with diacritics encoding tones, instead of the default representation `notone`);
- `pisa` corresponds to Model 3P (Stahlberg et al., 2012) with the best performing parameters for `giza++` (as segmentation results in Chapter 3 were evaluated only

on the first 500 sentences so as to keep numbers comparable for different corpora sizes, we re-evaluated `pisa`'s results on the whole Mboshi 5K here);

- "Attention" corresponds to the attention-based segmentation method proposed in Chapter 6 (with baseline version BASE, and auxiliary loss extension AUX).

When several runs were run, the results are averaged.

Method	BP	BR	BF	WP	WR	WF	LP	LR	LF	X	Avg. len
`dpseg`	55.72	76.21	64.38	31.02	40.52	35.14	43.14	10.91	17.42	2.04	3.21
AGs$^{\#}$	83.59	81.12	82.34	65.60	63.99	64.78	61.54	54.48	57.79	21.77	4.29
`pypshmm` (BASE)	84.49	60.73	70.66	53.11	40.69	46.08	44.51	47.91	46.15	9.55	5.47
`pypshmm`* (BASE)	69.28	69.24	64.39	43.24	37.52	39.38	38.20	37.95	36.44	8.97	4.23
`pypshmm` (MULTI)	84.60	62.79	72.08	55.05	43.24	48.44	45.96	49.04	47.45	10.64	5.33
`pypshmm`* (MULTI)	72.27	67.79	68.85	45.57	42.03	43.35	39.82	39.23	39.18	8.38	4.42
`pisa`	46.18	18.31	26.22	17.73	8.82	11.78	9.45	12.26	10.68	0.97	8.41
Attention (BASE)	47.10	65.93	54.94	22.68	30.21	25.90	24.40	37.81	29.65	1.93	3.14
Attention (AUX)	53.49	57.27	55.31	28.41	30.08	29.22	22.72	38.86	28.67	2.79	3.95

Table 7.1: Precision, Recall and F-measure on boundaries (BP, BR, BF), tokens (WP, WR, WF), and types (LP, LR, LF), as well as sentence exact-match (X) and average token length, for various (see text) word discovery methods on the Mboshi 5K corpus. Monolingual methods are in the top half and bilingual methods in the bottom half. A star ($*$) indicates that a tonal representation was used (with diacritics), and a hash sign ($\#$), that expert linguistic knowledge was used.

Amongst all methods, Adaptor Grammars using expert knowledge obtain the best segmentation results. The two variants of `pypshmm` (BASE and MULTI) show degraded performance when using tonal representations (this is largely due to the presence of outliers, see Figure 5.2), but produce the strongest segmentation results after AGs, and improve the results of `dpseg` by a substantial margin on type F-measure (due to a better recall). For the bilingual methods, yielding less accurate segmentations than the monolingual ones, the improvement of our attention-based approach over `pisa` is significant. The variant with an auxiliary loss improves further the results, especially due to a better precision on boundaries and tokens.

7.2 Future work

In this section, we identify three main directions for future work: to focus on word alignment as much as on segmentation, to shift from graphemic transcriptions to speech representations, and to make more systematic and efficient use of weak supervision.

7.2.1 Word alignment

As we already pointed out, automatically discovered word units in an unknown language will not fully be useful for language documentation unless these words are reliably aligned to known words in the well-resourced language. This is crucial to help a linguist hint at the discovered units' semantic content, and progress in the documentation work. We worked in that direction with an attention-based segmentation method in Chapter 6, but were not able to evaluate alignments for lack of a gold standard. Reference alignments are currently being created, and a straightforward continuation of our work will be to evaluate automatic alignments produced by our baseline method (and its extensions) with the reference. We will also be able to compare our alignments to those obtained with `pisa` (Stahlberg et al., 2012), and an error analysis should then allow us to propose new improvements for our method. This also suggests an interesting side problem for CLD: finding a robust way of automatically creating reference word alignment annotations from interlinear glosses[1] (see Xia and Lewis, 2007). Such glosses are readily available for many corpora in the Pangloss collection[2] for instance (see also Section 1.2.3).

Related to moving towards improving the quality of both segmentation and alignment, the work of Cohn et al. (2016), modifying the attention mechanism with alignment biases inspired by the automatic word-based alignment literature, seems promising and should be explored in our context. Another promising avenue is the introduction of *structural* biases in the attention by Kim et al. (2017): as the authors explain, standard attention does not model any structural dependencies between source units, yet with limited data (the assumed condition in CLD), a neural network might not be able to learn these dependencies implicitly. The authors propose a general formulation of the attention as the probability of a latent variable z conditioned on the encoder's representations $h_1, \ldots, h_J$ and the decoder's query s_i. The context vectors c_i are defined here as the expectation of an "annotation function" under the probability distribution of z. In this formulation, Bahdanau's attention can be seen as a categorical latent variable whose sample space is defined by the set of source positions. But z can also be chosen as a vector of discrete latent variables whose distribution is specified by a conditional random field (CRF) (Lafferty et al., 2001) capturing dependencies in the source. This allows the modeling of a (weak) linguistic supervision (see Section 7.2.3).

7.2.2 Towards speech

In Section 3.1.1, we explained the choice made to use graphemic transcriptions created by linguists in the experiments presented in this thesis: we wanted to assess the impact of a number of factors on unsupervised word discovery (quantity and representation of the data, language modeling assumptions, usability of the bilingual supervision) without risking for our analysis to be obfuscated by noisy data, and to also keep results comparable throughout this document.

[1] See `https://www.eva.mpg.de/lingua/resources/glossing-rules.php` for the Leipzig glossing rules.

[2] `http://lacito.vjf.cnrs.fr/pangloss/index_en.html`.

That said, Chapter 1 made it clear that any research effort in CLD ultimately requires to work from speech. We have already been working in that direction. The results of a monolingual segmentation pipeline from speech data making use of the Dirichlet process-based language model of Goldwater et al. (2006a) were reported in (Godard et al., 2018a). In (Ondel et al., 2018), unsupervised word discovery was explored as an extrinsic criterion to measure the quality of acoustic units automatically discovered from speech. We also tested our baseline neural segmentation method on such acoustic units in (Godard et al., 2018d) and showed that it was more robust to noisy transcripts than a strong Bayesian word segmentation algorithm. The extensions proposed in Section 6.3.2 (word-length bias and auxiliary loss) should now be tested with acoustic units. Another idea would be to have linguists classify discovered acoustic units into vowels and consonants (or devise a way to automate this process), and to experiment with Adaptor Grammars: the methodology proposed in Chapter 4 would then be applicable to a realistic scenario (see also (Lee et al., 2015) already mentioned in Section 2.4.5).

A trend, as seen in recent contributions to CLD (Kamper et al., 2015; Duong et al., 2016; Anastasopoulos et al., 2017; Anastasopoulos and Chiang, 2018b; Adams et al., 2018; Bansal et al., 2018a), is to work directly from speech signal. For unsupervised word discovery, some preliminary speech-to-text translation experiments we conducted on the Mboshi 5K corpus in (Scharenborg et al., 2018a) showed, however, that this task was still very hard on such a small corpus; attention matrices (with Mboshi speech on the source side and French words on the target side) most of the time did not seem usable as soft alignments, and more work is required to demonstrate the usability of our attention-based segmentation method directly from speech.[3]

7.2.3 Leveraging weak supervision

Given the usefulness of integrating expert knowledge demonstrated in Chapter 4 (see also Table 7.1), future work should focus on collecting and better handling such knowledge or supervision. With Adaptor Grammars, we encoded grammatical hypotheses (e.g. dependencies between words or syllables, list of prefixes) but we did so in a static way. An interactive and iterative approach seems the right paradigm for CLD, and a linguist should be able to make a hypothesis, assess results, provide (partially) new information and obtain new results, etc. Such *online learning* scheme could draw inspiration from the work of Sirts and Goldwater (2013) showing how Adaptor Grammars can be used in a semi-supervised setting. The impact on word discovery of adapting or not a particular non-terminal should also be explored.[4] More generally, the goal would be to explore the structured set of all possible grammars, and use word discovery evaluation on a test set (reasonable in size for CLD, maybe a couple hundred of manually segmented sentences) to guide the linguist in the exploration.

In the attention-based segmentation method from Chapter 6, we believe that the weak supervision provided by the bilingual data could be put to better use. One idea would be to learn (from a limited number of sentence pairs manually segmented by a

[3]Recently, Anastasopoulos and Chiang (2018a) also left this for future work and experimented with acoustic units instead for the word discovery task.

[4]To our knowledge, no systematic study on the effect of the adaptation exists.

linguist) a regression model predicting the number of words in the target sentence from its number of characters and from the number of words and characters in the source sentence. This prediction could then be integrated in the auxiliary loss we proposed in Section 6.3.2. An important limitation to the work presented in Chapter 6 is that we evaluated our method solely on the Mboshi corpus. More insights could be gained experimenting with other languages and corpora. One straightforward idea would also be to train our models with pre-trained embeddings on the source (well-resourced) side; syntactic or semantic information could also be used, for instance projected dependency structures as in (Xia and Lewis, 2007).

We should finally consider more carefully the implications of aligning characters or phonemes (instead of words) to words. The work of Kudo (2018), for instance, indicates a way to use segmentation ambiguity to train an NMT model more robustly, using a regularization scheme.

7.3 Perspectives in CLD

We hope this research convincingly shows that working with real low-resource corpora, especially endangered languages, and encouraging collaboration between linguists and computer scientists, is crucial for the advance of CLD.

Working with realistic corpora should help computer scientists to come up with methods and tools that stand the test of helping an actual documentary linguist in their work. Recently, Adams et al. (2018); Michaud et al. (2018) have made an impressive demonstration of the applicability of the most recent research in speech and language technology into the linguist's workflow (in this case the phonemic transcription of a low-resource tonal language), and Adams (2017) acknowledges this was largely made possible due to a strong collaboration with linguist Alexis Michaud. Our best results for unsupervised word discovery relied on a method using expert knowledge gathered during many fruitful discussions with linguist Annie Rialland. In our own experience, one recurring difficulty has been our very limited knowledge of the Bantu languages we processed. If this prevented us from making biased decisions in various modeling choice, it also made any *qualitative* analysis of our data and results much harder. As discussed in Section 2.1.2, *quantitative* segmentation metrics, on the other hand, have some important limitations: their capacity to measure the linguistic soundness of discovered words or units is not nuanced (a word segmented in two of its morphemes, for instance, will yield the same results as a random segmentation). Therefore, a tight collaboration with linguists allows for a very precious qualitative error feedback. Besides, it gives an opportunity to make sure we do not depart from reasonable assumptions in the context of endangered language documentation.

Several currently active lines of research seem extremely relevant for CLD. One of these, inspired by the way infants acquire language, attempts to discover linguistic units using multi-modal inputs: in lieu of often lacking orthographic transcriptions, speech signal or images are used (Scharenborg et al., 2018a; Kamper and Roth, 2018). Embeddings of speech (see Chung et al., 2018) could also be a way to bypass transcription. NMT using transfer learning with limited bilingual data for low-resource languages has also been proposed by Zoph et al. (2016). Other approaches inspired by transfer learn-

ing include (Chen et al., 2017; Gu et al., 2018; Kocmi and Bojar, 2018; Neubig and Hu, 2018). This trend is related to another recent line of research attempting to perform *unsupervised* NMT (Lample et al., 2018; Artetxe et al., 2018), which could suit CLD well as monolingual corpora only are needed in this approach (and such corpora are easier to collect than bilingual ones), but applicability to low-resource settings still needs to be demonstrated. Zero-shot (i.e. assuming very minimal or no parallel data) cross-lingual transfer, as in (Artetxe and Schwenk, 2018) (sentence embeddings used for classification), (Xie et al., 2018) (named entity recognition) , or (Rijhwani et al., 2019) (entity linking), also constitutes a very promising avenue of research for CLD.

Such a swarming and fast-paced research landscape is truly heart-warming for those of us who care for the world's language diversity. It is still quite a tall order, and many reasonable forces concur to push us towards uniformization; I realize there is a certain irony in writing these lines in a language into which I was not born. If we are to meet the challenges of language endangerment, and find the means to scale up the output of documentary linguistics, we will undoubtedly need to engage in many kinds of collaborative and iterative processes, from collecting data to making cutting-edge machine learning algorithms applicable to CLD. I can only hope for this work to have modestly contributed to what I believe to be a meaningful effort.

Summary in French

Dans le **chapitre 1**, nous montrons que la diversité linguistique subit une pression considérable à travers le monde, et que la linguistique documentaire traditionnelle ne peut à elle seule relever le défi de la documentation pour toutes les langues en danger : parmi environ 7 000 langues connues dans le monde, la moitié pourraient disparaître d'ici la fin du siècle. La documentation linguistique est une activité chronophage, et l'ampleur de la menace qui pèse sur les langues exigerait un nombre de linguistes de terrain beaucoup plus important que celui dont nous disposons actuellement. C'est pourquoi un nouveau domaine, la documentation linguistique computationnelle (CLD[5]), est en train de voir le jour avec l'ambition d'aider les linguistes dans leur travail de documentation et de préservation, en leur fournissant des outils de traitement automatique. Les travaux présentés dans cette thèse partagent cette ambition, en se concentrant sur le problème particulier de la découverte non supervisée de mots, c'est-à-dire la segmentation en mots d'un flux non segmenté de phonèmes ou de caractères. Dans le scénario de collecte de données du projet BULB (voir section 1.2.2), un tel flux de phonèmes aura d'abord été transcrit automatiquement à partir du signal de parole, et une traduction en français aura également été recueillie et transcrite.

Après avoir défini formellement la tâche de segmentation en mots ainsi que la tâche d'alignement mot à mot associée, nous montrons dans le **chapitre 2** comment les deux tâches peuvent s'articuler dans une approche « segmenter, puis aligner », ou au contraire dans une approche « aligner pour segmenter ». Dans ce chapitre, nous étudions ensuite la littérature pertinente et introduisons les concepts ou méthodes utiles pour notre travail. En particulier, nous décrivons certains modèles bayésiens non paramétriques pour la segmentation en mots, ainsi que des modèles joints pour la segmentation et l'alignement non supervisés. Ceci nous amène à faire plusieurs observations :

- Au cours des deux dernières décennies, la communauté scientifique a progressivement abandonné diverses méthodes de segmentation en mots basées sur des statistiques locales ou des critères heuristiques, pour adopter des modèles et des techniques aux principes mieux définis. La modélisation bayésienne non paramétrique, et le formalisme des *Adaptor Grammars* (AGs) en particulier, ont permis d'améliorer considérablement la segmentation automatique en mots, et d'intégrer des connaissances *a priori* qui peuvent être disponibles dans le contexte de la CLD.
- Les méthodes automatiques d'alignement mot à mot dépendent fortement des hypothèses faites sur la granularité des mots du côté de la source ou de la cible.

[5] Computational Language Documentation

De plus, l'adoption d'un paradigme neuronal en traduction automatique au cours des cinq dernières années a significativement ralenti la recherche sur les modèles génératifs d'alignement.

- L'axe de recherche le plus prometteur pour la segmentation en mots consiste à modéliser conjointement la segmentation et l'alignement. Jusqu'à récemment, cette recherche était essentiellement motivée par la traduction automatique, mais plusieurs contributions ont récemment démontré les avantages de cette approche pour la CLD. La modélisation conjointe pourrait par exemple permettre une segmentation plus robuste lorsque l'on travaille sur des transcriptions automatiques (bruitées) du signal de parole.

Nous concluons le chapitre 2 par une liste de questions de recherche (voir section 2.7) qui motivent notre propre travail.

La première série de questions porte sur l'utilisation en pratique des méthodes de découverte non supervisée de mots lorsque les ressources sont très limitées, en particulier sur des corpus réels recueillis pour le besoin de la documentation linguistique. Afin de répondre à ces questions, nous introduisons dans le **chapitre 3** deux nouveaux corpus en langues bantoues peu dotées (mboshi et myene), et conduisons des expériences préliminaires de segmentation en mots avec plusieurs méthodes monolingues ou bilingues existantes. À l'aide d'un corpus français-anglais supplémentaire, nous évaluons également l'impact de la taille et de la représentation des données sur nos résultats. Ceci nous enseigne qu'avec des quantités de données très limitées (environ 5 000 phrases), les modèles les plus expressifs peuvent se révéler moins efficaces que les modèles les plus simples, surtout lorsqu'une calibration fine n'est pas envisageable en pratique. À titre d'exemple, l'estimation plus complexe des paramètres de la méthode bilingue que nous utilisons semble responsable d'une dégradation des résultats obtenus avec cette méthode. Par conséquent, tirer parti d'un signal de supervision bilingue s'avère moins simple à mettre en oeuvre que nous l'espérions. De façon plus générale, les travaux présentés dans le chapitre 3 établissent de solides arguments en faveur d'une recherche en CLD s'appuyant sur des corpus réalistes pour la linguistique documentaire. Ceux-ci exhibent en effet généralement des propriétes différentes de celles des corpus construits pour simuler artificiellement une faible quantité de données, ce qui peut conduire à des résultats contre-intuitifs. Fonder l'approche expérimentale sur des corpus correspondant à une véritable collecte documentaire constitue selon nous une nécessité si l'on veut élaborer des outils efficaces pour la CLD.

Une seconde série de questions interroge la manière dont des connaissances expertes pourraient être mises à profit pour améliorer la découverte non supervisée de mots, et quel type d'information utile pourrait être fourni au linguiste dans le cadre d'un processus de documentation. Dans le **chapitre 4**, et en collaboration avec des linguistes, nous montrons que la segmentation en mots peut être grandement améliorée en utilisant le formalisme des *Adaptor Grammars* et en spécialisant une grammaire générique avec des connaissances linguistiques spécifiques. Nous montrons également que des expériences systématiques utilisant un champ grammatical soigneusement conçu peuvent fournir des informations linguistiques utiles. Dans l'ensemble, ces travaux prouvent qu'une collaboration étroite entre linguistes et informaticiens peut être bénéfique pour le processus de documentation et devrait être encouragée.

Nous interrogeons également l'utilité de l'information tonale pour la segmentation en mots, et le **chapitre 5** étudie spécifiquement cette question. Nous montrons que l'information tonale s'avère en effet utile dans un cadre supervisé, ce qui nous amène à concevoir une extension tonale pour le modèle bayésien non paramétrique le plus performant étudié au chapitre 3. La capacité de ce nouveau modèle à capturer un signal tonal sur des données synthétiques est démontrée, mais nous ne parvenons malheureusement pas à mettre en évidence l'efficacité du modèle sur des données réelles. Si la faible quantité de données disponibles constitue certainement un facteur important pour expliquer ce résultat négatif (notre modèle tonal requiert l'estimation d'un plus grand nombre de paramètres que sa contrepartie non tonale), nous croyons également que l'incapacité de notre modèle à capturer certaines régularités structurelles au niveau grammatical, et pas uniquement au niveau lexical, joue un rôle important. En l'absence de modèles plus performants, nous recommandons par conséquent de supprimer l'information tonale pour la segmentation en mots dans un scénario de documentation linguistique computationnelle.

Enfin, et dans la mesure où la supervision bilingue semblait initialement difficile à exploiter (chapitre 3), nous nous demandons si nous pouvons trouver de nouvelles façons d'utiliser ces données bilingues pour la segmentation en mots. Dans le **chapitre 6**, nous proposons et perfectionnons une méthode neuronale de segmentation en mots. Nous montrons que l'une de nos extensions mettant en jeu une fonction objectif auxiliaire améliore significativement la précision des segmentations. Les résultats obtenus n'atteignent toutefois pas ceux des méthodes bayésiennes monolingues les plus efficaces, mais nous rappelons qu'il est crucial d'apprendre des alignements entre les mots de la source et les unités découvertes en cible, et que ceci requiert une approche bilingue. Faute de références annotées manuellement, l'évaluation de ces alignements devra être réalisée dans de futurs travaux, mais nous montrons de grandes améliorations dans la segmentation par rapport à l'unique autre méthode bilingue à notre disposition, ce qui rend notre méthode prometteuse. D'autres travaux (Godard et al., 2018d) ont également suggéré que cette méthode est plus robuste aux conditions bruitées que les méthodes bayésiennes.

Nous présentons dans le tableau 7.1 une vue synthétique des principaux résultats de découverte non supervisée de mots sur le corpus Mboshi 5K (section 3.2.2) obtenus dans cette thèse. Dans ce tableau, lorsque plusieurs évaluations ont été réalisées les résultats sont moyennés. Ils correspondent aux méthodes suivantes :

- `dpseg` est le modèle de langue bigramme de Goldwater et al. (2006a) basé sur des processus de Dirichlet (voir section 2.4.3) ;
- « AGs » correspond aux résultats obtenus avec la grammaire A3-B3-C3-D1+. (voir section 4.3.2) au chapitre 4 ;
- `pypshmm` est le modèle décrit dans le chapitre 5, avec BASE la version baseline, et MULTI la variante tonale (les étoiles indiquent les résultats obtenus sur la représentation `tone` qui encode les tons avec des signes diacritiques, par contraste avec la représentation par défaut `notone`) ; `pisa` correspond au Model 3P (Stahlberg et al., 2012) avec les paramètres les plus performants pour `giza++`. Étant donné que les résultats de segmentation du chapitre 3 ont été évalués uniquement

sur les 500 premières phrases du corpus afin de fournir des résultats comparables pour différentes tailles de corpus, nous avons réévalué les résultats de `pisa` sur l'ensemble du corpus Mboshi 5K ici ;

- « Attention » correspond à la méthode de segmentation neuronale basée sur l'attention proposée au chapitre 6 (avec la version baseline BASE, et l'extension avec fonction objectif auxiliaire AUX).

Parmi toutes les méthodes considérées, les *Adaptor Grammars* utilisant des connaissances expertes obtiennent les meilleurs résultats de segmentation. Les deux variantes de `pypshmm` (BASE et MULTI) conduisent à des performances dégradées lorsque des représentations tonales sont utilisées (ceci est largement dû à la présence de valeurs aberrantes, voir figure 5.2), mais produisent les meilleurs résultats de segmentation après les AGs, et améliorent substantiellement les résultats de `dpseg` en termes de F-mesure sur les types (grâce à un meilleur rappel). Pour les méthodes bilingues, qui produisent des segmentations moins précises que les méthodes monolingues, l'amélioration obtenue grâce à notre approche basée sur l'attention par rapport à `pisa` est significative. Notre variante utilisant une fonction objectif auxiliaire améliore encore les résultats, notamment grâce à une meilleure précision sur les frontières et les tokens.

Bibliography

Oliver Adams. *Automatic Understanding of Unwritten Languages*. PhD thesis, The University of Melbourne, 2017. *[pages 9, 10, and 129]*

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. Inducing Bilingual Lexicons from Small Quantities of Sentence-Aligned Phonemic Transcriptions. In *12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015. *[pages 45 and 50]*

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. Learning a Translation Model from Word Lattices. In *Proceedings of INTERSPEECH*, pages 2518–2522, 2016a. *[page 52]*

Oliver Adams, Graham Neubig, Trevor Cohn, Steven Bird, Quoc Truong Do, and Satoshi Nakamura. Learning a Lexicon and Translation Model from Phoneme Lattices. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2377–2382, Austin, Texas, November 2016b. Association for Computational Linguistics. *[pages 49 and 52]*

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. Evaluating Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation. 2018. *[pages 9, 128, and 129]*

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitri Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. Breaking the Unwritten Language Barrier: The Bulb Project. In *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia, 2016. *[page 9]*

Tamer Alkhouli and Hermann Ney. Biasing Attention-Based Recurrent Neural Networks Using External Alignment Information. In *Proceedings of the Second Conference on Machine Translation*, pages 108–117, 2017. *[page 108]*

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. On The Alignment Problem In Multi-Head Attention-Based Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, Belgium, Brussels, October 2018. Association for Computational Linguistics. *[page 103]*

Alexandre Allauzen and François Yvon. Méthodes statistiques pour la traduction automatique. *Gaussier, E. et Yvon, F., éditeurs: Modèles statistiques pour l'accès à l'information textuelle*, 7:271–356, 2011. *[page 40]*

Célestin Amboulou. *Le Mbochi: Langue Bantu Du Congo-Brazzaville (Zone C, Groupe C20)*. PhD thesis, INALCO, Paris, 1998. *[page 83]*

Odette Ambouroue. *Éléments de Description de l'orungu, Langue Bantu Du Gabon (B11b)*. PhD thesis, Université Libre de Bruxelles, 2007. *[page 58]*

Antonios Anastasopoulos. *Computational Tools for Endangered Language Documentation*. PhD thesis, University of Notre Dame, 2019. *[page 9]*

Antonios Anastasopoulos and David Chiang. Tied Multitask Learning for Neural Speech Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. *[pages 9, 59, and 128]*

Antonios Anastasopoulos, Sameer Bansal, David Chiang, Sharon Goldwater, and Adam Lopez. Spoken Term Discovery for Language Documentation using Translations. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 53–58, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. *[pages 10 and 128]*

Antonis Anastasopoulos and David Chiang. Leveraging translations for speech transcription in low-resource settings. *arXiv:1803.08991 [cs]*, March 2018b. *[pages 9, 59, and 128]*

Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Ann. Statist.*, 2(6):1152–1174, November 1974. *[page 31]*

Mark Aronoff. *Word Formation in Generative Grammar*. Linguistic Inquiry Monographs. MIT Press, Cambridge, MA, 1976. *[page 28]*

Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *arXiv:1812.10464 [cs]*, December 2018. *[page 130]*

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised Neural Machine Translation. In *ICLR 2018*, 2018. *[page 130]*

Eric Auer, Peter Wittenburg, Han Sloetjes, Oliver Schreer, Stefano Masneri, Daniel Schneider, and Sebastian Tschöpel. Automatic annotation of media field recordings. *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, 2010. *[page 9]*

Peter K. Austin and Julia Sallabank. *The Cambridge Handbook of Endangered Languages*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, Cambridge, 2011. *[pages 6 and 7]*

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. *[pages 103, 105, 106, 108, and 113]*

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. Towards speech-to-text translation without speech recognition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 474–479, Valencia, Spain, April 2017. Association for Computational Linguistics. *[page 56]*

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Low-Resource Speech-to-Text Translation. In *Interspeech 2018*, pages 1298–1302. ISCA, September 2018a. *[pages 9, 56, and 128]*

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv:1809.01431 [cs]*, September 2018b. *[pages 9 and 59]*

Patricia Bedrosian. The Mboshi noun class system. *Journal of West African Languages*, 26(1):27–47, 1996. *[page 58]*

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. Painless Unsupervised Learning with Features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590, Los Angeles, California, June 2010. Association for Computational Linguistics. *[page 50]*

James Bergstra, Daniel Yamins, and David Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA, June 2013. PMLR. *[page 62]*

Laurent Besacier, Bowen Zhou, and Yuqing Gao. Towards Speech Translation of Non Written Languages. In *Spoken Language Technology Workshop, 2006.*, pages 222–225. IEEE, 2006. *[pages 8, 24, 46, and 47]*

Jeff A. Bilmes and Katrin Kirchhoff. Factored Language Models and Generalized Parallel Backoff. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 2, pages 4–6, Edmonton, Canada, 2003. Association for Computational Linguistics. *[page 94]*

Frédéric Bimbot, Sabine Deligne, and François Yvon. Unsupervised decomposition of phoneme strings into variable-length sequences, by multigrams. In *International Conference of PHonetic Sciences (ICPHS)*, Stockholm, Sweden, 1995. *[page 25]*

Steven Bird. *A Scalable Method for Preserving Oral Literature from Small Languages*. Springer, 2010. *[page 8]*

Steven Bird. Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology*, 6:1–16, 2011. *[page 8]*

Steven Bird and David Chiang. Machine Translation for Language Preservation. In *Proceedings of COLING 2012: Posters*, pages 125–134, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. *[page 8]*

Steven Bird, Florian R. Hanke, Oliver Adams, and Haejoong Lee. Aikuma: A mobile app for collaborative language documentation. *ACL 2014*, 2014. *[page 8]*

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. Parallel Speech Collection for Under-resourced Language Studies Using the LIG-AIKUMA Mobile Device App. *Procedia Computer Science*, 81:61–66, 2016. *[pages 9 and 59]*

David Blackwell and James B. MacQueen. Ferguson Distributions Via Polya Urn Schemes. *Ann. Statist.*, 1(2):353–355, March 1973. *[page 33]*

Wilhelm Heinrich Immanuel Bleek. *De Nominum Generibus: Linguarum Africae Australis, Copticae, Semiticarum Aliarumque Sexualium...* apud Adolphum Marcum, 1851. *[page 58]*

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. *[page 39]*

James P. Blevins. Word-Based Morphology. *Journal of Linguistics*, 42(03):531–573, 2006. *[page 28]*

Marcely Zanon Boito, Antonios Anastasopoulos, Marika Lekakou, Aline Villavicencio, and Laurent Besacier. A small Griko-Italian speech translation corpus. In *Proceedings of SLTU*, 2018. *[page 9]*

Benjamin Börschinger and Mark Johnson. Exploring the Role of Stress in Bayesian Word Segmentation using Adaptor Grammars. *Transactions of the Association of Computational Linguistics*, 2:93–104, 2014. *[page 39]*

Jan A. Botha and Phil Blunsom. Adaptor Grammars for Learning Non-Concatenative Morphology. In *EMNLP*, pages 345–356, 2013. *[page 39]*

Michael R. Brent. An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning*, 34(1-3):71–105, 1999. *[page 24]*

Matthias Brenzinger. *Language Diversity Endangered*. Trends in Linguistics. Studies and Monographs [TiLSM]. De Gruyter, 2008. *[page 7]*

Matthias Brenzinger, Arienne Dwyer, Tjeerd de Graaf, Colette Grindevald, Michael Krauss, Osahito Miyaoka, Nicholas Ostler, Osamu Sakiyama, María E. Villalón, Akira Y. Yamamoto, and Ofelia Zepeda. Vitalité et disparition des langues. Technical report, UNESCO Report (Groupe d'experts spécial de l'UNESCO sur les langues en danger)., 2003. *[page 7]*

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993. *[pages 40 and 42]*

James Brunning. *Alignment Models and Algorithms for Statistical Machine Translation.* PhD thesis, 2010. *[page 40]*

Franck Burlot and François Yvon. Morphology-Aware Alignments for Translation to and from a Synthetic Language. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT'15, pages 188–195, Da Nang, Vietnam, 2015. *[page 46]*

Franck Burlot and François Yvon. Learning Morphological Normalization for Translation from and into Morphologically Rich Languages. *Prague Bulletin of Mathematical Linguistics*, (108):49–60, 2017. *[page 46]*

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. A Teacher-Student Framework for Zero-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada, 2017. Association for Computational Linguistics. *[page 130]*

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014a. Association for Computational Linguistics. *[pages 103 and 105]*

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014b. Association for Computational Linguistics. *[page 103]*

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014. *[page 113]*

Tagyoung Chung and Daniel Gildea. Unsupervised Tokenization for Machine Translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 718–726, Singapore, August 2009. Association for Computational Linguistics. *[pages 48 and 49]*

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces. In *Advances in Neural Information Processing Systems 31*, pages 7365–7375, 2018. *[page 129]*

Shay B. Cohen, David M. Blei, and Noah A. Smith. Variational inference for adaptor grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 564–572. Association for Computational Linguistics, 2010. *[page 39]*

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. Incorporating structural alignment biases into an attentional neural translation model. *arXiv preprint arXiv:1601.01085*, 2016. *[pages 102, 108, 110, 121, and 127]*

Marta R. Costa-jussà and José A. R. Fonollosa. Character-based Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany, August 2016. Association for Computational Linguistics. *[page 45]*

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, Hoboken, NJ, 2. ed edition, 2006. *[page 26]*

Mathias Creutz. Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Sapporo, Japan, July 2003. Association for Computational Linguistics. *[page 26]*

Mathias Creutz and Krista Lagus. Unsupervised Discovery of Morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics, July 2002. *[page 26]*

Mathias Creutz and Krista Lagus. Induction of a Simple Morphology for Highly-Inflecting Languages. In *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pages 43–51, Barcelona, Spain, July 2004. Association for Computational Linguistics. *[page 26]*

Mathias Creutz and Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. In *In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, 2005. *[page 27]*

Mathias Creutz and Krista Lagus. Unsupervised Models for Morpheme Segmentation and Morphology Learning. *ACM Trans. Speech Lang. Process.*, 4(1):3:1–3:34, 2007. *[page 27]*

David Crystal. *Language Death*. Cambridge University Press, 2000. *[pages 6 and 7]*

Adrià de Gispert and José B. Mariño. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*, 50(11-12):1034–1046, November 2008. *[page 46]*

Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual*

Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pages 73–76, Boulder, Colorado, June 2009. *[page 46]*

Carl de Marcken. Linguistic Structure As Composition and Perturbation. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 335–341, Santa Cruz, California, 1996. Association for Computational Linguistics. *[page 26]*

Hervé Déjean. Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora. In *Proceedings of the Workshop on Paradigms and Grounding in Natural Language Learning*, pages 295–299, Adelaide, Australia, 1998. *[page 24]*

Sabine Deligne and Frédéric Bimbot. Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference On*, volume 1, pages 169–172. IEEE, 1995. *[pages 24 and 26]*

Sabine Deligne, Francois Yvon, and Frédéric Bimbot. Variable-Length Sequence Matching for Phonetic Transcription Using Joint Multigrams. In *Fourth European Conference on Speech Communication and Technology*, 1995. *[page 25]*

Sabine Deligne, François Yvon, and Frédéric Bimbot. Selection of Multiphone Synthesis Units and Grapheme-to-Phoneme Transcription Using Variable-Length Modeling of Strings. In *Data-Driven Techniques in Speech Synthesis*, pages 125–147. Springer, 2001. *[page 24]*

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39 (1):1–38, 1977. *[page 43]*

John DeNero, Dan Gillick, James Zhang, and Dan Klein. Why Generative Phrase Models Underperform Surface Heuristics. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 31–38, New York City, June 2006. Association for Computational Linguistics. *[pages 49 and 51]*

John DeNero, Alexandre Bouchard-Côté, and Dan Klein. Sampling Alignment Structure under a Bayesian Translation Model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 314–323, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. *[pages 49 and 51]*

Yonggang Deng and William Byrne. HMM Word and Phrase Alignment for Statistical Machine Translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 169–176, Vancouver, British Columbia, Canada, 2005. Association for Computational Linguistics. *[pages 47 and 49]*

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 1370–1380, Baltimore, Maryland, June 2014. Association for Computational Linguistics. *[page 103]*

Mark Dingemanse, Jeremy Hammond, Herman Stehouwer, Aarthy Somasundaram, and Sebastian Drude. A high speed transcription interface for annotating primary linguistic data. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 7–12, Avignon, France, April 2012. Association for Computational Linguistics. *[page 9]*

Lise M. Dobrin and Jeff Good. Practical Language Development: Whose Mission? *Language*, 85(3):619–629, 2009. *[page 7]*

Laura J. Downing. On the ambiguous segmental status of nasals in homorganic NC sequences. In *The Internal Organization of Phonological Segments*, pages 183–216. 2005. *[page 84]*

Markus Dreyer and Jason Eisner. Graphical Models over Multiple Strings. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 101–110, Singapore, 2009. Association for Computational Linguistics. *[page 28]*

Markus Dreyer and Jason Eisner. Discovering Morphological Paradigms from Plain Text Using a Dirichlet Process Mixture Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 616–627, Edinburgh, United Kingdom, 2011. Association for Computational Linguistics. *[page 28]*

Markus Dreyer, Jason R. Smith, and Jason Eisner. Latent-variable Modeling of String Transductions with Finite-state Methods. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1080–1089, Honolulu, Hawaii, 2008. Association for Computational Linguistics. *[page 28]*

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. The Zero Resource Speech Challenge 2017. In *Automatic Speech Recognition and Understanding (ASRU), 2017 IEEE Workshop On*. IEEE, 2017. *[page 8]*

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. An attentional model for speech translation without transcription. In *Proceedings of NAACL-HLT*, pages 949–959, 2016. *[pages 56, 102, 110, and 128]*

Ilknur Durgar El-Kahlout and François Yvon. The pay-offs of preprocessing for German-English Statistical Machine Translation. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the Seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 251–258, Paris, France, 2010. *[page 46]*

Chris Dyer. Using a maximum entropy model to build segmentation lattices for MT. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages

406–414, Boulder, Colorado, June 2009. Association for Computational Linguistics. *[page 46]*

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. *[pages 43 and 108]*

Christopher J. Dyer. The "Noisier Channel": Translation from Morphologically Complex Languages. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 207–211, Prague, Czech Republic, June 2007. Association for Computational Linguistics. *[page 46]*

Halvor Eifring and Rolf Theil. *Linguistics for Students of Asian and African Languages*. 2004. *[page 22]*

Georges Martial Embanga Aborobongui. *Processus Segmentaux et Tonals En Mbondzi – (Variété de La Langue Embosi C25)*. PhD thesis, Université Paris 3 Sorbonne Nouvelle, 2013. *[pages 57, 58, and 83]*

Ramy Eskander, Owen Rambow, and Tianchun Yang. Extending the Use of Adaptor Grammars for Unsupervised Morphological Segmentation of Unseen Languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 900–910, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. *[pages 75 and 76]*

Nicholas Evans and Stephen C. Levinson. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32 (05):429, October 2009. *[pages 7 and 85]*

Daniel L. Everett. Cultural Constraints on Grammar and Cognition in Pirahã: Another Look at the Design Features of Human Language. *Current Anthropology*, 46(4):621–646, August 2005. *[page 85]*

Elif Eyigöz, Daniel Gildea, and Kemal Oflazer. Simultaneous Word-Morpheme Alignment for Statistical Machine Translation. In *HLT-NAACL*, pages 32–40, 2013. *[page 46]*

Shi Feng, Shujie Liu, Mu Li, and Ming Zhou. Implicit Distortion and Fertility Models for Attention-based Encoder-Decoder NMT Model. January 2016. *[page 108]*

Mark Fishel and Harri Kirik. Linguistically Motivated Unsupervised Segmentation for Machine Translation. In *Proceedings of the Language Ressources and Evaluation Conference*, La Valette, Malta, 2010. *[pages 46 and 47]*

J.A. Fishman. *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Multilingual Matters. Multilingual Matters, 1991. *[page 6]*

Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. Whyisenglishsoeasytosegment. *Proc. of CMCL*, 2013. *[pages 69 and 80]*

Alexander Fraser and Daniel Marcu. Measuring Word Alignment Quality for Statistical Machine Translation. *Computational linguistics*, 33(3):293–303, 2007. *[page 21]*

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. Modeling Inflection and Word-Formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 664–674, Avignon, France, April 2012. Association for Computational Linguistics. *[page 46]*

Kuzman Ganchev, João V. Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 99:2001–2049, 2010. *[page 44]*

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional Sequence to Sequence Learning. *arXiv:1705.03122 [cs]*, May 2017. *[page 103]*

Hamidreza Ghader and Christof Monz. What does Attention in Neural Machine Translation Pay Attention to? *arXiv:1710.03348 [cs]*, October 2017. *[page 106]*

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, May 2010. PMLR. *[page 114]*

Pierre Godard. *Typologie Pour l'alignement Multilingue*. Mémoire de Master, Université Sorbonne Nouvelle – Paris 3, 2014. *[page 16]*

Pierre Godard and François Yvon. Enlightening the Bulb : Unsupervised learning of morphology for word and subword alignments. Technical report, 2016. *[page 16]*

Pierre Godard, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Laurent Besacier, Hélène Bonneau-Maynard, Guy-Noël Kouarata, Kevin Löser, Annie Rialland, and François Yvon. Preliminary Experiments on Unsupervised Word Discovery in Mboshi. In *Proceedings of Interspeech*, San-Francisco, California, USA, 2016. *[pages 55, 62, and 67]*

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélène Maynard, Markus Müller, Annie Rialland, Sebastian Stüker, François Yvon, and Marcely Zanon Boito. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. In *Proceedings of LREC*, Miyazaki, Japan, 2018a. *[pages 11, 55, 56, 59, and 128]*

Pierre Godard, Laurent Besacier, François Yvon, Martine Adda-Decker, Gilles Adda, Hélène Maynard, and Annie Rialland. Adaptor Grammars for the Linguist: Word Segmentation Experiments for Very Low-Resource Languages. In *Proceedings of the*

15th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON), Brussels, Belgium, 2018b. *[page 73]*

Pierre Godard, Kevin Loser, Alexandre Allauzen, Laurent Besacier, and Francois Yvon. Unsupervised Learning of Word Segmentation: Does Tone Matter? In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, Hanoi, Vietnam, 2018c. *[pages 54 and 89]*

Pierre Godard, Marcely Zanon Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, and Laurent Besacier. Unsupervised Word Segmentation from Speech with Attention. In *Proceedings of Interspeech*, Hyderabad, India, 2018d. *[pages 11, 56, 101, 102, 109, 113, 114, 121, 125, 128, and 133]*

John Goldsmith. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198, June 2001. *[pages 22 and 27]*

Sharon Goldwater. *Nonparametric Bayesian Models of Lexical Acquisition*. PhD thesis, Brown University, 2006. *[pages 20, 26, 29, 31, 32, 33, 35, 76, and 80]*

Sharon Goldwater and David McClosky. Improving Statistical MT through Morphological Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. *[page 46]*

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Contextual Dependencies in Unsupervised Word Segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680, Sydney, Australia, July 2006a. Association for Computational Linguistics. *[pages 33, 38, 48, 62, 83, 125, 128, and 133]*

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Interpolating Between Types and Tokens by Estimating Power-Law Generators. In *Advances in Neural Information Processing Systems 18*, pages 459–466, Cambridge, MA, 2006b. MIT Press. *[page 38]*

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. A Bayesian Framework for Word Segmentation: Exploring the Effects of Context. *Cognition*, 112(1):21–54, 2009. *[pages 62, 65, 67, 80, and 83]*

Li Gong, Aurélien Max, and François Yvon. Improving bilingual sub-sentential alignment by sampling-based transpotting. In *Proceedings of the International Workshop on Spoken Language Translation*, page 8, Heidelberg, Germany, 2013. *[page 51]*

João Graça, Kuzman Ganchev, and Ben Taskar. Learning tractable word alignment models with complex constraints. *Computational Linguistics*, 36(3):481–504, 2010. *[page 44]*

Joao Graça, Kuzman Ganchev, and Ben Taskar. Expectation Maximization and Posterior Constraints. In *NIPS*, volume 20, pages 569–576, 2007. *[page 44]*

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. *[page 27]*

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. *[page 130]*

Malcolm Guthrie. *The Classification of the Bantu Languages*. Oxford University Press, 1948. *[page 57]*

Malcolm Guthrie. *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages*. Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages. Gregg, 1967. *[page 57]*

Nizar Habash and Fatiha Sadat. Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 49–52, New York City, USA, June 2006. Association for Computational Linguistics. *[page 46]*

Geoffrey Haig, Nicole Nau, Stefan Schnell, and Claudia Wegener. *Documenting Endangered Languages: Achievements and Perspectives*. De Gruyter Mouton, 2011. *[page 11]*

Ken Hale. On endangered languages and the safeguarding of diversity. *Language*, 68 (1):1–3, 1992. *[page 7]*

Ken Hale, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. Endangered Languages. *Language*, 68(1):1–42, 1992. *[page 5]*

Fatima Hamlaoui and Emmanuel-Moselly Makasso. Focus marking and the unavailability of inversion structures in the Bantu language Bàsàá. *Lingua*, 154:35–64, 2015. *[page 78]*

Harald Hammarström and Lars Borin. Unsupervised Learning of Morphology. *Computational Linguistics*, 37(2):309–350, June 2011. *[page 22]*

Florian R. Hanke and Steven Bird. Large-scale text collection for unwritten languages. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1134–1138, 2013. *[pages 8 and 9]*

Zellig S. Harris. From Phoneme to Morpheme. *Language*, 31(2):pp. 190–222, 1955. *[page 24]*

K. David Harrison. *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge*. Oxford Studies in Sociolinguistics Series. Oxford University Press, 2007. *[pages 6 and 7]*

Martin Haspelmath. Should linguistic diversity be conserved like biodiversity? https://dlc.hypotheses.org/195, 2012. *[page 7]*

Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 298(5598):1569–1579, November 2002. *[page 85]*

Robert K. Herbert. *Language Universals, Markedness Theory, and Natural Phonetic Processes*. De Gruyter Mouton, Berlin, Boston, 1986. *[page 84]*

Jahn Heymann, Oliver Walter, Reinhold Haeb-Umbach, and Bhiksha Raj. Iterative Bayesian word segmentation for unsupervised vocabulary discovery from phoneme lattices. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference On*, pages 4057–4061. IEEE, 2014. *[page 63]*

Nikolaus P. Himmelmann. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors, *Trends in Linguistics. Studies and Monographs [TiLSM]*. Mouton de Gruyter, Berlin, New York, January 2006. *[page 10]*

Leanne Hinton. Approaches to and Strategies for Language Revitalization. In Kenneth L. Rehg and Lyle Campbell, editors, *The Oxford Handbook of Endangered Languages*. Oxford University Press, 2018. *[page 8]*

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. *[page 103]*

Yu Hu, Irina Matveeva, John Goldsmith, and Colin Sprague. Refining the SED Heuristic for Morpheme Discovery: Another Look at Swahili. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 28–35, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. *[page 28]*

Larry M. Hyman. Lexical vs. Grammatical Tone: Sorting out the Differences. In *Tonal Aspects of Languages 2016*, pages 6–11, Buffalo, New York, USA, 2016. *[page 99]*

Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard Rose, et al. A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition. 2013. *[page 52]*

Mark Johnson. Unsupervised Word Segmentation for Sesotho Using Adaptor Grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, Ohio, 2008a. Association for Computational Linguistics. *[pages 39, 75, 80, and 90]*

Mark Johnson. Using Adaptor Grammars to Identify Synergies in the Unsupervised Acquisition of Linguistic Structure. In *Proceedings of ACL-08: HLT*, pages 398–406, Columbus, Ohio, 2008b. Association for Computational Linguistics. *[pages 38, 39, and 75]*

Mark Johnson and Katherine Demuth. Unsupervised phonemic Chinese word segmentation using Adaptor Grammars. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 528–536. Association for Computational Linguistics, 2010. *[page 90]*

Mark Johnson and Sharon Goldwater. Improving Nonparameteric Bayesian Inference: Experiments on Unsupervised Word Segmentation with Adaptor Grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, Boulder, Colorado, 2009. Association for Computational Linguistics. *[pages 36, 39, 65, 75, 76, and 80]*

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. Bayesian Inference for PCFGs via Markov Chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146, Rochester, New York, April 2007a. Association for Computational Linguistics. *[page 37]*

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. Adaptor Grammars: A Framework for Specifying Compositional Nonparametric Bayesian Models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648, Cambridge, MA, 2007b. MIT Press. *[pages 36, 37, 39, and 74]*

Mark Johnson, Anne Christophe, Emmanuel Dupoux, and Katherine Demuth. Modelling Function Words Improves Unsupervised Word Segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 282–292, Baltimore, Maryland, 2014. Association for Computational Linguistics. *[pages 39 and 75]*

Herman Kamper. *Unsupervised Neural and Bayesian Models for Zero-Resource Speech Processing*. PhD thesis, University of Edinburgh, 2016. *[page 9]*

Herman Kamper and Michael Roth. Visually Grounded Cross-Lingual Keyword Spotting in Speech. In *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 248–252. ISCA, August 2018. *[page 129]*

Herman Kamper, Aren Jansen, and Sharon Goldwater. Fully Unsupervised Small-Vocabulary Speech Recognition Using a Segmental Bayesian Model. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. *[page 128]*

Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, and Simon Kirby. Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52, August 2017. *[page 110]*

Timothy Kempton and Roger K. Moore. Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. *Speech Communication*, 56:152–166, January 2014. *[page 8]*

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured Attention Networks. In *5th International Conference on Learning Representations*, page 21, 2017. *[pages 108, 121, and 127]*

Tom Kocmi and Ondřej Bojar. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Belgium, Brussels, October 2018. Association for Computational Linguistics. *[page 130]*

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, Cambridge; New York, 2010. *[pages 20, 46, 51, and 102]*

Philipp Koehn. Neural Machine Translation. *arXiv:1709.07809 [cs]*, September 2017. *[page 102]*

Philipp Koehn and Kevin Knight. Empirical methods for compound splitting. In *EACL '03: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, pages 187–193, Budapest, Hungary, 2003. Association for Computational Linguistics. *[page 46]*

Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics. *[pages 106 and 108]*

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003. *[page 40]*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007. *[page 56]*

Oskar Kohonen, Sami Virpioja, and Krista Lagus. Semi-Supervised Learning of Concatenative Morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86, Uppsala, Sweden, July 2010. Association for Computational Linguistics. *[page 27]*

Guy-Noël Kouarata. *Variations de Formes Dans La Langue Mbochi (Bantu C25)*. PhD thesis, Université Lumière Lyon 2, 2014. *[pages 57 and 83]*

Michael Krauss. The world's languages in crisis. *Language*, 68(1):4–10, 1992. *[pages 5 and 8]*

Shaohui Kuang, Junhui Li, António Branco, Weihua Luo, and Deyi Xiong. Attention Focusing for Neural Machine Translation by Bridging Source and Target Embeddings. *arXiv:1711.05380 [cs]*, November 2017. *[page 108]*

Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *arXiv:1804.10959 [cs]*, April 2018. *[page 129]*

Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. Morpho Challenge Competition 2005–2010: Evaluations and Results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95. Association for Computational Linguistics, 2010. *[page 24]*

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. *[page 127]*

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised Machine Translation Using Monolingual Corpora Only. In *ICLR 2018*, page 14, 2018. *[page 130]*

Adrien Lardilleux, François Yvon, and Yves Lepage. Hierarchical sub-sentential alignment with anymalign. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*, pages 279–286, 2012. *[page 51]*

Thomas Lavergne, Olivier Cappé, and François Yvon. Practical Very Large Scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513, Uppsala, Sweden, July 2010. Association for Computational Linguistics. *[page 60]*

Hai-Son Le, Alexandre Allauzen, and François Yvon. Continuous Space Translation Models with Neural Networks. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–48, Montréal, Canada, June 2012. Association for Computational Linguistics. *[page 103]*

Chia-ying Lee, Timothy J. O'Donnell, and James Glass. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403, 2015. *[pages 39 and 128]*

Marika Lekakou, Valeria Baldissera, and Antonis Anastasopoulos. Documentation and analysis of an endangered language: Aspects of the grammar of Griko. Technical report, John S. Latsis Public Benefit Foundation, 2013. *[page 9]*

M Paul Lewis. *Ethnologue: Languages of the World*. SIL international, 16th edition, 2009. *[page 6]*

M Paul Lewis. *Ethnologue: Languages of the World*. SIL international, 18th edition, 2015. *[page 6]*

M Paul Lewis. *Ethnologue: Languages of the World*. SIL international, 21st edition, 2018. *[page 6]*

M Paul Lewis and Gary F Simons. Assessing Endangerment: Expanding Fishman's GIDS. *Revue roumaine de linguistique*, 55(2):102–120, 2010. *[page 6]*

Molly L. Lewis and Michael C. Frank. The length of words reflects their conceptual complexity. *Cognition*, 153:182–195, August 2016. *[page 110]*

Percy Liang, Ben Taskar, and Dan Klein. Alignment by Agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111. Association for Computational Linguistics, 2006. *[page 44]*

Junyang Lin, Xu Sun, Xuancheng Ren, Muyu Li, and Qi Su. Learning When to Concentrate or Divert Attention: Self-Adaptive Attention Temperature for Neural Machine Translation. *arXiv:1808.07374 [cs]*, August 2018. *[page 110]*

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Neural Machine Translation with Supervised Attention. *arXiv:1609.04186 [cs]*, September 2016. *[page 108]*

Adam Lopez. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49, August 2008. *[page 20]*

Kevin Löser and Alexandre Allauzen. Une méthode non-supervisée pour la segmentation morphologique et l'apprentissage de morphotactique à l'aide de processus de Pitman-Yor. 2016. *[pages 36, 63, 83, 89, and 93]*

Bogdan Ludusan, Gabriel Synnaeve, and Emmanuel Dupoux. Prosodic boundary information helps unsupervised word segmentation. In *Annual Conference of the North American Chapter of the ACL*, pages 953–963, Denver, Colorado, USA, 2015. *[pages 54 and 90]*

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015. *[page 108]*

Daniel Marcu and Daniel Wong. A Phrase-Based,Joint Probability Model for Statistical Machine Translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 133–139. Association for Computational Linguistics, July 2002. *[pages 49 and 51]*

I. Dan Melamed. Manual Annotation of Translational Equivalence: The Blinker Project. *arXiv preprint cmp-lg/9805005*, 1998. *[page 21]*

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. Supervised Attentions for Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2283–2288, Austin, Texas, November 2016. Association for Computational Linguistics. *[page 108]*

Alexis Michaud, Oliver Adams, Trevor Cohn, Graham Neubig, and Séverine Guillaume. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12:393–429, 2018. *[page 129]*

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048. ISCA, 2010. *[page 103]*

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics, 2009. *[pages 35, 36, 49, 50, 52, 63, 65, and 94]*

Robert C. Moore. Improving IBM word-alignment model 1. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 518. Association for Computational Linguistics, 2004. *[page 42]*

Markus Müller, Sebastian Stüker, and Alex Waibel. Language Adaptive DNNs for Improved Low Resource Speech Recognition. In *Interspeech 2016*, pages 3878–3882, September 2016. *[page 10]*

Markus Müller, Jörg Franke, Sebastian Stüker, and Alex Waibel. Towards Phoneme Inventory Discovery for Documentation of Unwritten Languages. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017. *[page 10]*

Jason Naradowsky and Kristina Toutanova. Unsupervised Bilingual Morpheme Segmentation and Alignment with Context-rich Hidden Semi-Markov Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 895–904, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. *[page 49]*

Daniel Nettle and Suzanne Romaine. *Vanishing Voices: The Extinction of the World's Languages*. Oxford University Press, 2000. *[page 6]*

Graham Neubig. Simple, Correct Parallelization for Blocked Gibbs Sampling. Technical report, Nara Institute of Science and Technology, 2014. *[page 63]*

Graham Neubig and Junjie Hu. Rapid Adaptation of Neural Machine Translation to New Languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium, 2018. Association for Computational Linguistics. *[page 130]*

Graham Neubig, Masato Mimura, Shinsuke Mori, and Tatsuya Kawahara. Learning a language model from continuous speech. In *INTERSPEECH*, pages 1053–1056. Citeseer, 2010. *[pages 35, 52, and 63]*

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. An Unsupervised Model for Joint Phrase Alignment and Extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 632–641, Portland, Oregon, 2011. *[pages 49, 50, and 51]*

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. Machine Translation without Words through Substring Alignment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 165–174, Jeju Island, Korea, July 2012. Association for Computational Linguistics. *[pages 49 and 52]*

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang. XNMT: The eXtensible Neural Machine Translation Toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts, USA, 2018. *[page 112]*

ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. Nonparametric Word Segmentation for Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 815–823, Beijing, China, 2010. Association for Computational Linguistics. *[page 49]*

Sonja Nießen and Hermann Ney. Toward hierarchical models for statistical machine translation of inflected languages. In *Proceedings of the Workshop on Data-Driven Methods in Machine Translation*, volume 14, pages 1–8, Toulouse, France, 2001. Association for Computational Linguistics. *[page 46]*

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics, 2000. *[page 40]*

Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003. *[pages 21, 44, and 48]*

David Odden. Bantu Phonology. *Oxford Handbooks Online*, December 2015. *[page 84]*

Timothy J. O'Donnell, Joshua B. Tenenbaum, and Noah D. Goodman. Fragment Grammars: Exploring Computation and Reuse in Language. Technical report, Massachusetts Institute of Technology, 2009. *[page 39]*

Kemal Oflazer and Ilknur Durgar El-Kahlout. Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32, Prague, Czech Republic, June 2007. Association for Computational Linguistics. *[page 46]*

Lucas Ondel, Pierre Godard, Laurent Besacier, Elin Larsen, Mark Hasegawa-Johnson, Odette Scharenborg, Emmanuel Dupoux, Lukas Burget, François Yvon, and Sanjeev Khudanpur. Bayesian Models for Unit Discovery on a Very Low Resource Language. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018. *[pages 11, 54, 56, 70, and 128]*

Gabriele Pallotti. A Simple View of Linguistic Complexity. *Second Language Research*, 31(1):117–134, January 2015. *[page 85]*

Naomi Palosaari and Lyle Campbell. Structural aspects of language endangerment. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, Cambridge Handbooks in Language and Linguistics, pages 100–119. Cambridge University Press, 2011. *[page 7]*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. *[page 43]*

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017. *[page 112]*

Jan-Thorsten Peter, Arne Nix, and Hermann Ney. Generating Alignments Using Target Foresight in Attention-Based Neural Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):27–36, June 2017. *[page 106]*

Steven T. Piantadosi, Harry Tily, and Edward Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9): 3526–3529, March 2011. *[page 110]*

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, December 2013. *[page 56]*

Nima Pourdamghani, Marjan Ghazvininejad, and Kevin Knight. Using Word Vectors to Improve Word Alignments for Low Resource Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 524–528, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. *[page 102]*

Annie Rialland and Martial Embanga Aborobongui. How intonations interact with tones in Embosi (Bantu C25), a two-tone language without downdrift. In *Intonation in African Tone Languages*, volume 24. De Gruyter, Berlin, Boston, 2016. *[page 58]*

Annie Rialland, Georges Martial Embanga Aborobongui, Martine Adda-Decker, and Lori Lamel. Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in Embosi. In *Proceedings of the 44th ACAL Meeting*, pages 221–230, Somerville, 2015. Cascadilla. *[page 58]*

Annie Rialland, Martine Adda-Decker, Guy-Noël Kouarata, Gilles Adda, Laurent Besacier, Lori Lamel, Élodie Gauthier, Pierre Godard, and Jamison Cooper-Leavitt. Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville). In *Proceedings of LREC*, Miyazaki, Japan, 2018. *[No citation]*

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. Zero-shot Neural Transfer for Cross-lingual Entity Linking. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, Honolulu, Hawaii, January 2019. *[page 130]*

Jorna Rissanen. *Stochastic Complexity in Statistic Inquiry*. Series in Computer Science - Vol 15. World Scientific Publishing, 1989. *[pages 22 and 25]*

Suzanne Romaine. Language Endangerment and Language Death. In *The Routledge Handbook of Ecolinguistics*, chapter chapter3. Routledge, 2017. *[page 6]*

J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical Learning by 8-Month-Old Infants. *Science*, 274(5294):1926–1928, December 1996. *[page 24]*

Jeanette Sakel and Daniel L. Everett. *Linguistic Fieldwork: A Student Guide*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2012. *[page 8]*

Baskaran Sankaran, Haitao Mi, Yaser Al-Onaizan, and Abe Ittycheriah. Temporal Attention Model for Neural Machine Translation. August 2016. *[page 108]*

Odette Scharenborg, Laurent Besacier, Alan W. Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux. Linguistic Unit Discovery from Multi-Modal Inputs in Unwritten Languages: Summary of the "Speaking Rosetta" JSALT 2017 Workshop. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018a. *[pages 56, 128, and 129]*

Odette Scharenborg, Patrick Ebel, Mark Hasegawa-Johnson, and Najim Dehak. Building an ASR System for Mboshi Using A Cross-Language Definition of Acoustic Units Approach. In *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 167–171. ISCA, August 2018b. *[page 59]*

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, volume 12, pages 44–49. Manchester, UK, 1994. *[page 60]*

M. Schuster and K.K. Paliwal. Bidirectional Recurrent Neural Networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November 1997. *[page 113]*

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. *[page 45]*

Claude Elwood Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. *[page 69]*

Kairit Sirts and Sharon Goldwater. Minimally-Supervised Morphological Segmentation Using Adaptor Grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266, 2013. *[pages 75 and 128]*

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. Morfessor 2.0: Toolkit for Statistical Morphological Segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University, 2014. *[page 60]*

Benjamin Snyder and Regina Barzilay. Cross-lingual Propagation for Morphological Analysis. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 848–854, Chicago, Illinois, 2008a. AAAI Press. *[pages 49 and 51]*

Benjamin Snyder and Regina Barzilay. Unsupervised Multilingual Learning for Morphological Segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio, June 2008b. *[pages 49 and 51]*

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, January 2014. *[page 114]*

Felix Stahlberg, Tim Schlippe, Stephan Vogel, and Tanja Schultz. Word segmentation through cross-lingual word-to-phoneme alignment. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 85–90. IEEE, 2012. *[pages 42, 45, 49, 50, 63, 66, 102, 125, 127, and 133]*

Felix Stahlberg, Tim Schlippe, Stephan Vogel, and Tanja Schultz. Word segmentation and pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. *Computer Speech & Language*, 2014. *[page 50]*

Sebastian Stüker, Laurent Besacier, and Alex Waibel. Human Translations Guided Language Discovery for ASR Systems. In *10th International Conference on Speech Science and Speech Technology (InterSpeech 2009)*, pages 1–4, Brighton (UK), 2009. Eurasip. *[page 8]*

Sebastian Stüker, Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitri Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso,

Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. Innovative Technologies for Under-Resourced Language Documentation: The Bulb Project. In *Proceedings of CCURL (Collaboration and Computing for Under-Resourced Languages : Toward an Alliance for Digital Language Diversity)*, Portorõz, Slovenia, 2016. *[No citation]*

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014. *[pages 103 and 105]*

Gabriel Synnaeve, Isabelle Dautriche, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. Unsupervised Word Segmentation in Context. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2326–2334, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. *[page 39]*

Yee Whye Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics, 2006. *[pages 34 and 95]*

Jörg Tiedemann. *Bitext Alignment.* Synthesis Lectures on Human Language Technologies. 2011. *[page 44]*

Kristina Toutanova and Michel Galley. Why Initialization Matters for IBM Model 1: Multiple Optima and Non-Strict Convexity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 461–466, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. *[page 44]*

Tasaku Tsunoda. *Language Endangerment and Language Revitalization: An Introduction.* Mouton de Gruyter, 2006. *[page 8]*

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany, August 2016. Association for Computational Linguistics. *[page 108]*

Nicola Ueffing and Hermann Ney. Using POS Information for SMT into Morphologically Rich Languages. In *10th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2003. *[page 46]*

Clara Vania and Adam Lopez. From Characters to Words to in Between: Do We Capture Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada, July 2017. Association for Computational Linguistics. *[page 23]*

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, June 2017. *[page 103]*

Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. The zero resource speech challenge 2015. In *Proc. of Interspeech*, 2015. *[page 8]*

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of the Machine Translation Summit XI*, pages 491–498, Copenhagen, Denmark Copenhagen, Denmark, 2007. *[pages 46 and 47]*

Sami Virpioja, Jaakko Väyrynen, Andre Mansikkaniemi, and Mikko Kurimo. Applying Morphological Decompositions to Statistical Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 195–200, Uppsala, Sweden, July 2010. Association for Computational Linguistics. *[page 46]*

Stephan Vogel. PESA: Phrase Pair Extraction as Sentence Splitting. 2005. *[page 51]*

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 836–841, Morristown, NJ, USA, 1996. *[pages 43 and 47]*

Ronald J. Williams and David Zipser. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2):270–280, June 1989. *[page 104]*

Anthony C. Woodbury. Defining documentary linguistics. In Peter K. Austin, editor, *Language Documentation and Description*, volume 1, pages 35–51. London, 2003. *[page 9]*

Anthony C. Woodbury. Language Documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, Cambridge Handbooks in Language and Linguistics, pages 159–186. Cambridge University Press, Cambridge, 2011. *[page 8]*

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403, 1997. *[page 51]*

Fei Xia and William Lewis. Multilingual Structural Projection across Interlinear Text. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 452–459, Rochester, New York, April 2007. Association for Computational Linguistics. *[pages 127 and 129]*

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. Neural Cross-Lingual Named Entity Recognition with Minimal Resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

pages 369–379, Brussels, Belgium, 2018. Association for Computational Linguistics. *[page 130]*

Jia Xu, Richard Zens, and Hermann Ney. Partitioning Parallel Documents Using Binary Segmentation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 78–85, New York City, June 2006. Association for Computational Linguistics. *[page 51]*

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. Bayesian Semi-Supervised Chinese Word Segmentation for Statistical Machine Translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1017–1024, Manchester, UK, August 2008. Coling 2008 Organizing Committee. *[pages 48 and 49]*

Zichao Yang, Zhiting Hu, Yuntian Deng, Chris Dyer, and Alex Smola. Neural Machine Translation with Recurrent Attention Modeling. *arXiv preprint arXiv:1607.05108*, 2016. *[page 108]*

Reyyan Yeniterzi and Kemal Oflazer. Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL'10, pages 454–464, 2010. *[page 46]*

Marcely Zanon Boito, Alexandre Berard, Aline Villavicencio, and Laurent Besacier. Unwritten Languages Demand Attention Too! Word Discovery with Encoder-Decoder Models. In *Automatic Speech Recognition and Understanding (ASRU), 2017 IEEE Workshop On*. IEEE, 2017. *[pages 101, 102, 109, 113, and 115]*

Ke Zhai, Jordan Boyd-Graber, and Shay B. Cohen. Online adaptor grammars with hybrid inference. *Transactions of the Association for Computational Linguistics*, 2: 465–476, 2014. *[page 39]*

Ying Zhang and Stephan Vogel. Competitive Grouping in Integrated Phrase Segmentation and Alignment Model. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 159–162, Ann Arbor, Michigan, June 2005a. Association for Computational Linguistics. *[page 51]*

Ying Zhang and Stephan Vogel. An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora. In *In Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT-05)*, pages 30–31, 2005b. *[page 51]*

Ying Zhang, Stephan Vogel, and Alex Waibel. Integrated phrase segmentation and alignment algorithm for statistical machine translation. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 567–573, October 2003. *[pages 49 and 51]*

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on*

Empirical Methods in Natural Language Processing, pages 1568–1575, Austin, Texas, November 2016. Association for Computational Linguistics. *[page 129]*

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Découverte non-supervisée de mots pour outiller la linguistique de terrain

Mots clés : apprentissage non-supervisé; segmentation automatique en mots; alignement bilingue; modèles bayésiens; langues peu dotées

Résumé : La diversité linguistique est actuellement menacée : la moitié des langues connues dans le monde pourraient disparaître d'ici la fin du siècle. Cette prise de conscience a inspiré de nombreuses initiatives dans le domaine de la linguistique documentaire au cours des deux dernières décennies, et 2019 a été proclamée Année internationale des langues autochtones par les Nations Unies, pour sensibiliser le public à cette question et encourager les initiatives de documentation et de préservation. Néanmoins, ce travail est coûteux en temps, et le nombre de linguistes de terrain, limité.
Par conséquent, le domaine émergent de la documentation linguistique computationnelle (CLD) vise à favoriser le travail des linguistes à l'aide d'outils de traitement automatique. Le projet *Breaking the Unwritten Language Barrier* (BULB), par exemple, constitue l'un des efforts qui définissent ce nouveau domaine, et réunit des linguistes et des informaticiens. Cette thèse examine le problème particulier de la découverte de mots dans un flot non segmenté de caractères, ou de phonèmes, transcrits à partir du signal de parole dans un contexte de langues très peu dotées. Il s'agit principalement d'une procédure de segmentation, qui peut également être couplée à une procédure d'alignement lorsqu'une traduction est disponible.
En utilisant deux corpus en langues bantoues correspondant à un scénario réaliste pour la linguistique documentaire, l'un en Mboshi (République du Congo) et l'autre en Myene (Gabon), nous comparons diverses méthodes monolingues et bilingues de découverte de mots sans supervision. Nous montrons ensuite que l'utilisation de connaissances linguistiques expertes au sein du formalisme des Adaptor Grammars peut grandement améliorer les résultats de la segmentation, et nous indiquons également des façons d'utiliser ce formalisme comme outil de décision pour le linguiste. Nous proposons aussi une variante tonale pour un algorithme de segmentation bayésien nonparamétrique, qui utilise un schéma de repli modifié pour capturer la structure tonale. Pour tirer parti de la supervision faible d'une traduction, nous proposons et étendons, enfin, une méthode de segmentation neuronale basée sur l'attention, et améliorons significativement la performance d'une méthode bilingue existante.

Title : Unsupervised Word Discovery for Computational Language Documentation

Keywords : unsupervised learning; automatic word segmentation; bilingual alignment; Bayesian models; low-resource languages

Abstract : Language diversity is under considerable pressure: half of the world's languages could disappear by the end of this century. This realization has sparked many initiatives in documentary linguistics in the past two decades, and 2019 has been proclaimed the International Year of Indigenous Languages by the United Nations, to raise public awareness of the issue and foster initiatives for language documentation and preservation. Yet documentation and preservation are time-consuming processes, and the supply of field linguists is limited.
Consequently, the emerging field of computational language documentation (CLD) seeks to assist linguists in providing them with automatic processing tools. The *Breaking the Unwritten Language Barrier* (BULB) project, for instance, constitutes one of the efforts defining this new field, bringing together linguists and computer scientists. This thesis examines the particular problem of discovering words in an unsegmented stream of characters, or phonemes, transcribed from speech in a very-low-resource setting. This primarily involves a segmentation procedure, which can also be paired with an alignment procedure when a translation is available.
Using two realistic Bantu corpora for language documentation, one in Mboshi (Republic of the Congo) and the other in Myene (Gabon), we benchmark various monolingual and bilingual unsupervised word discovery methods. We then show that using expert knowledge in the Adaptor Grammar framework can vastly improve segmentation results, and we indicate ways to use this framework as a decision tool for the linguist. We also propose a tonal variant for a strong nonparametric Bayesian segmentation algorithm, making use of a modified backoff scheme designed to capture tonal structure. To leverage the weak supervision given by a translation, we finally propose and extend an attention-based neural segmentation method, improving significantly the segmentation performance of an existing bilingual method.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France