ABSTRACT

Title of dissertation: MULTILINGUAL USE OF TWITTER: LANGUAGE CHOICE AND LANGUAGE BRIDGES IN A SOCIAL NETWORK

Irene Eleta, Doctor of Philosophy, 2014

Dissertation directed by: Professor Jennifer Golbeck
College of Information Studies

Social media is international: users from different cultures and language backgrounds are generating and sharing content. But language barriers emerge in the communication landscape online. In the quest for language diversity and universal access, the vision of a cosmopolitan Internet has stumbled over the language frontier.

Expatriates, minorities, diasporic communities, and language learners play an important role in forming transnational networks, creating social ties across borders. Many users of social media are multicultural and multilingual; they are mediating between language communities. In the microblogging site Twitter, information spreads across languages and countries. How are multilingual users of Twitter connecting language groups? What are the factors influencing their language choices? This research advances a step towards understanding the network structures and communication strategies that enable intercultural dialog, cross-language sharing of information, and awareness of global problems.

This dissertation research aims at: (1) exploring the ways in which multilingual users of Twitter are connecting different language groups in their social network; (2) modeling how the network influences their language choices; (3) and exploring what the textual features of their posts can elicit about language choices and mediation between groups.

This dissertation goes beyond survey information about multilingualism and provides a deeper understanding about the structural relations between language communities in Twitter. This research work is one of the few that apply social network analysis to the study of sociolinguistic questions on the Internet. Focusing on the social networks of multilingual users, this dissertation contributes an original classification of network types based on the patterns of connections between language groups. Also, it applies the novel idea of modeling the influence of network factors in the language choices of the user. Finally, this dissertation tests the hypothesis that the type of exchange influences language choice, and explores with a theme analysis how other textual features might elicit cross-cultural awareness. These results can inform the design of social media platforms.

MULTILINGUAL USE OF TWITTER: LANGUAGE CHOICE AND LANGUAGE BRIDGES IN A SOCIAL NETWORK

by

Irene Eleta Mogollón

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Jennifer Golbeck, Chair/Advisor
Professor Benjamin B. Bederson
Professor Jordan Boyd-Graber
Professor Kari M. Kraus
Professor Ira Chinoy

En ese océano que separa los continentes y las vidas,

aún no te he perdido en la tormenta.

A Pepe.

Acknowledgments

First and foremost I would like to thank my advisor, Prof. Jennifer Golbeck, for her inspiring lessons on social network analysis, for encouraging me to develop my own original ideas and guiding me through that difficult process, always caring for my motivation. I have reached this milestone thanks to her support in the moments of adversity. She also made it possible by financing this dissertation research.

I owe my gratitude to the other committee members of my dissertation, Prof. Ben Bederson, Prof. Jordan Boyd-Graber, Prof. Kari Kraus, and Prof. Ira Chinoy, who have provided valuable feedback in their diverse areas of expertise to make this dissertation a more solid and complete research work.

I would also like to thank Dr. Judith Klavans and Prof. Doug Oard for their advice and mentoring in the early stages of my doctoral endeavor. They transformed the graduate student I was into a researcher. It was an immense privilege to count with them.

I would also like to acknowledge help from Tony Rogers, who joined Prof. Jennifer Golbeck and I in our search for multilingual users of Twitter.

My peers and the professors at the College of Information Studies, the *iSchool*, and at the Human-Computer Interaction Lab (HCIL) have enriched my graduate experience in many ways, providing inspiration and support.

I owe my deepest thanks to Fulbright for sponsoring my doctoral studies and for the financial support in the first years of my program. Also, they have enriched my stay in the United States by giving me the opportunity to participate in the many

cultural and social events they organize, including academic workshops, where I have met many new friends. The Fulbright community has had an enormous influence in my vision of the world and in this research work. They are the most inspiring example I know of multicultural social ties between the world's nations.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

> [G]lobalization is characterised by unprecedented flows of information, exchanges among different groups and networks that transcend the local and national [116, p. 9].

As the number of Internet users from different parts of the world grows [58], so does the use of a wealth of languages online [89]. The Internet is not accessed only through computers, but also through cellphones and tablets; this trend is enabling more people in developing countries and speakers of a plethora of languages to access it [58]. While access to the Internet and communication flows are greater than ever before, there is evidence of fragmentation due to language and national borders on the Web [44], and on the blogosphere [53, 47]. Also, many authors warn about the existence of a "linguistic digital divide" that prevents many users of the Internet from having access to relevant information in their languages [78, 67, 68, 8].

In the past years, social media has emerged as horizontal networks of communication, where a complex interplay takes place between mainstream media, journalists, political actors, grassroots activists, citizens and technology [13, 77]. On the one hand, there are powerful social actors shaping the linguistic landscape of the Internet with a top-down approach, like national and supranational institutions, broadcasting media, and companies with interests in transnational business [30]. On

the other hand, users of social networks and content-sharing platforms constitute a counter-power [13], reshaping this linguistic landscape with their contributions.

Social media has enabled valuable social outcomes such as spontaneous organization during humanitarian crisis [98], public denunciations of human rights violations [85], creation of relevant content for communities that are underserved in terms of information on the Internet and in their languages [102], and foreign language practice and participation in transnational interest communities and diaspora communities [100].

Many researchers and media outlets are turning their attention to the microblogging site Twitter. They have realized the potential of Twitter for spreading information of unfolding events in real-time across languages and geographic regions [55, 77]. But how are the news traveling across language frontiers?

In this dissertation, I study how multilingual users of Twitter mediate between language groups in their social network, focusing on social connections and language choice. My long-term goal is to advance our understanding of the network structures and communication strategies that enable intercultural dialog, cross-language sharing of information, and awareness of global problems.

This research goes beyond survey information about multilingualism: I apply social network analysis to gain a deeper understanding about the structural relations between language communities in Twitter. I focus on the social networks of multilingual users and contribute a classification of network types based on the patterns of connections between language groups. Also, I propose and apply the novel idea of modeling the influence of network factors in the language choices of the user.

1.1 What is Twitter?

In Twitter, users share posts with followers; these posts are limited to 140 characters and often include links to webpages, images and other resources. Twitter has characteristics of a social network —although relationships do not need to be reciprocal— and an information-sharing network, where both mainstream media and user-generated content are disseminated publicly [69, 77]. The posts of the people a user follows are laid out in a vertical stream, in inverse chronological order, i.e. the most recent posts are at the top and the user can scroll down the screen to read the previous messages. Twitter posts can be of three types:

1. an original comment by the author;

2. a reposting of a comment authored by someone else, the user that passes the message along can do it either by means of the button "Retweet" or preceding copied text by "RT", "rt", or other markers of attribution (see figure 1.1);

3. a reply or comment addressed to a particular user by means of a "mention", the @ sign followed by a username.

The key to the success and novelty of Twitter is due to the speed of information dissemination and the fact that most of this information is publicly accessible. For instance, it takes a "tweet" less than one hour on average to be reposted and, if it gets beyond that first hop, it will be reposted almost instantly in subsequent hops, reaching an average of 1000 people [69].

Figure 1.1: Two Twitter repostings. The authors' usernames at the top of each message have been erased for privacy reasons, as well as the usernames next to "Retweeted by"; the later users reposted them by clicking the Retweet button. The message at the top was previously reposted copying the text and preceding it by RT and the mention of the original author's username, which is partially shown (@k). Screenshot taken in 2011 reflecting interface design at the time of data collection. For more images and details on the evolution of Twitter's interface until 2011, see [114]: `http://blog.bufferapp.com/how-twitter-evolved-from-2006-to-2011`.

As a consequence, there is an emergent body of research literature studying how to leverage Twitter's tremendous potential for "participatory sensing" and collaboration [28], enabling "situational awareness" in emergency events [105], and functioning as an "awareness system" for journalism [51].

1.2 Motivation

The underlying motivation for my research is promoting language diversity and facilitating access to multilingual information on the Internet, thus everybody can benefit from it for communicating, learning, making business, sharing ideas and resources. However, multilingualism also brings new challenges, like the segregation

of information and communication spheres, which can hinder the potential of the Internet for discovery, cross-cultural awareness, intercultural dialog, and transnational collaboration to find solutions for local conflicts and global problems.

To support my views, I highlight below the relevant points of the "Geneva Declaration of Principles" [115] for an inclusive Information Society, approved at the *World Summit of the Information Society* in 2003:

- The international management of the Internet should facilitate access for all, taking into account **multilingualism**.
- The Information Society should foster and respect **cultural and linguistic diversity**, **dialogue among cultures** and civilizations, and encourage **international cooperation**.
- Everyone should have the right to seek, receive, and impart information and ideas through any media and **regardless of frontiers**.

These principles are based on prior international declarations, such as the UNESCO's "recommendation concerning the promotion and use of multilingualism and universal access to cyberspace" [103].

The Declaration of Principles states the importance of a "rich public domain [...] for the growth of the Information Society, creating multiple benefits such as an educated public, new jobs, innovation, business opportunities, and the advancement of sciences" [115, p. 4]. Similar supporting arguments come from the Internet Society, which is an non-profit international organization that provides leadership

for Internet policy and technology standards [59]. The Internet Society's vision was remarkably conveyed by Vint Cerf in his 1999 speech "The Internet is for Everyone" [14].

Also, I was inspired by Zuckerman's comparison of cosmopolitan cities with the Internet [119]; using this metaphor, he proposed to plan and design technology for creating the structure that fosters social contact, vibrant communities, and discovery, to fulfill the vision of an internet that constitutes a truly cosmopolitan space.

Unfortunately, there are innumerable challenges to achieve these goals that go beyond the technical aspects, such as socioeconomic inequality, lack of infrastructure in rural areas and disadvantaged parts of the world [34, 91], restrictive governmental or private controls [107]. Indeed, the Internet has many types of frontiers and barriers, but I am particularly interested in the language frontier.

1.2.1 Language bubbles?

The first language frontier many potential users encounter prevents them from using the services on the Internet: the interfaces are not localized into their language, writing system, and cultural conventions. There is a wealth of literature, mostly practice-oriented, about localization of interfaces for improving usability and accessibility of software products and websites targeting a global market [117, 50, 92]. Additionally, there are other —more subtle— language frontiers.

Drawing similarities with the "filter bubble" problem, Scott Hale [46] writes about "language bubbles" on the Internet. As an example, he shows the different set of search results obtained for the query "Tiananmen Square" in English and Chinese using the search engine Google [46]. *The filter bubble* [86] was a very discussed book warning the public about the widespread use of algorithms for personalization of search results and news feeds online without the knowledge or control of the end-user [86].

This book, and other works expressing similar concerns, have triggered debate about the decisions shaping the design of information systems and social networks, and the impact they have on society. Maybe as a result of this debate, the recommender systems community is making an effort to incorporate the values of *diversity* and *novelty* into the recommendation models and algorithms [1]. However, as Hale points out [46], system designers might not be taking into account the dimensions of culture and language yet.

A research study of 25 language versions of Wikipedia by Hecht and Gergle [49] serves to illustrate this ignored challenge. Wikipedia is an online encyclopedia built with user contributions and revolves around the principle of reaching consensus on concepts' descriptions. Hecht and Gergle [49] found that more than 74% of concepts in Wikipedia are described in only one language and there is a surprisingly small overlap of concepts in different languages. For instance, in the case of two mature language editions, the authors report that 51% of concepts in English are covered in German, but only 16% of concepts in German are also in the English Wikipedia [49].

One implication is that we are seeing a substantially different knowledge repository depending on the language we use. This is not only a matter of insufficient translation, but of concepts that are culture-specific or not considered relevant in other languages, e.g. city districts, national sport teams [49].

A survey on the topology of web links determined that the number of hyperlinks that cross international borders is significantly lower than the number of domestic hyperlinks [44]. Similarly, the blogosphere is fragmented into language communities [80, 53, 47]. We see a different Internet depending on the language we use, which is hindering our capabilities for sharing and learning, but this research problem remains unexplored for the most part.

Notably, the field of multilingual information retrieval has a solid literature body [87], including the design of multilingual search interfaces and the specific problem of cross-language information retrieval, but is narrowly focused on search. Also, there is a growing body of literature in the Semantic Web field about multilingual ontologies, cross-language linking of data and resources [3], but its application is limited to certain domains.

In general, there are scarce efforts to understand what are the network structures and communication environments that foster intercultural dialog, cross-language sharing of information and resources, awareness of global problems, and international collaboration.

1.2.2 The bridges between the local and the global

Coupland [18] highlights how social relations become possible across distance with the help of Information and Communication Technologies (ICTs) in this time of unprecedented numbers of mobile trajectories and flows of populations. Expatriates, migrants, minorities, diaspora communities, and language learners play an important role in forming transnational networks and cultural bridges between nations and communities.

Many users of the Internet are multicultural and multilingual. They sometimes act as invisible translators. For instance, during casual daily interactions, they might be passing information from one language community to the other, without strictly translating, but re-contextualizing a story in a new language and culture [6]. Some ground-breaking initiatives are already taking advantage of multilingual users' language skills for raising international awareness about local conflicts, human rights violations, and advocacy causes. Such is the case of Global Voices, "an international community of bloggers who report on blogs and citizen media from around the world" [37].

In 2009, the Berkman Center for Internet and Society mapped the Arabic blogosphere and described a key concept that has motivated this dissertation work. They identified English and French "language bridges" on the Arabic blogosphere, consisting of bloggers that wrote in English or French and their native (Arabic) language, which connected the different national blogospheres with the international one [32].

Understanding what is the impact of these "language bridges" and how social media is used for "reaching out to the world" and drawing the attention of international broadcasting media are still open questions of particular interest after the popular uprisings during 2011 [16].

In the microblogging site Twitter, information spreads across languages and countries [75, 76, 77] and, as I will show, this is possible thanks to multilingual users that are mediating between language communities. "[T]he greatest connecting power is the will of the users who want to be connected" [45], like in the example of bloggers in Arab countries connecting with an international audience [32] or the self-denominated "voluntweeters" after the earthquake in Haiti [98].

A world that faces global challenges, could benefit from leveraging the interconnections of its population for finding and sharing solutions from the local level to the international level.

1.2.3 Values in the design of communication platforms

The localization of the interface into a diversity of languages and cultural codes, support for non-latin scripts and bidirectional text displays, as well as providing assistive technologies for translation, are basic requirements for a globally accessible communication platform [117, 50].

The debate about filter bubbles uncovers that information systems, online social networks, content-sharing and communications platforms are not neutral tools.

The values and design decisions that underlie these systems [95] have an impact on the users' perceptions of the world and their behavior.

The design of Twitter and other information-sharing platforms comes with an embedded set of values, like sharing, dissemination, being public and participative, etc. In accordance with these values, research can shed light on how to leverage the language skills and multicultural background of its users to promote dissemination of information across language frontiers.

Even after designers identify the values they want to imprint in the system, they still need to understand the challenges associated. For example, if we want a communication and information-sharing platform that enables intercultural dialog and collaboration, cross-language link sharing, and awareness of global problems, we need to study how the system might be constraining the linguistic decisions of multilingual users and impairing their ability to cross online frontiers. Also, we should acknowledge and respect that, in some cases, certain communities might have reasons for concealing information or resources.

1.3 An Ultimate Goal

The overarching goal that motivates my research is to advance our understanding of the network structures and communication strategies that foster intercultural dialog, cross-language sharing of information, and awareness of global problems. We could leverage this knowledge to reduce the impact of language frontiers online, to encourage social contacts and links to resources across languages, and to promote

the use of multiple languages, i.e. instead of constraining multilingual users to one language choice, empowering them to mediate between cultures.

Ultimately, who are the people and what are the reasons that connect different cultural and linguistic groups? What can we do to foster and leverage these cross-cultural connections for building a cosmopolitan space?

These are very broad and ambitious questions, and my research path has barely started. In the next section, I narrow the scope to provide a founding ground for this area of inquiry.

1.4 Objectives and Research Questions

This dissertation research aims at: (1) exploring the ways in which multilingual users of Twitter are connecting different language groups in their social network; (2) modeling how the network influences their language choices; (3) and exploring what the textual features of their posts can elicit about language choices and mediation between language groups.

This dissertation focuses on the microblogging site Twitter because it constitutes an example of a social and information-sharing network where information is disseminating across languages and countries (see subsection 1.2.2). Also, the interface is available in a diversity of languages, supports various non-latin scripts, bidirectional text, and it does not filter the posts by language. Therefore, it potentially exposes the user to a multilingual conversation if she/he chooses to follow people writing in different languages.

Four questions drive this research:

1. In what ways are multilingual users of Twitter connecting language groups?

2. How is the social network of multilingual users in Twitter influencing their choice of language?

3. Does the type of exchange in Twitter (i.e. public post, reply) influence the language choice of multilingual users?

4. What the themes and textual features in the posts of multilingual users reveal about cross-cultural awareness or international dialogue?

Inspired by an expanded paradigm of Web Content Analysis proposed by Herring [52], this research includes social network analysis, natural language processing for automatic language identification, theme and exchange analyses.

In this dissertation, the research subjects are Twitter users authoring posts in English and at least another language. Focusing on the social network of these multilingual users, the methodology combines a qualitative approach to social network analysis and network statistics to present a taxonomy of network types based on the patterns of intersections and connections between language groups. The resulting theoretical constructs or categories answer the first research question. A factor analysis based on two regression models will answer the second research question on the social network influence in the language choices of multilingual users.

To answer the third research question, I test the hypothesis that the textual feature indicating addressivity (@ sign) within the posts of multilingual users in-

fluences their language choice. Finally, regarding the fourth research question, a generic theme analysis will provide preliminary findings on topics that might help in raising cross-cultural awareness, and on the reasons for using English keywords in non-English posts.

1.5 Contributions and Audiences

The main contribution of this dissertation is that it goes beyond survey information about multilingualism and provides a deeper understanding about the structural relations between language communities in a social network online. In particular, this research proposes new specific network statistics to enhance the definitions of original theoretical constructs: the types of intersections between language groups in social networks.

Inspired by previous studies on the blogosphere, I propose to apply social network analysis to study sociolinguistic questions on the Internet. Adapting the *Ecology of Language* theoretical framework from sociolinguistics to the social network context, this research conceives of the social network of multilingual users as a micro-scale language ecology, influencing their communication strategies and language choices. This conceptualization leads to a second key contribution, which is the novel idea of modeling the influence of social network factors in the language choices of the user.

Other contributions include: the confirmation of previous empirical observations pointing to addressivity as a factor for language choice in Twitter, the iden-

tification of themes that might be raising cross-cultural awareness, and the identification of certain types of hashtags (keywords preceded by the # sign) and related contexts that could encourage multilingual conversations.

Regarding the audiences that this dissertation addresses, the research area of information diffusion in social networks could benefit from the findings about the structural relations between language groups. More broadly, this work is relevant to the fields of Information Studies and Social Informatics.

The lessons learned in this work could inform the design of socio-technical systems. This research contributes to "understanding users" in the field of Human-Computer Interaction, especially, in the areas of computer-supported cooperative work and technology-mediated social participation. Also, this dissertation might be of interest in the field of Language Technologies for potential applications.

This dissertation was inspired by works in Computer-Mediated Communication and constitutes another example of applying social network analysis in this field, which is still rarely used.

Finally, other audiences include the fields of Digital Humanities, Sociolinguistics, and Communications Studies, especially in relation to the Internet. Researchers in these areas might find inspiration in this dissertation for exploring their research questions with the new lens of social network analysis, and the use of automatic language processing.

Chapter 2

Theoretical Framework

I begin this chapter by reviewing relevant theoretical perspectives from Sociolinguistics in the context of globalization. The *Global Language System* theory proposed by De Swaan provides a macro-scale perspective on language; it was later reinterpreted by Calvet in his *Ecology of World Languages*, which includes an amalgam of views conforming the Ecology of Language approach. This approach comprises different levels of analysis in the study of languages: the macro-scale language dynamics, described as language ecologies, emerge as a result of the interactions of individuals, and their language choices at micro-scale level. However, Sociolinguistics theories and methods remain too fragmentary.

Inspired by previous studies on the blogosphere, in this chapter I propose to apply social network analysis to study sociolinguistic questions on the Internet. Social network analysis enables us to understand the influence of micro-scale interactions into macro-scale social dynamics; this analytical approach could enrich the Ecology of Language perspective.

Finally, narrowing the scope to the particular environment of the Internet and social media, I discuss and define the concepts that relate to interactions of users and technology, with particular attention to multilingualism, language choice, and *language-switching*.

2.1 The Global Language System

De Swaan's theory of communication potential and language competition, called the "Global Language System" [26, 25, 27], illustrates the dynamics of the world's languages with a constellation metaphor: English constitutes the hypercentral sun of the global constellation of languages. At the top level, supranational subsystems —like Spanish, French, and Arabic— compete with English as languages of global communication. There are a dozen supranational languages in this constellation and a hundred national languages orbiting around them like planets. This pattern appears at different levels in the system, starting in the periphery with local languages surrounding a national language like the satellites of planets. The central languages in each subsystem or cluster have a mediation function between local languages.

A key point in De Swaan's theory that is relevant to this research is the connection between the language groups through polyglots and interpreters. Multilingualism and translation constitutes the gravitational force that provides cohesion to the system, enabling communication and interaction between different language groups. At the same time, speakers are confronted by multiple and competing linguistic options. Individual and collective choices shape the system, but are themselves influenced by the spheres of politics, economics, and culture [25].

De Swaan [26] proposes a formula to determine the communication potential of a language, which could influence the decision of people to learn it. The factors that the formula takes into account are: the number of speakers of the language and

the number of multilingual speakers that know the language. The number of multilingual speakers is related to the centrality of the language in the system. These multilingual speakers increase the communication potential value of a language because they enable connections with other languages in the system. For instance, in the language subsystem of the European Union (E.U.), German has a high communication potential value due to the large number of speakers in the region, but English has a higher value due to the lager number of multilingual speakers competent in English, which provides the opportunity to communicate with people from many different countries in the E.U. [27].

De Swaan already mentions network concepts that will be introduced in section 2.3: the "centrality" of a language in the system, which accounts for the mediation potential between that language and others thanks to the number of polyglots speaking it; these polyglots are facilitating "linkages or connections" among languages, which are necessary for the "cohesion" of the system.

The constellation metaphor serves to illustrate the following concepts: *languages of local communication* (described as "satellites" or "the periphery"), *languages of regional communication* (illustrated as "planets"), and *languages of global communication* (represented as the "central stars"). When considering the linkages between language groups and the communication potential, the following terms are used given a particular context or region: a *vernacular language* is the first language of the majority of its users, a *lingua franca* is the second language of the majority of its users, who speak different first languages, and in the case of a *vehicular language*

there is a balance between the number of speakers that use it as a first language and the number of speakers that use it as a second language [2].

2.2 The Ecology of Language

While De Swaan's theory places English as a hyper-central sun and uses socio-economic concepts to assign a "value" to languages, the Ecology of Language approach adds nuanced views regarding language hierarchies. For instance, this approach addresses the phenomena of "ethnic revivals" as a form of counter-power in the language dynamics of a globalized world [56, 30, 13], where the socio-economic approach of De Swaan falls short.

Borrowing concepts from the field of Ecology, Haugen introduced the Language Ecology approach and the notion of the co-evolution of languages and their interdependence within a social system [19, 71, 48]. Hornberger [56] synthesizes this analytical approach as taking into account all languages in a given ecosystem, recognizing their social spaces and contexts. Adapting the Ecology of Language approach to Twitter, this research conceives of the social network of multilingual users as a micro-scale language ecology, influencing their communication strategies and language choices.

Under the ample umbrella of this approach and its various names (language or linguistic ecology, ecology of languages, and ecolinguistics), there is a diverse group of authors and works painting the multifaceted social, political, and cultural reality surrounding the languages of the world. However, this painting is, for the most part,

fragmentary in nature. While the works of the Ecology of Language generally focus on micro-sociology problems, such as managing multilingualism in South African and Bolivian classrooms [56], De Swaan proposes a planetary vision. In essence, the difference between the two perspectives is based on the scale of the analysis.

Calvet provides an integrative perspective in his book *Towards and Ecology of World Languages* [11]. Calvet reinterprets De Swaan's theory as the "gravitational model", which describes the global ecosystem of languages, or the macroscopic scale, and complements it with other models that account for phenomena in lower scales, such as internal regulation of languages, social and official functions of languages, and identity [71]. In this way, the Ecology of Language does not exclude the Global Language System theory by De Swaan, but embraces it as part of this inclusive approach.

In summary, the Ecology of Language comprises various levels of study in sociolinguistics, from microscopic to macroscopic scale: from the individuals interacting to the population, and ultimately, the ecosystem of society, policy, economics, and communication media [71]. However, this framework lacks a connecting tool between those levels of analysis, i.e. how global language dynamics emerge from individual interactions and language choices. Inspired by *Language Networks on LiveJournal* [53] —one of the first studies in taking a network analysis approach to study languages in social media—, I propose to apply social network analysis to facilitate our understanding of the connections between micro-scale and macro-scale sociolinguistic questions.

2.3 Networks

> Building on a long tradition of network analysis in sociology and anthropology [...] and an even longer history of graph theory in discrete mathematics [...], the study of networks and networked systems has exploded across the academic spectrum [109](243).

The science of networks provides a new framework to understand complex systems in biology, sociology, communication technologies, business, etc [7]. Network structure is "thought to influence individual (micro) and collective (macro) behavior, as well as the relationships between the two", and has provided useful insights in the study of the spread of disease and information dissemination [109](256). Similarly, networks could help us gain a deeper understanding of language use in society, in communication technologies, and of language effects on information dissemination.

2.3.1 Concepts of Social Network Analysis

Social network analysis (SNA) studies the network structure of relations between people to understand social phenomena, instead of categorizing human behavior based on individual inner forces [80]. In SNA, people are represented as *nodes* of a social graph. The nodes are connected by *edges*, or social *ties*, that could be reciprocal or just a one way relation, like the "follower of" relation in Twitter.

In this work, I will use an important type of subgraph: the *egocentric network*. The egocentric network is obtained by selecting an individual node, called the *ego*, and all of its connections [38]. In other words, it constitutes the personal social

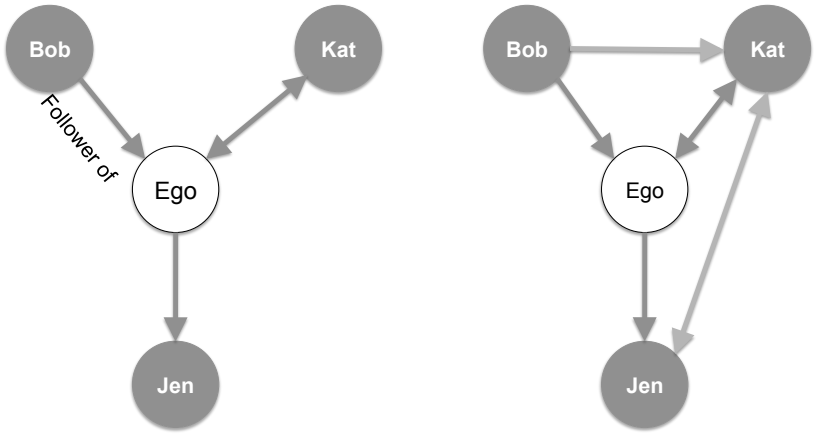


Figure 2.1: Egocentric network with degree 1 (left) and with degree 1.5 (right).

network of an individual with his or her contacts. An egocentric network that includes only the connections with the ego has degree 1. More frequently, researchers are interested in including the connections among the ego's contacts, in this case the egocentric network has degree 1.5 [38]. Figure 2.1 illustrates these basic concepts. The egocentric network has become a standard unit of measurement for studying small scale interactions (or micro-sociology) [41, 80].

An edge that constitutes the only connection between two groups of nodes is called a *bridge* [38], which is a mathematical concept. Bridges are of special importance for the dissemination of information from one group to the other [41, 42].

In this context, *gatekeepers* are people enabling communication between two groups. Multilingual users of Twitter might be in a position of their social network where information necessarily has to pass through them to reach the other language group. This has several implications: the gatekeeper's role is critical for spreading news between communities and for rising cross-cultural awareness, but they could

broker information to their advantage [80], or they could be conservative in their decisions to transfer information to one group for cultural reasons [82]; in either case, places where we find gatekeepers could be considered *structural holes* between communities.

Communities or *clusters* are composed by nodes that are more connected to one another than with the rest of the social graph. For example, imagine a town with different neighborhoods, where almost everybody that lives in the same neighborhood knows each other, but they know fewer people from the other neighborhoods. There are a variety of automatic methods to detect clusters or communities based on network structure [38]. Figure 2.2 shows a schematic view of how people are connected in clusters. In Twitter, clusters form due to language, geography, or topic of interest, but this research focuses on language clusters.

The *cohesion* of a social graph is a count of the minimum number of edges that prevent the entire graph from breaking in isolated components [38]. These types of edges that connect clusters of people are critical for providing cohesion to the society, building a sense of community, and for effective self organization and collective action across language groups [41, 42].

Centrality is a core concept in SNA, and measures how "central" a node is in the network to estimate its importance [38]. There are different centrality measures that account for various reasons why a node might be important. For instance, *degree centrality* counts the number of edges or connections a node has. In this work, and in De Swaan's theory (explained in section 2.1), the relevant centrality measure is *betweenness centrality*, which captures how important a node is in the

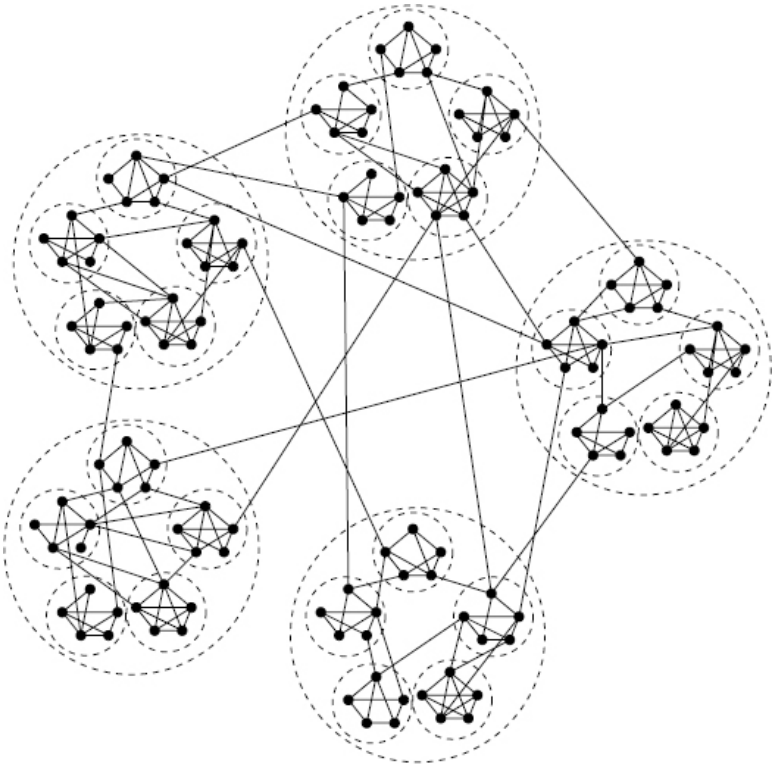


Figure 2.2: Schematic view of a network with clusters. Clusters (or communities) are composed by nodes that are more connected to one another than with the rest of the social graph. This dissertation focuses on language clusters. This visualization is an extract from a open access journal article [63]: M. Kaiser, M. Görner, and C. C. Hilgetag. (2007). Criticality of spreading dynamics in hierarchical cluster networks without inhibition. *New Journal of Physics*: `http://iopscience.iop.org/1367-2630/9/5/110/fulltext/`

flow of information from one part of the network to another [38]. In other words, betweenness centrality reveals the potential mediators or gatekeepers.

2.3.2 A network perspective on Sociolinguistics

In this dissertation, I focus on multilingual users of Twitter as mediators between language clusters. For example, imagine a user posting in English and Spanish. In Twitter, she follows the updates of researchers posting in English, but also the Twitter accounts of Spanish local media. Her friends in Spain are connected with her in Twitter, and also her colleagues in the United States. This user sometimes posts in Spanish commenting on local news or replying to some Spanish friend. Often, she posts in English to disseminate research content. She might post in English to draw international attention about important events in Spain.

This dissertation hypothesizes that the language choices this user makes every time she writes a post will be influenced by the language composition of her social network and, in turn, will have an impact on it. In section 2.2, adapting the Ecology of Language approach to Twitter, I propose to conceptualize the social network of multilingual users as a micro-scale language ecology, influencing their communication strategies and language choices.

Also, this dissertation investigates what is the structural relation between the language clusters in the social network of this user to develop methods for detecting gatekeepers or structural holes. Future research on information dissemination could benefit from these methods that account for the language effect.

An early work that inspired the question of multilingual users as mediators, at the micro-socioloy scale, was the research on gatekeepers by Metoyer-Duran in a variety of ethnolinguistic communities in the United States (American Indian, Chi-

nese, Japanese, Korean, and Latino) [82]. She studied their profiles (multilingual and multi-literate), their behavior as information providers in their respective communities and how they utilize their interpersonal network and new technologies [82]. Her study identified the profiles that facilitated access to information resources for underserved communities.

At the macroscopic scale, giant clusters in the social network might represent language communities at international level, in some cases roughly corresponding with national borders. For illustrating this idea, I include a visualization of European language communities in Twitter from a recent study [83]; in figure 2.3 the dots represent Twitter posts with geolocation information and the colors differentiate the languages of the posts. However, at the micro-scale, there are pockets of expatriates and diverse ethnolinguistic communities (similar to those studied by Metoyer-Duran) immersed in these giant clusters, where multilingualism is present.

A pair of language communities that share more connections through multilingual individuals or translations than other pairs would have a "communication highway" between those two languages. Continuing with the metaphoric theme of roadways, if only a few multilingual gatekeepers and translations connect both language communities, there would be a "rope bridge" crossing the structural hole. However, these relationships between languages are rarely reciprocal, in other words, the communication highway might be one way only. For instance, statistics on literature translation in Europe reflect the dominance of English as a source (origin of the translation) and, in the other extreme, low percentages of non-European lan-

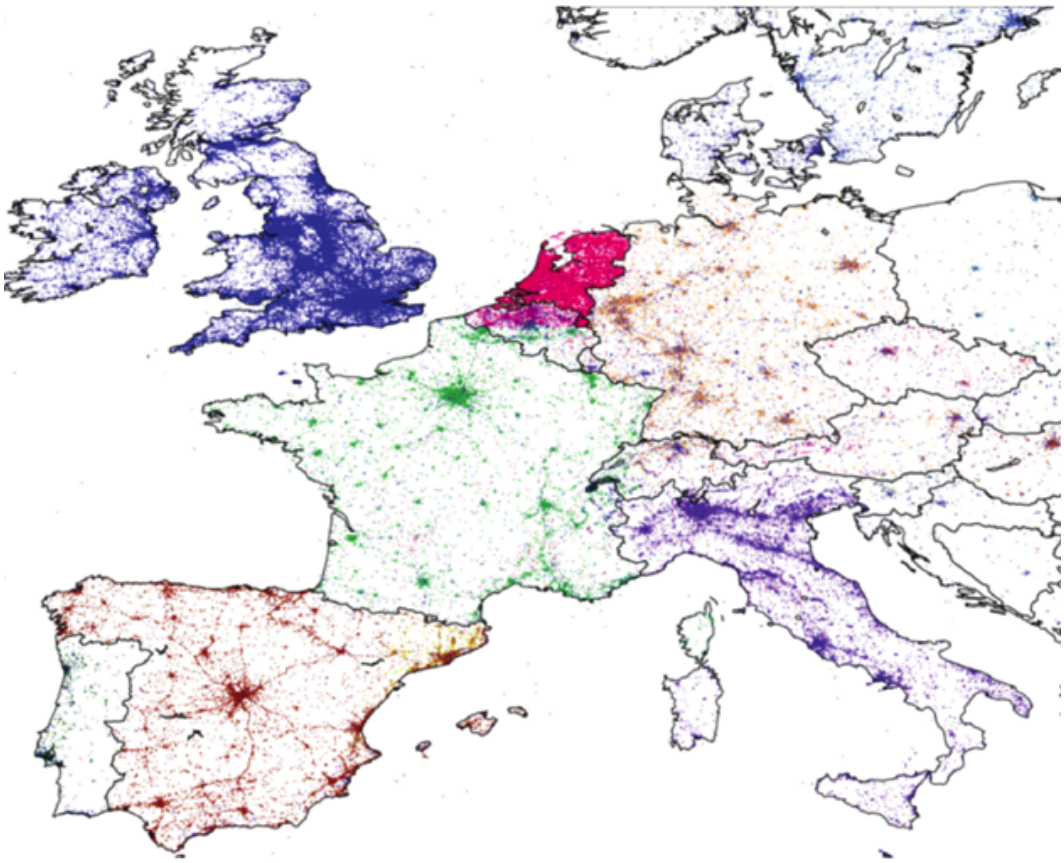


Figure 2.3: European language communities in Twitter. The colored dots represent Twitter posts with geolocation information and their colors differentiate the languages of the posts. Giant language clusters roughly overlap with national borders at the macro-scale. However, at the micro-scale, there are pockets of expatriates and minority languages immersed in these giant clusters. This visualization is an extract from a journal article published under creative commons license [83]: Mocanu D, Baronchelli A, Perra N, Gonalves B, Zhang Q, et al. (2013) The Twitter of Babel: Mapping World Languages through Microblogging Platforms. PLoS ONE 8(4): e61981

guages as a source [74]. For this reason, studying language choice is important to understand the directionality of information flows.

Network science connects the microscopic scale (eg. the multilingual user interacting with her network and switching between languages) and the macroscopic scale (eg. the language dynamics in the Twitter system). The problem of modeling language competition exemplifies the connection between the microscopic and macroscopic scales in sociolinguistics.

When looking at the different outcomes these models predict over time, there are cases of language death, language dominance, language coexistence, language fragmentation into multiple languages, which reminds of the Language Ecology approach. Vazquez et al. [104] described the idea underlying these models: collective social phenomena are studied in terms of interacting agents, which are represented as nodes in a network of social interactions; nodes can change their language according to specified rules of interaction with the neighbors in the network. The models include probabilities to switch languages determined by the local density of speakers of the opposite language, prestige of the language and other parameters [104].

The rules of interaction and probabilities to switch languages belong to the microscopic level, but the simulation of interacting agents generates macroscopic results. For example, the Bilinguals Model with three types of people —speakers of language X, speakers of language Y, and bilingual speakers— shows how the social structure influences the final outcome: the lower the cohesion of the network, the higher the chances of evolving into one dominant language [104].

Using social network analysis in Sociolinguistics is not straightforward. Representing people as nodes requires careful thinking for assigning individual attributes, like the languages they understand, the languages they use, level of language competence and literacy, race or ethnic identity, genre, etc. Additionally, nodes can have a location attribute, which overlaid on a map can distinguish the expatriates and migrant populations. Also, nodes can represent other actors in society, like organizations and media outlets. The edges connecting the nodes might be face to face interactions, interactions mediated by technology (like phone or email), affective or affiliation relations, to name a few examples. Ultimately, the socio-linguist has to conceptualize and interpret what the network measures —like centrality and cohesion— reveal.

2.4 The Internet as a Sociolinguistic Ecology

There is an ongoing debate about the dominance of English on the Internet and its impact on language diversity [20, 39, 33]. The United States' leading role in developing the Internet had consequences like the initial use of English only, protocols devised for the Roman alphabet, and a telecommunications infrastructure that was economically dominated by U.S. companies [33].

However, the Internet is evolving very fast. As other nations started to come into play, and users of different countries gained access to the Internet, a wealth of languages blossomed online [89]. At the same time that online content was increasing exponentially, the percentage of English content diminished to 45% in 2005,

in favor of other languages, while the estimated online content in Chinese grew to 9% in 2008, followed by German and Spanish [89]. The UNESCO's "recommendation concerning the promotion and use of multilingualism and universal access to cyberspace" [103] and new standards like Unicode, enabling the use of different written systems, intensified the trend towards a multilingual Internet.

Despite this progress, non-English speaking users perceive the scarcity of online resources in their first language and are generally appreciative when they can find information in their language [8]. If users have sufficient knowledge of English as a second language, they might search in English because they perceive there is more content in this language and of better quality [8].

Organizations interested in transnational business have realized the importance of adapting to local cultures to be competitive in a global economy [30]. They are translating and localizing (adapting to the culture) their products and services on the Internet. Dor [30] warns against leaving the standardization of vernacular languages in the hands of software, media, and advertising industries, in detriment of the users key role on language change, identity, and maintenance.

Even though this preoccupation is well founded, when Dor wrote his article the participatory Web was still in its infancy. Recently, the wide array of content-sharing and social media platforms, blogs, wikies, and social networking sites that conform the so-called "Web 2.0" has lowered the barriers for users to become producers of content too [6]. The social networking site Facebook broke with the top-down approach of language standardization in interface localization and implemented one where users seek consensus about the translation of terms in the interface [73]. How-

ever, the model of inviting users to translate the interface of a site is not transferable to every company.

The Wold Wide Web relied on the information retrieval paradigm, were users search and read content generated by institutions, organizations, broadcasting media, etc., while interpersonal communication happened via email, Internet Relay Chat (IRC), and newsgroups [6]. With the advent of Web 2.0 environments, which encouraged participation and sharing, there was a paradigm shift. Users have become consumers and producers of content at the same time, blurring the boundaries between professional and user-generated discourse, individual and collective authorship, and various communication modes co-existing in a single platform: personal messages, instant messaging or chat, public posts, etc [6].

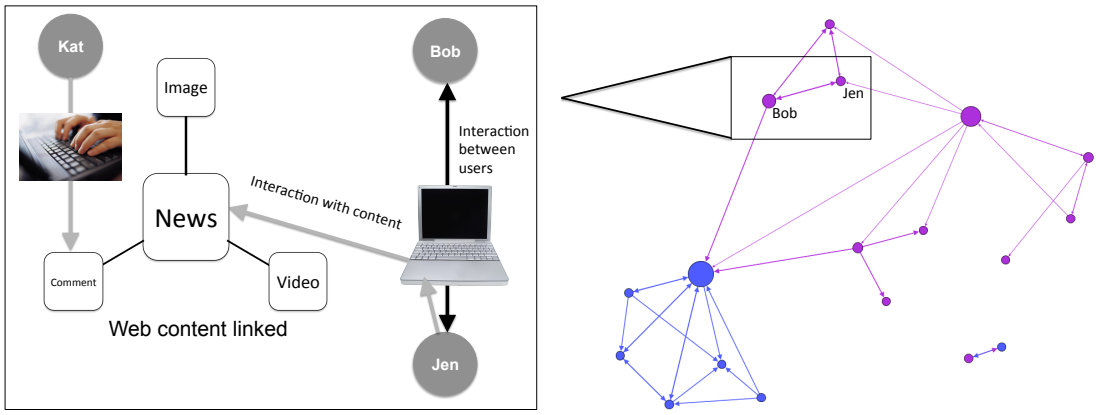
Thanks to the changes brought by the participatory Web, there is a growing body of literature documenting the increased visibility of vernaculars [6], the creation of relevant content in minority languages [102], and foreign language practice and participation in transnational interest communities and diaspora communities [100]. Other studies in the field of computer-mediated communication focus on the reproduction in written form of patterns associated with spoken language, the use of slang or dialect features, playful uses of orthography and typography [23], and describe the informal adaptations to the Roman alphabet of languages with other writing systems, like Arabic [108]. These characteristics of the written language in social media pose a challenge for the automatic analysis of text, which I will discuss in detail in chapter 4.

2.4.1 The mediation of technology and the cosmopolitan space

Instead of just thinking of a global language and its impact on local ones, Androutsopoulos [6] proposes to direct our attention to the circulation of cultural artifacts across national and ethnolinguistic borders and how social media platforms enable the negotiation of local responses, and the appropriation of those artifacts in new socio-cultural environments.

In these participatory environments, a network of users interacting with other users and with digital resources emerges. Digital resources like videos, still images, speech, music, and text can be labelled (tagged) by users, who are collectively building taxonomies [88], or even creating multilingual knowledge repositories like *Wikipedia* [49]. Very importantly, users are now finding information through social recommendations or cues (like tags) left by other people [88]. As a result of this overlapping networks of content and users, information and resources circulate in different ways across countries [6].

In the sociolinguistic ecology of the Internet, interactions between users are constraint by the mediation of technology [6]. The design of keyboards, displays, interfaces, standards that support writing systems, and features of communication platforms have an impact on the users' language choices and translation behaviors. As explained in section 2.3, language choice can affect the directionality of information flows between language groups. In social media, social interactions are not as clearly delimited from interactions with content, since user comments on a digital object (text, photo, video), and repostings, also constitute an interaction with



(a) Interactions mediated by technology.

(b) Bilingual social network.

Figure 2.4: **(a)** Focuses on a few users, Jen, Bob and Kate, who are interacting between them and with Web content through the mediation of technology. Two networks overlap: the network of digital objects that are interlinked and the social network of users. And **(b)** illustrates the social network to which Jen and Bob belong. Pink nodes represent people who use English, blue nodes represent people who use Chinese, and the edges represent the"follower of" relationship in Twitter.

the user that posted it. Figure 2.4 illustrates technology and network structure conditioning users' communication strategies.

Zuckerman [119] used the metaphor of cities to provide a vision of an internet that aspires to be a cosmopolitan space, enabling the contact with the unfamiliar, the serendipity that propitiates learning. In cities, urban planning can create the structure for social contact, vibrant communities, and discovery [119]. An urban planning for a vibrant language ecology on the Internet has to challenge the existing structure of the network of hyperlinks and social connections, and consider the capabilities for sharing multimedia, the length of text permitted, the language technologies available, the functionality for managing audiences, the flexibility for users to reinvent purposes and adapt content.

Inspired by this vision, I focus on the problem of the social network structure in multilingual egocentric networks and on the factors influencing the flexible language choices of multilingual users.

2.4.2 Overview and remarks

In summary, the sociolinguistic ecology of the Internet is determined by powerful social actors like national and supranational institutions, broadcasting media, companies with interests in transnational business, and also by the contributions and interactions of users in social networks, content-sharing platforms, blogs, wikis, etc. At the microscopic scale, the interactions of users are mediated by technology, constrained by it and the network structure. At the macroscopic scale, the Internet is facilitating transnational communication, the flow of information and digital artifacts across language and national borders, and language learning. The growing language diversity of the Internet seems to be enabling access to information in minority languages and encouraging participation across a wider spectrum of society.

However, the flows, access and participation are hindered due to socioeconomic reasons, reduced bandwidth and lack of infrastructure in rural areas and disadvantaged parts of the world [34, 91], or even by governments that purposefully seek to maintain their nation constraint into an isolated information and communication sphere [107].

2.5 Micro-Sociology Focus: Conceptualizing Multilingual Users and Language Choice in Twitter

Adapting the Ecology of Language approach to the social network context, this dissertation focuses on the social network of the multilingual user, conceptualized as a micro-scale language ecology influencing the user's language choices. As an application of this conceptualization, I propose the novel idea of modeling the influence of social network factors in the language choices of the user. In the rest of this section, I describe key concepts underlying this research related to multilingualism, language choice and mediation in the context of Twitter.

Mediators. In section 2.1, I explained the importance that De Swaan gave to multilingual speakers, who are linking language groups and providing cohesion to the system of languages [25]. In section 2.3, I introduced the work of Metoyer-Duran focusing on gatekeepers of ethnolinguistic communities, where she described them as being multilingual and multi-literate [82]. In the context of the Internet, Androutsopoulos [6] highlighted an additional condition to become a mediator between global resources and local audiences: adequate technology access and competence. Bringing these characteristics together, we can draw a very basic profile of mediators between language groups in Twitter: multilingual, multi-literate, and technologically literate.

There are many degrees of language competence and literacy, some users might understand a second language but are unable to speak it or write it; others abstain from using one of the languages they know in certain contexts, like documented cases

of Internet users who preferred to search in English instead of their first language [8]. A discussion on the various degrees of bilingualism, and what Hornberger called "the continua of biliteracy" [56], falls outside the scope of this work. For a comprehensive discussion and classification of the types of bilinguals, consult *The Bilingualism Reader* [111]. In this dissertation, I focus on the multilingual users that write in at least two languages on Twitter.

Language choice. A recurrent theme in the Sociolinguistics literature regarding multilingualism is *language choice*, or how multilingual speakers make decisions about which language to use in each situation and interaction. These small scale decisions have an impact on the global dynamics when aggregated. Simplifying the outcomes of numerous research studies, Androutsopoulos [5] identified *setting*, *participants*, and *topic* as the main factors influencing language choice in bilingual online communities. Other works that I will review in this section and in chapter 3 highlight the influence of the *audience*, the social context, prestige of a language, identity, etc. Figure 2.5 represents the main factors theoretically influencing the language choices of multilingual Twitter users. One of the contributions of this dissertation is to study, for the first time, social network factors in language choice.

Code-switching. A term that frequently appears associated to language choice and bilingualism is *code-switching*. In this work, I use the definition of Joshi [62], which considers two types of code-switching: *intra-sentential*, when the user alternates from one language to another within the same sentence, and *inter-sentential*, when the change of language happens at the same time that the sentence finishes and a new one starts. When studying code-switching, we need to identify

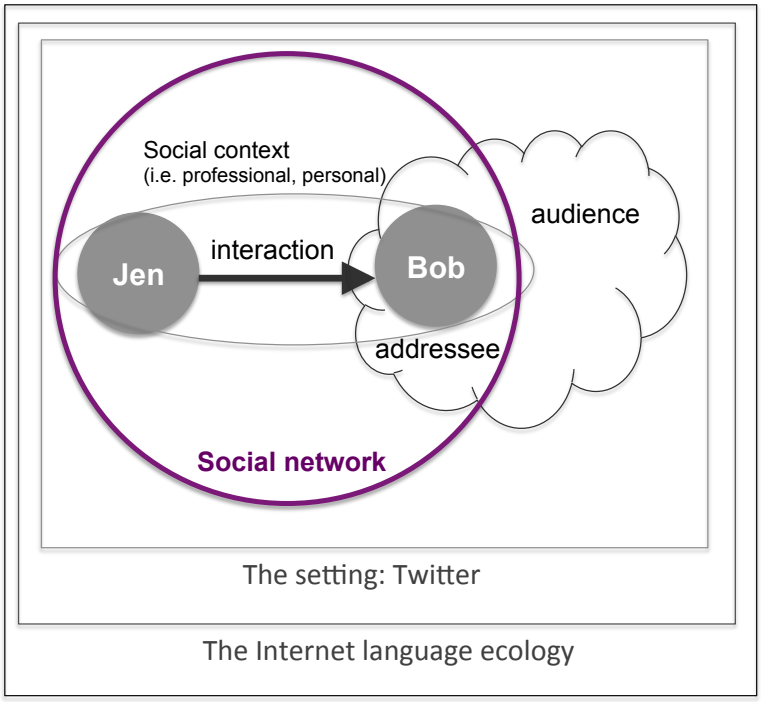


Figure 2.5: Aside from topic and identity, important **factors influencing the language choices** of a multilingual Twitter user are: the addressee in the interaction, the imagined audience, the social context, and the social network. Also, the setting and the Internet influence language choice due to the language ecologies associated. There is an overarching global language ecosystem.

the *matrix language*, which usually provides the grammatical structure and more lexical items, and the *embedded language* [62].

In Twitter, posts are so short that we could consider inter-sentential code-switching when the language changes from one post to the next, while bilingual posts would be cases of intra-sentential code-switching. In chapter 6, the factor analysis represents language choices of the users as counts of inter-sentential code-switching. In chapter 7, the theme analysis includes cases of intra-sentential code-switching, where there are embedded English keywords in other languages.

The setting. Although not directly addressed in this dissertation, it is important to acknowledge the setting as an underlying factor for language choice. In this work, the setting is the Twitter platform and is characterized by its design features, like the limitation of text to 140 characters in every interaction, the default mode of communication being public, the languages available to interact with the interface, the display of messages posted by other users, the possibility to share links, the asynchronous nature of communication, the features for users to manage their social network, etc. Also, conventions and social uses of Twitter have emerged over time among its users [66].

The Twitter setting has a specific language ecology derived from sociopolitical factors. Twitter is a company based in the United States, which has an impact on its adoption across the world, or lack thereof in certain countries like China [83]. Since the micro-blogging service launched in 2006, it was rapidly adopted in many countries; as early as 2007, Java et al. [60] reported about its international adoption in North America, Europe, and Asia (mainly in Japan), but they estimated that 45% of the social network lied within North America.

Not surprisingly, English became a dominant language, with various estimates ranging from 51% of posts [55] to 53% [90]. A recent large-scale study selected the 20 most active countries in Twitter and showed the percentage of English use in each country against the percentage of their corresponding vernaculars [83], illustrating the weight that English has in communications via Twitter. Also, Poblete et al. [90] selected the 10 most active countries in Twitter and unveiled that the U.S. was

the country that concentrated more connections from overseas. For these reasons, I selected multilingual users who have English as one of their language options.

The participants or interlocutors. The participants or interlocutors in an interaction can vary from a one-to-one exchange in an online chat to a one-to-many question posed in a forum for an entire community. In the micro-blogging site Twitter posts are generally public, but there are also posts addressed to specific individuals, and the possibility to send private messages.

The audience. In Twitter, there are different levels of reach a user could have. First, the posts addressed to one or few individuals could be seen by common friends, and potentially found in search results of the platform by others; second, public posts can be seen by the network of people that "follows" the user, and potentially, these posts can reach anyone in Twitter. Re-using a theoretical framework from the field of communication, Johnson classifies Twitter audiences as addressees, auditors, over-hearers and eavesdroppers [61]. Marwick and boyd [81] described the concept of the "imagined audience" in Twitter, or how the user conceptualizes his or her audience to be able to make linguistic choices, even though the real audience that reads the post might be different.

In chapter 7, I test the hypothesis that addressing a message to one interlocutor or a public audience influences the choice of language of the multilingual user.

The egocentric network. Wether Twitter users address posts to the members of their social network explicitly or not, in this dissertation I argue that the egocentric network has an impact on the choice of language. Not only it could have an influence as a perceived audience, but also as a source of information.

Topics or interests. Java et al. [60] identified Twitter communities based on the social network structure and, analyzing the words in the posts, they observed the common topics or interests that differentiated the communities. In the context of the Internet, where the perception of distance and territory blurs, the experience of identity becomes multi-layered. In addition to ethnic identities, there are dimensions of shared "feelings", "knowledge", or "activities" across distance [18]. Because of this multi-layered identities, the user can belong to different communities and choose the language accordingly. This dissertation includes preliminary work related to topics in the theme analysis (chapter 7), but a complete analysis of topics as a factor for language choice is left for future research.

Other factors. There are other important factors surrounding the multi-faceted reality of language choice online. Kelly-Holmes [65] argues that the *prestige* and international importance of a language encourages its use online. Also, language choice could relate to the availability of online resources in a language, or lack thereof [65]. The *social context* of the interaction is another factor, for example, English being used for professional emails and a vernacular language for personal communications [108]. As mentioned above, *identity*, as a marker of social and cultural differences, play an important role in language choice [108, 5]. However, this dissertation does not include them in the analysis to keep this research work to a reasonable scope.

Chapter 3

Related Work

This chapter comprises a review of the literature informing the present research work in one or more of the following themes: language choice and code-switching on the Internet, a network approach to language on the Internet, and multilingualism on Twitter.

This dissertation contributes a classification of network types based on the patterns of connections between language groups, which goes beyond survey works about multilingualism on Twitter. I used a network approach after being inspired by works analyzing language networks on the blogosphere. The literature about language choice on the Internet serves to frame my novel proposal of modeling the influence of network factors in the language choices of the user. Also, the literature about language choice is relevant to chapter 7, where I test the hypothesis that English is used more in public messages than in replies to individuals. In the theme analysis, I include cases of code-switching.

3.1 Language Choice and Code-Switching Online

A survey gathering answers by 2267 students in high schools and universities of eight countries (France, Italy, Indonesia, Macedonia, Oman, Poland, Ukraine, and Tanzania) revealed the complex reality of language choices of Internet users [65].

Most of the participants reported some knowledge of English in addition to their native language, or language of education, and often, they also reported competence in a third language. The study found that bilingual or trilingual Internet sessions were somewhat frequent, that language choice could relate to the availability of online resources in a language, or lack thereof, and to the prestige and international communication potential of a native language [65]. Other research studies support the observation that the perceived scarcity of online resources in a native language influences behavior and attitudes of its users when searching online [8, 67].

Also, domain knowledge influences language choice in online searching because higher expertise on a topic facilitates understanding of relevant texts in a second language [67]. Combining the factors topic and context, there are reported cases where looking for international news and doing academic work encourages the use of English, while personal communication is conducted more often in native languages [65].

Along the same lines, a study on email and Internet chat in Egypt documented the use of English for professional emails, while in personal emails and chat users preferred a romanized form of Egyptian Arabic, which was mostly used orally before the advent of the Internet [108]. This informal transliterations include the numbers 2, 3 and 7 for rendering additional phonemes from Arabic into the Roman alphabet [108]. On the other hand, the use of Classical Arabic and Arabic script was somewhat relegated [108]. Informal transliterations pose a challenge when doing automatic text analysis, as I will explain in detail in chapter 4.

The same study observed cases of code-switching between English and romanized Egyptian Arabic; the later was used for greetings, humor, sarcasm, food, holidays and religion [108]. A case study on code-switching between English and Spanish, and English and Indonesian in Internet chat provided evidence of borrowed English terms related to computers, such as "e-mail", "attachment", and "PC" [12]. When chatting about friendship and relationships, the subjects preferred their first language instead of English [12]. Another study on the language choices of Greek, Turkish and Persian diaspora online communities in Germany found that topics on politics and technology disfavored the use of home and minority languages in newsgroups and web forums, while music and poetry favored it [5].

These works studying topics and contexts (i.e. professional versus personal) as factors for language choice provide some basis for the theme analysis in chapter 7 and for future research about the influence of interest communities in the language choices of Twitter users. The remaining works that I review in this section study the selection of English in multilingual settings online.

A longitudinal study of a Swiss forum with participants of three different native languages detected the increase in the use of English over time, even though English was not the first language of any of the users [31]. This finding suggests that the presence of a multilingual or international audience might encourage the use of English as a *lingua franca*. In view of this previous finding, I propose a multilingual index of the social network as a potential predictor of English use by the multilingual user (chapter 6). Other research works on email and mailing lists

[31, 65] studied the impact on language choice of addressing a message to one person or to a multilingual audience; in the later case English was preferred.

Recent works on the use of social networking sites by bilinguals argue that the intended audience determines the language choice. In Facebook, Welsh-English bilingual high school students write the status updates more often in English to ensure that all their friends feel included, whereas they use Welsh for one-to-one messages with other Welsh-speaking friends [21].

In Twitter, Welsh-English bilinguals use proportionally more English in public posts (53%) than in replies to individuals (44%) in a sample of 500 posts [61]. The reason for using Welsh or English in replies is related to the language profile of the addressee to some extent; sometimes English is used to communicate between Welsh-English bilinguals [61]. In relation to this finding, Johnson [61] speculates that the use of English is encouraged in Twitter for its potential to reach a wider audience. Finally, the later study on Twitter reported very few cases of bilingual posts [61]. In chapter 7, I test that English is used more in public messages and I observe that bilingual posts are scarce.

As a closing note to this section, I would like to acknowledge an existing body of literature about "context collapse" in social media, and particularly in Twitter [81] and Facebook [106]. Professional and personal contexts merge in the same communication environment, where users seek to balance the different identity presentations [81] and, as a result, possibly their linguistic choices.

3.2 Networked Languages

The literature on multilingual computer-mediated communication is very helpful for raising awareness about the multifaceted reality of language choice and how the context and the mediating technology influences it, but generally does not address the potential transnational impact. *Language Networks on LiveJournal* [53] was possibly one of the first studies in taking a network analysis approach to study the language demographics of a social media site, a blog hosting service in particular. Apart from studying the robustness of non-English language networks, they identified blogs that were bridging language communities and described some characteristics of their authors (students of foreign languages, expatriates, multilingual and multicultural) and topics (images, content with international appeal).

Two years later, the Berkman Center identified English and French "language bridges" on the Arabic blogosphere, consisting of bloggers that wrote in English or French and their native (Arabic) language, and connected the different national blogospheres with the international one [32]. However, they did not explore further into the connections with the international blogosphere or their motivations for language choice. These questions are important to understand how people draw international attention using their transnational networks and how information disseminates across language borders.

Hale [47] tackled the aspect of cross-language linking among blogs in English, Spanish, and Japanese. Focusing on the topic of the earthquake in Haiti in 2010, he was able to quantify the increase in foreign content awareness over time and

detect patterns of cross-lingual linking among blogs [47]. Most notably, blogs in English linked much less to foreign content than the blogs in Spanish and Japanese, the largest single destination of cross-lingual links being a collection of photos [47]. Interestingly, *Global Voices*, an international blogging community that promotes translation of content, created 15% of all cross-lingual linking in the dataset [47]. This finding illustrates that designing for multilingualism and cross-cultural awareness has a impact on the network structure.

This network approach to languages has been applied to the blogosphere, but not yet to the microblogging service Twitter. A study on the influence of distance in the formation of Twitter ties [99] tangentially included language. The most interesting finding related to this research is the observation of cross-language ties in a sample of 1768 pairs of nodes: English-Other (7.4%), Other-English (3.1%), Other-Other (1.1%) [99]. However, the authors identified the users' language using only one post, and made some questionable assumptions in their interpretations, like users being monolingual and equating country with language use, which lead them to be skeptical of their own observation of 8% geolocated subjects in Brazil using English [99]. Actually, this percentage of English use on Twitter in Brazil is very close to the estimate provided by a more solid and large-scale survey [83].

Like in [99], there are examples of rough research assumptions about languages, geography, and text analysis on Twitter that manifest the need for a deeper understanding of these interrelated areas of study.

Shifting the focus from languages to countries, it is possible to find research that looks at transnational ties in Twitter. Poblete et al. [90] illustrated with a

detailed network graph the strength of ties between the ten most active countries in Twitter. These ties are the aggregated results of users' "following" relationships in Twitter. Apart from the expected stronger connections among countries that share the same language (U.S., U.K., Canada, and Australia), the graph also unveils that the U.S. attracts the most international attention, while it pays little attention outside its borders [90]. Also, South Korea and Brazil have little connections overseas [90].

Ideally, an holistic view of the language ecology in Twitter will require an analysis of the languages in the overlapping networks of users' attention ties (follower relationship), interaction ties (replies and retweets), topics and linked resources. As a starting point, this dissertation focuses on attention networks at small scale.

3.3 Multilingual Twitter

In section 3.2, I noted that there are currently no research works that take a network approach to languages on Twitter, which would be useful to understand cross-language ties and communication connections between language communities. On the positive side, there are a number of works that study language on Twitter for different purposes and, as the body of research literature on Twitter is growing fast, it might be a matter of time that works on networked language communities become published.

A large-scale study on the languages used in Twitter (more than 62 million posts over a period of four weeks) reveals that almost 49% of the posts are written

in a language different from English and provides a ranking of the top 10 languages [55]. This study records the number of URLs and hashtags (keywords preceded by the @ sign) shared by people from different language communities [55]. Their closing reflection encourages the study of bilingual brokers and how information flows across language communities [55]. An even larger-scale survey of Twitter (380 million posts over 564 days), conducted later, also presents a ranking of top languages among the 78 detected [83].

When comparing the language rankings of the first study and the later survey, we can observe that European languages are loosing positions against Asiatic languages, except for the increase of Spanish and the unchallenged dominance of English [83]. Twitter is a fast-changing environment, where the language ecologies and their impact on communication behavior are constantly in the process of negotiating their contexts and finding their balances.

The same survey, *The Twitter of Babel* [83], compares the percentage of English use on Twitter in 20 countries versus the vernacular language use with an illuminating graph (figure 3.1). This graph should dismiss the assumptions that equate country and language use. Another piece of evidence against such an assumption is the survey's focus on Belgium. Mocanu et al. [83] compare Belgium census data with Twitter data, and observe that the Flemish-speaking population (or Belgian Dutch speakers) is over-represented in Twitter by comparison to the Walloon French speaking population. The researchers connect it with the finding that Twitter has higher penetration in the Netherlands than in France [83], which might attract more users of Dutch language variants regardless of geographical borders [83].

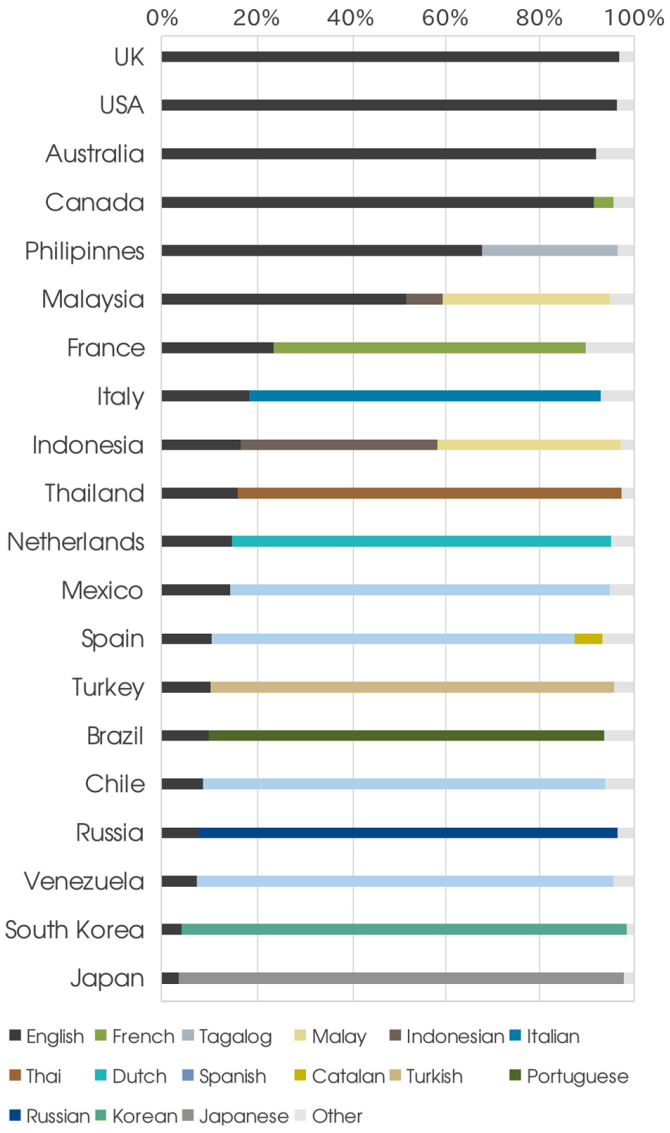


Figure 3.1: Language share of the top 20 most active countries on Twitter, ordered by number of English posts. This graph is an extract from the journal article [83]: Mocanu D, Baronchelli A, Perra N, Gonalves B, Zhang Q, et al. (2013) The Twitter of Babel: Mapping World Languages through Microblogging Platforms. PLoS ONE 8(4): e61981

The Twitter of Babel [83] constitutes a very valuable large-scale survey with examples at different scales, from country level, to city level and neighborhood. However, it does not mention or count cross-lingual connections and bilingual users, even when being implicit in the multilingual situations they describe. Here is where a network approach could provide more insights about transnational influence.

Drawing attention to methodological challenges, Graham et al. [40] compare common geolocalization and language identification methods used for Twitter data analysis. One of the issues the authors encountered is the difficulty in classifying text correctly as Arabic when the Twitter posts were written in Roman alphabet [40], like the cases reported in section 3.1. In particular, Compact Language Detector failed to classify romanized Arabic in 89% of cases [40]. Also, Bergsma et al. [9] detail the challenges in automatic language identification of Twitter posts. The researchers used people to annotate the language of the posts and build a test collection of Twitter texts in languages with non-roman scrip, i.e. Arabic, Farsi, Hindi, and Urdu [9]. The aim is improving automatic language identification in these languages [9].

Looking at the different use of Twitter depending on the language, at least two studies document the different frequency of features in Twitter posts, such as URLs, hashtags, repostings, and user mentions [110, 55]. The findings of these works suggest that Twitter is used more for conversational purposes in some languages, like Indonesian, while in other languages is more common to use it for sharing resources, like German [110, 55]. This dissertation proposes to study in future work particular languages as factors influencing the communication behavior of multilingual users.

In conclusion, the studies about languages in Twitter are descriptive, or of survey type, and only implicitly one can guess there are multilingual users playing a role in the language landscape they describe. Even when comparing the different uses of Twitter depending on the language, these studies do not investigate further into the interactions between language communities and their relation to language choice.

Chapter 4

Methodology

Inspired by an expanded paradigm of Web Content Analysis proposed by Herring [52], who also pioneered a network approach to the study of languages on social media [53], this research includes social network analysis, natural language processing for automatic language identification, theme and exchange analyses. This expanded paradigm proposes broadening the construct of content analysis for accommodating new techniques of analysis appropriate for the evolving landscape of the Internet, and enumerates link and exchange analyses, topic analysis, feature analysis, image analysis, language analysis, etc.

In this dissertation, I apply social network analysis to answer the first research question about the egocentric networks of multilingual users; I use two regression models in the factor analysis to answer the second research question on social network factors that affect language choice; finally, I test the hypothesis that the addressivity feature (@ sign) influences language choice, and explore with a theme analysis how other textual features might be facilitating cross-cultural awareness.

This chapter starts with a brief introduction of the research design, followed by an account of the collection and processing of data that underlies the four studies of the dissertation. The details of the analysis are described in the chapters corresponding to each study.

4.1 Research Design

The research design is composed of four sequential studies of the same datasets, focusing on complementary facets of mediation between language communities and language choice. Table 4.1 shows a schematic view of the four studies and the corresponding chapters.

First, I identified Twitter users authoring posts in English and another language. I collected their last 50 posts and their egocentric network with degree 1.5. The egocentric network with degree 1.5 includes the people connecting with the multilingual subject or *ego* and the connections among the people directly connected with the ego (see section 2.3 for social network concepts). Also, I analyzed automatically the last 30 posts of all the users within the egocentric networks to identify the language they are using in Twitter.

In summary, the data comprises a list of 92 egos, with 50 posts each, and a list of contacts associated with every ego, with a language label, and their linkages in the form of an adjacency list. Figure 4.1 illustrates the components of the datasets.

In chapter 5, the *social network analysis* combines a qualitative approach and network statistics to generate a taxonomy of network types based on the patterns of intersections and connections between language groups. The study follows an exploratory design, with a first qualitative phase that takes a grounded theory approach, and a second quantitative phase that consolidates the qualitative findings. The unit of the analysis is the egocentric network of multilingual users. I visualized the 92 networks with the Gephi social network analysis tool and identified the groups

Problem	Facet	Chapter	Objective	Approach
multilingual Twitter users as mediators	social network	5	classification of egocentric networks	social network analysis QUAL+QUAN
	content	7	exploring cross-cultural themes	theme analysis QUAL
language choices of multilingual Twitter users	social network	6	influence of network in language choice	factor analysis, regression QUAN
	content	7	influence of addressivity in language choice	hypothesis test QUAN

Table 4.1: Research design divided in four studies of the same datasets, looking at the research problem of multilingual Twitter users as mediators and their language choices with different foci. In chapter 5, I use social network analysis for classifying egocentric networks of multilingual Twitter users. Chapter 6 consists of a factor analysis, using regression models to detect if the social network influences language choices of multilingual users. Chapter 7 focuses on textual content; firstly, I test the hypothesis that addressivity influences language choice and, secondly, I look for international themes and other textual features that might indicate cross-cultural awareness.

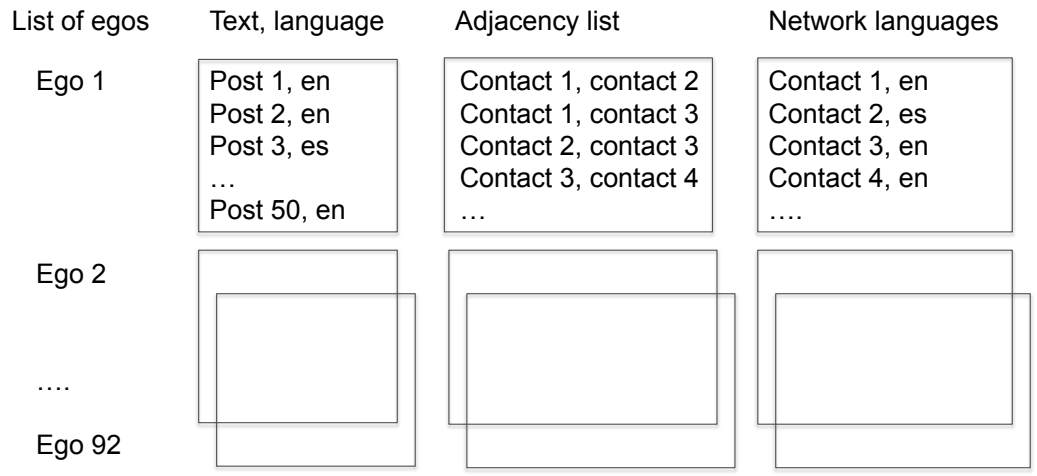


Figure 4.1: The data comprises a list of 92 egos, with 50 posts each and language labels, a list of their contacts with a language label, and the social network links in the form of an adjacency list.

of people that write in different languages. Focusing on the structural relationships of these language groups, I complemented the qualitative study of visualizations with network statistics specifically created to provide a robust definition of network types. Finally, I used machine learning for testing the results.

In chapter 6, the *factor analysis* models the influence of a set of factors related to the social network in the language choices of multilingual users. The dependent variables considered are the proportion of English use and non-English use within the 50 posts of the ego (language choice of the ego). The factors included are the proportion of English and non-English language use in the social network of the ego, and the degree of multilingualism of the social network. The relative importance of factors, or their weight, is represented by the coefficients obtained by fitting two different generalized linear models to the dataset (linear and logistic regression).

In chapter 7, *exploring textual features*, I shift attention from the social network to the content of the posts written by the egos. First, I look at the textual feature of the @ sign at the beginning of a post as an indicator of addressivity. Based on this indicator, I test the hypothesis that the type of exchange (public post versus reply to an individual) influences the choice between English and other languages.

In a second study included in chapter 7, I look at content with the objective of detecting themes that might help in creating cross-cultural awareness, where the multilingual users could be acting as mediators from the point of view of their messages. I identify themes related to non-English speaking countries or communities in English posts and, also, I identify English hashtags (keywords preceded by the # sign) inserted in non-English posts. Using a generic theme analysis, this study serves as an explorative qualitative phase to inform the design of future studies after this dissertation work.

4.2 Sampling and Data Collection

I identified potential multilingual Twitter users with the help of Prof. Jennifer Golbeck and Tony Rogers. We started by issuing queries to the Google search engine, restricted to the Twitter domain, that combined one English word ("between" or "tomorrow") and one of the words in the list of figure 4.2. For instance, a query was "tomorrow" and "también" (which means "also" in Spanish). The words were selected from lists of "stop words" for every language. Stop words are very common words in a language. There are many lists of stop words created for natural language

Language	Words
Arabic	عادي , مدار
Chinese	关
French	alors, très
German	zusammen, gern
Greek	περίπου
Hebrew	טוב
Italian	molto, peggio
Japanese	これ , まで
Korean	에
Polish	właśnie, chyba, albo
Portuguese	muito
Russian	к , о
Spanish	desde, también

Figure 4.2: Common words in different languages used for querying in combination with English words.

processing, usually to filter them for various purposes. In this case, I used these common words to represent each language. The main selection criteria was that the word should not be identical or similar to any other word in a different language, in order to avoid ambiguity about the language the word represents.

In sections 2.5 and 3.3, I reviewed studies that document the dominant use of English in Twitter and how this relates to the weight that the United States has in the social network [60, 90] and to its use in many non-English speaking countries that are active on Twitter [83]. When trying bilingual combinations in our initial search, it was very difficult to find bilinguals that did not use English as one of the active languages; this realization is supported by the findings of a study [99] commented in section 3.3. For this reason, we decided to limit the sampling to multilingual users who wrote in English and, at least, one other language.

The search results directed to the users' profile pages on Twitter. The ordered ranking of users' profiles given by Google could be placing more popular users first, biasing our initial selection, but we ignore the actual criteria used by the search engine. We visited these profile pages, read the last posts, and checked that the subjects were actually using two languages. We established clear written instructions for selecting them. In particular, we did not select users whose:

- posts in one language were automatically generated (i.e. users posting in Spanish with only Foursquare checkins in English) or were spam,
- posts in one language were only named entities, like song titles, names of books, etc.
- posts in one language were only reposted content (or "retweets").

Note that reposting on Twitter does not prove any active knowledge in a language, as it only requires to click on a button or copy text. Moreover, if users just repost the same text, the message stays concealed in the same language community.

Also, the instructions for selecting a user required that he or she had written at least one post entirely in a second language, had more than 30 posts (excluding "retweets"), and had between 4 and 5,000 followers. I discarded potential subjects that had more than 5000 followers due to the computational workload required for processing large social networks and the policy limitations of the Twitter API for extracting data.

We identified 175 potential multilingual users. After this first selection of users, I retrieved the last 50 posts of each one of them by means of the Twitter

API. The API only allows one to extract data from public user accounts in Twitter. For this reason, accounts made private by the users are not included in the sample of multilingual subjects, and when extracting their contacts automatically, private accounts render no data.

As in the previous selection phase, I did not include the repostings, with the exception of those that had a comment added by the user; in such cases, it might be possible to find bilingual text and translation. Specifically, the 50 posts of every user did not include the repostings that were shared clicking on a "retweet" button, or those posts that started with the characters "RT" or "rt" (abbreviations of retweet), but could include the posts that had some text written before RT or rt.

Based on the data, I selected only those users who had written at least 4 non-automatically generated posts in a second language, to ensure that the language was well represented. During the data collection process, I had to discard some users because they made their accounts private or closed them, and one user started posting spam. The data collection process spanned from October 3 to November 7, 2011.

Finally, my sample contains 92 multilingual users that write in 19 languages (Arabic, Basque, Catalan, Chinese, Dutch, English, French, Galician, German, Greek, Hebrew, Italian, Japanese, Korean, Mongolian, Polish, Portuguese, Russian, Turkish), usually two or three languages per person.

Figure 4.3 shows the purpose of different components of the dataset. I kept the last 50 posts of the final multilingual users for studying their language choices and conducting the theme analysis. Also, with the help of Prof. Jennifer Golbeck,

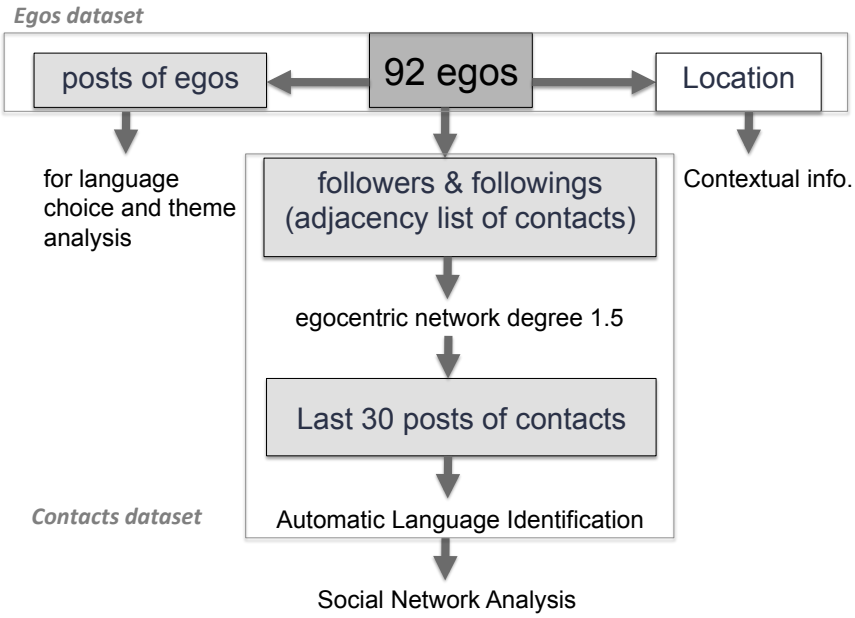


Figure 4.3: Data collection and purpose of different datasets extracted from Twitter.

I extracted the location of the multilingual users from their profiles as contextual information. For every multilingual user (ego) we extracted the egocentric networks with degree 1.5 in the form of adjacency lists of followers and followings (contacts) for the social network analysis, as explained in section 4.1. In total, there are 25,556 contacts within the 92 egocentric networks. Finally, I retrieved the last 100 posts from the contacts' accounts to identify the languages they use in Twitter, with the exception of private accounts and accounts with no posts (5,950 cases). As previously explained, only 30 posts per user were analyzed and the repostings included have text before the characters "RT" and "rt".

4.3 Methods for Assigning Language Labels to Users

In this section, I introduce the options I considered to assign language labels to every user in the egocentric networks with the purpose of completing the contacts dataset illustrated in figure 4.3. First, I had to automatically detect the language(s) used in a number of their posts and, secondly, I had to determine their language profile based on those multiple posts.

The egos dataset also required the automatic identification of the languages of posts, but the egos were not assigned a language label. In the factor analysis, the language profile of the multilingual egos is conceptualized differently, as pairs or frequencies for the two main languages of the user.

I considered two main options in order to assign a language label to a user in Twitter: (1) extracting the language code that the user has selected on the Twitter interface, and (2) automatically identifying the language of a certain number of posts that the user has written. In the case of extracting the language code of the interface, the immediate problem is that this code is not accurate for bilingual users. Also, as I will explain in section 4.4, a test suggested that the interface language has a very high error rate in representing the actual languages of the user. One reason could be that many users do not change the interface language given by default (i.e. English) because they can understand it, but prefer to write in their native language.

As a result of these considerations, I chose the option of automatically identifying the language of users' posts. As reviewed in section 3.2, related research

[99] only used one post per person to assign a language to a user. This is insufficient to determine bilingual and multilingual use, and also problematic in a noisy environment such as Twitter, with frequent cases of automatic posting and spam.

A question that I will address in the next section is how many Twitter posts are enough to determine the language(s) of the user. Having more than one post and language label for a user requires a process to determine which language label fits best for that user.

4.3.1 Tools for automatic language identification

The first step was to identify the language of the users' posts. I considered three tools: Google Language Identification tool (Google's proprietary option), Chorme browser Compact Language Detector (partly open source code by Google), and Python Language Detector (an open source module for programming language Python). Google's language detection algorithm uses quadgrams —or four character tokens— [118, 97] and Python uses trigrams [43].

Briefly, the process of using the language identification tool works as follows: I send an input file that contains the posts of a user after eliminating mentions, hashtags, URLs and symbols, and the language identification tool returns an output file with the **language labels** and **confidence levels** for every post.

4.3.2 Algorithm for assigning a language label to a person

In a second step, I elaborated the rules for assigning a language or languages to a person. For a given user, with a list containing pairs of language and confidence level, the heuristics of the algorithm are:

- Discard all pairs with confidence level below 0.1. The purpose of this rule is to eliminate noise or inaccuracies in the language assignation method. If no pair remains, the language label assigned to the user is "unknown".

- For each remaining language, compute the frequency (number of posts in that language) and select the highest confidence value of all posts in that language, thereafter called "maximum confidence".

- Discard all languages with a frequency below 10% of the total number of posts for that user. The purpose of this rule is to eliminate languages that are not well represented in the profile of the user, due to automatic posting, etc. If no language passes the frequency threshold, the label assigned to the user is "unknown".

- Determine if the user is monolingual or multilingual. If more than one language has maximum confidence equal or greater than 0.7, the language label assigned to the user is a multilingual label. Otherwise is monolingual. Note that the "maximum confidence" is the highest confidence level of all posts in a language for one user. Therefore, the requirements for considering a user multilingual are: (1) at least two languages with 10% minimum frequency (2) and maximum

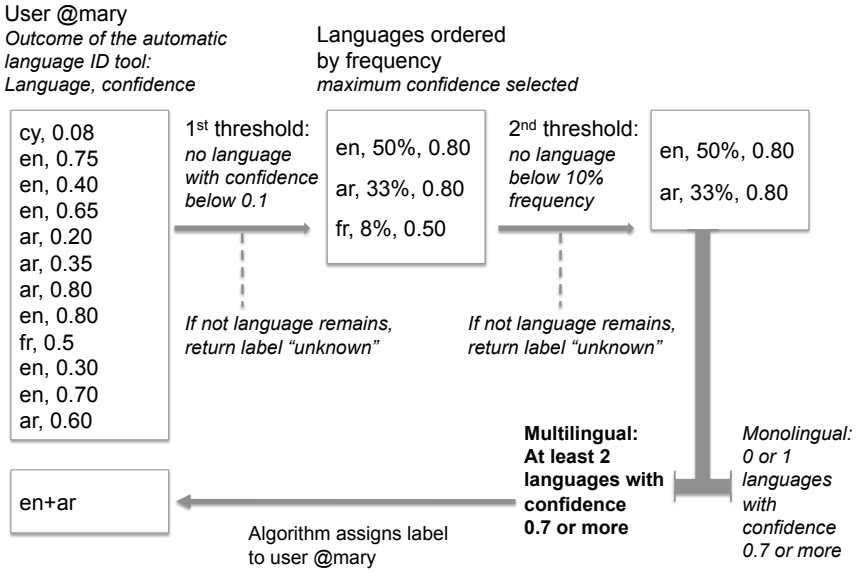
confidence equal to or greater than 0.7. This multilingual label is composed by the code of the most frequent language and the code of the second most frequent language.

- In the monolingual case, if only one language has maximum confidence equal to or greater than 0.7, that is the language assigned to the user.
- In the monolingual case, if no language has maximum confidence equal to or greater than 0.7, the language assigned to the user is the one with the highest frequency.

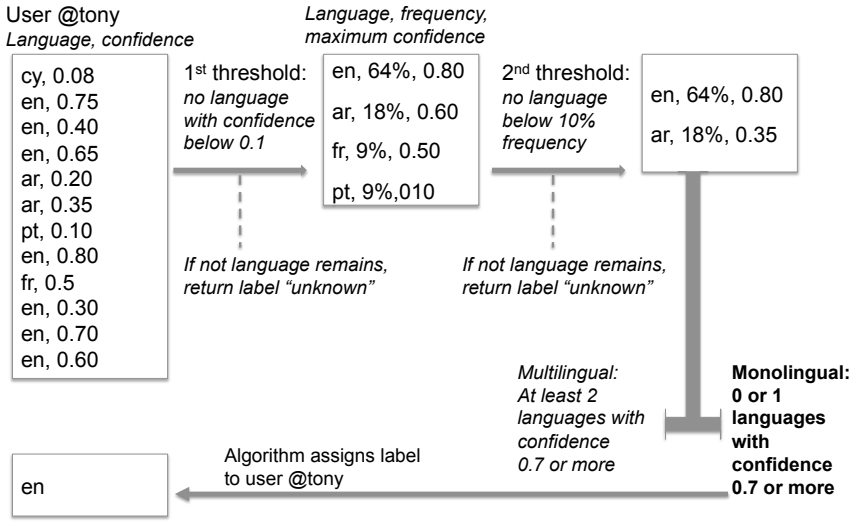
Figures 4.4a and 4.4b illustrate the process of assigning a language label to a user with examples.

Note that the thresholds are based on the assumption that the confidence level represents a probability between 0 and 1, but they could be tailored for each tool. The confidence level of 0.7, used as threshold to determine multilingualism, was selected after observing issues derived from transliteration.

For instance, Arabic was sometimes written in the Arabic scrip, but also in a romanized form that the tool identified incorrectly and with low confidence as some other language. In the cases where Arabic was one of the two languages used, the romanized form (with an inaccurate language label and low confidence) had enough frequency to displace the Arabic as one of the actual languages that composed the multilingual label. Adding the requirement of high confidence (equal to or greater than 0.7) eliminated the instances with incorrect labeling and favored the Arabic that was correctly classified because it was not transliterated.



(a) Example where the algorithm assigns a multilingual label to a user.



(b) Example where the algorithm assigns a monolingual label to a user.

Figure 4.4: Two examples of how users are assigned language labels by the algorithm after the language of their posts have been automatically identified.

4.4 Testing Methods for Assigning Language Labels to Users

In this section, I explain the creation of a test dataset and a baseline for comparing the results of different language identification tools, and for comparing the results of the language assignation algorithm versus a human making that assignation.

Finally, I compare the test results using a number of posts per user between 1 and 100 to answer the question: how many Twitter posts are enough to determine the language(s) of the user?

4.4.1 The test dataset

For testing the tools, I prepared a sample of users. I randomly sampled 10 egos from the 92, and 20 contacts for each ego —or all contacts, whichever is greater— obtaining a total of 190 users from the list 25,556 contacts. From those users, only 177 had data available. The other 13 users had no data either because their account was private or they did not post anything. Finally, I extracted the last 100 posts of the 177 users, including repostings that had an added comment by the user. In total, I extracted 15,973 posts. This is my test dataset.

4.4.2 The baseline

I created a baseline as close as possible to human labeling. Given the time and resource constrains, I decided to use one of the automatic language identification tools to assign a language to each post as an initial guess and manually revise the

results. In this task, I took advantage of my skills in English, Spanish, and French, as well as my familiarity with some Romanic languages, such as Portuguese, Italian, Catalan, and Galician. The number of posts that were sent to the language identifier after eliminating mentions, hashtags, URLs and symbols is 15,856 (from the test dataset). Then, I revised the language labels of the posts following these criteria:

- A. A post is labelled monolingual if any of the following are true:

    1. All words are in one language.

    2. Only one word in a second language within a post of five or more words.

    3. There is a title or named entity in a second language and fewer than five words in the most dominat language of the user.

- B. A post is labelled bilingual if any of the following are true:

    1. There is one word in a second language when a post has fewer than five words.

    2. There are two words in a second language when the post has between five and ten words included, unless is case A3.

    2. If the post has more than ten words and there are at least three words in a second language, unless is case A3.

    3. A title or named entity is in a second language but there are at least five or more words in the other.

- C. A post is labelled automatic if any of the following are true:

1. There are two identical posts.

2. There are at least three posts that are nearly identical except for a number or a word.

3. There are sentences like: "Posted a picture on Facebook", "liked a photo on Facebook", "favorited a Youtube video", "I am at something @ name of place" (foursquare).

- D. A post has a non-identifiable language if any of the following are true:

1. There is a named entity that could correspond to more than one language.

2. There are only symbols, emoticons, or random letters.

3. Other reasons.

Named entities —a term coined in the field of Information Extraction— are information units that consist of rigid designators for a referent, like proper names of people or organization names, locations or times, among many types [84].

In the cases where Arabic, Hebrew and Mongolian were transliterated, the tool did not identified the language correctly. In those cases, I considered the language of the post to be the language of the user other than English (Arabic, Hebrew or Mongolian). This language was determined in other posts written by the same user in non transliterated form, where the tool is more accurate.

Finally, the criteria for classifying the language profile of the user for the baseline was to use a 10% frequency threshold do determine if the user was monolingual

or multilingual. If fewer than 10% of the posts were in a second language, the user was considered monolingual and assigned the label of the most frequent language. If the second language passed the threshold, the user was assigned a multilingual label composed by the codes of the most frequent language and the second most frequent language. Automatic posts and non-identifiable language posts did not add to the frequency count of any language. Bilingual posts added 0.5 frequency points (instead of 1) for each of the two languages.

To create the baseline, I decided to use Google Language Identification tool because the Compact Language Detector (CLD) has two additional disadvantages and Python Language Detector has one additional disadvantage. Unlike Google's proprietary option, CLD cannot detect the Mongolian language in the dataset, and the confidence values do not represent probability. The confidence values range from 0 to over 100, but the maximum value is unknown. The lack of interpretability of confidence values poses a challenge to use the algorithm that assigns a language label to a person.

Python Language Detector is able to detect Mongolian and the confidence values are between 0 and 1, which might indicate probability. In practice, the confidence values are biased towards the range 0.9–1. The minimum is only 0 when the post is empty. Instead, 0.17 acts as the minimum confidence value, for instance, in cases when the post has just a symbol and any guess should return a 0 confidence value. This biased behavior of the confidence values would affect the performance of the algorithm. I considered tailoring the thresholds of the algorithm that assigns a language to a person to account for this, but given the disproportionate amount of

posts with confidence level 0.9, this value has little discriminatory power. For these reasons, I expected that the Google Language Identification tool would perform better.

In summary, I considered my baseline to be the results of Google Language Identification with a subsequent revision, and I assigned the language labels according to a set of criteria described in this subsection 4.4.2.

4.4.3 Testing the language identification tools and the algorithm that assigns language labels to users

I tested the performance of the language assigning algorithm, comparing the baseline (Google Language ID, manual revision, criteria-based language assignation) with the results of the automated language assignation (Google Language ID, manual revision, language assigning algorithm). The manual revision is not performed in the actual analysis of the contacts dataset, but serves for testing the performance of the language assigning algorithm, changing only one variable with respect to the baseline. To obtain the estimated error rate, I divided the number of cases where the automated results did not coincide with the baseline by 177 total cases. The resulting estimated error rate is 6.78%, with 1.13% false negatives (missing multilinguals), and 5.65% false positives. Therefore, the algorithm tends to overrepresent multilinguals.

Subsequently, I tested the performance of Google Language Identification tool in combination with the language assigning algorithm, eliminating the human re-

vision (Google Language ID, language assigning algorithm). These are the actual conditions of the analysis performed with the contacts dataset. I computed the estimated error rate with respect to the baseline. Figure 4.5 shows how increasing the number of posts used for assigning a language to a person diminishes the estimated error rate.

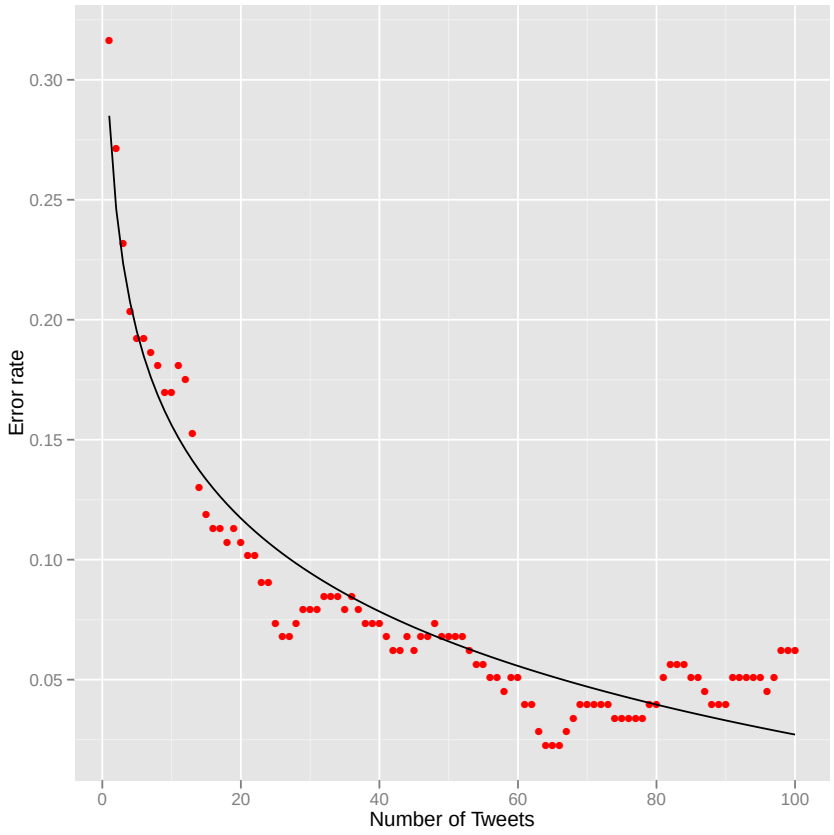


Figure 4.5: The *y axis* represents the estimated error rate of using the Google Language ID method with respect to the baseline and the *x axis* represents the number of posts per person used for language assignation. As the number of posts increases, the estimate error rate diminishes like a negative logarithmic function.

Using a regression model, I obtained function 4.1 fitting the estimated error rate as a function of the number of posts per person used for language assignation. The variable x is the number of posts per user.

$$f(x) = 0.285 - 0.056 \times log(x) \tag{4.1}$$

Compared to Google Language ID, the Compact Language Detector does not detect the language in many more instances: Google did not identify the language in 46 cases, while CLD did not identify the language in 3,071 cases of 15,856 in the test dataset.

In the case of the Python Language Detector (LangPy), there are 3,590 different outcomes compared to the Google Language ID (from a total of 15,856 posts). I compared the performance of LangPy in combination with the algorithm that assigns language labels to users (LangPy, language assigning algorithm) with the baseline (Google Language ID, manual revision, criteria-based language assignation). Figure 4.6 shows the estimated error rate of using LangPy with respect to the baseline, as the number of posts used for assigning a language to a person increases. It also displays, side by side, the previous results of the estimated error rate using Google Language ID (Google Language ID, language assigning algorithm) with respect to the baseline. In the case of Google Language ID, the estimated error rate is always lower, partially due to the fact that the baseline uses this tool as a starting point. Also, the biased confidence values of Python Language Detector constitute a challenge for the algorithm that assigns language labels to users.

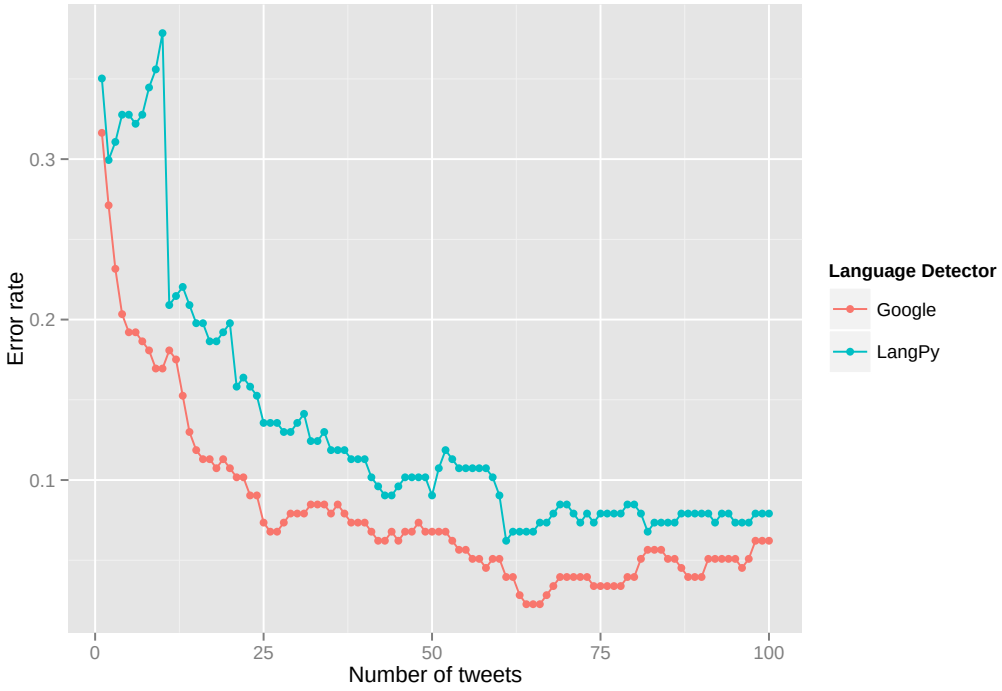


Figure 4.6: The *y axis* represents the estimated error rate with respect to the baseline and the *x axis* represents the number of posts per person used for language assignation. The tools compared are Google Language ID and Python Language Detector (langPy). As the number of posts increases, the estimate error rate diminishes. In the case of Google Language ID, the estimated error rate is always lower, also due to the fact that the baseline uses this tool as a starting point.

As explained before, looking at Twitter's language code to identify a language has the highest estimated error of all methods considered: 0.418. From 177 users in the test dataset, 25.99% were multilingual but the interface does not offer them the option to select more than one language. Another 15.82% of users were using a language different from the language selected on the interface, which was English in all of these cases.

Estimated Error Rate	Number of Posts per User	Estimated Cost ($)
Below 0.15	15	363.11
Below 0.10	30	704.40
Below 0.05	70	1547.85

Table 4.2: Overview of different budget options: estimated error rate in the automatic language analysis associated with the number of posts used per person, and the corresponding analysis costs for the entire contacts dataset. The use of Google Language ID tool costs $20 per one million characters of text.

4.4.4 Deciding the number of posts per user

Once I decided to use Google Language ID tool, a question remained: "How many Twitter posts are enough to determine the language(s) of the user?" In essence, this is a budget question. The cost of using Google Language ID tool is $20 per one million characters of text.

Drawing from the estimated error rate results shown in figure 4.5 and the estimated error rate function 4.1, I selected three error rate options paired with the number of posts per person needed. Based on this number, I used the character count to estimate the cost of analysis for the contacts dataset, which comprises 19606 contacts with up to 100 posts per user. Table 4.2 provides an overview of estimated cost versus estimated error rate in the automatic language analysis of the contacts dataset.

With Prof. Jennifer Golbeck's advice, we selected an estimated error rate 0.10, a cost of $704.4 for the automatic language identification of the contacts dataset using 30 posts per person. Aside from the contacts dataset, there is also a small dataset of 92 egos with 50 posts per user (figure 4.3).

4.5 Assigning Language Labels to Users

For the social network analysis and the factor analysis, the contacts dataset required language labels for 25,556 people in the 92 egocentric networks. However, only 19,606 contacts had available data. The rest was assigned the language label "unknown". As explained in section 4.4.4, I decided to extract 30 posts per person to determine the language or languages they are using in Twitter. The text extracted from the posts was processed through a pipeline for automatic language identification. The first stage in the pipeline involved the elimination of URLs, hashtags (keywords preceded by the # sign), replies and mentions (usernames preceded by the @ sign), and other symbols. In the second stage, I used the Google API to identify the language of each processed post and the confidence value.

Subsequently, every user was represented by a file that contained the languages and confidence values of their posts. This file was processed by the algorithm that assigns language labels to users; the details of the heuristics can be found in section 4.3.2. The labels could be: "unknown"; a code for one language in the case of monolingual users (i.e. "en" for English); or two language codes joined by the symbol "+" in the case of multilingual users (i.e. "en+ar" for English and Arabic).

In the case of the egos dataset, composed of 50 posts from 92 multilingual users, the text was similarly processed for automatic language identification. The purpose of this dataset being different, I automatically processed the data to obtain a percentage of use of the two most frequent languages for every person, with the corresponding language codes, and identified those egos that used a third language in at least 10% of the posts. The results served to describe the characteristics of this dataset, and to quantify the language choices. The egos dataset is composed of 87 bilingual users and 5 trilingual users, all of them use English as fist or second most frequent language (which was a condition in the sampling method). Also, they use one or two of the following 18 languages: Arabic, Basque, Catalan, Chinese, Dutch, French, Galician, German, Greek, Hebrew, Italian, Japanese, Korean, Mongolian, Polish, Portuguese, Russian, Turkish.

4.6 Scope

When looking at the links between users in Twitter, there are two types of networks where the methodology could focus: (1) the network generated by the exchange of messages between people, like replies and repostings, which represents a communication network and a transient social network, dynamically evolving around a topic of interest [64]; (2) the network of "follower of" relationships between people, representing a relatively more stable social network, spanning diverse topics and communities. In both cases, the networks could reflect a static moment in time or an evolution, and the data collection has to be planned accordingly.

In section 1.3, *An Ultimate Goal*, I wonder what generates connections across languages communities and enables cross-cultural communication? Answering that question completely requires different approaches and collecting data to look at both types of networks. There are many complementary aspects that can be analyzed using the transient communication networks and the attention (or followers) social networks. However, in the interest of keeping this research project to a reasonable scope, I decided to focus on attention social networks of multilingual users, as captured at one point in time. In future research, I would like to broaden the scope to account for dynamic networks, topic-based and communication networks.

As documented in section 3.1, language choice online is influenced by many simultaneous factors, such as the cultural and linguistic context in a particular region, the social context, the users' perception of the availability of online resources in a language, and the topic, to name a few factors that this dissertation will not consider. The distinctive approach of this research consists on shifting the focus to factors derived from the social network where the user is immersed. Also, *participants* and *audience* are implicit factors when studying the textual feature of the @ sign. Regarding the *setting* factor, the social networking site Twitter could be considered one variable for comparison, but this work will not expand into other social networks with different characteristics, like Facebook.

Methodologically, Androutsopoulos recommends to take into account the digital surroundings when analyzing written text, for instance, looking at the pictures and videos that are linked [6]. Although this strategy is particularly relevant to

Twitter and would enrich the qualitative analysis, this work will focus just on textual themes and hashtags due to time constraints in the final stage.

According to Rotman et al. [93], this type of research work would be considered an exploratory step prior to embarking in "extreme ethnography", which is a new approach to ethnographic methods for the study of human behavior in large scale online environments. Indeed, adding detailed geographic information and cultural backgrounds of the nodes in the social networks would provide a fascinating overview of international communication patterns among individuals. However, such endeavor will require a wealth of resources, and a long time.

4.7 Limitations

Due to the policy limitations of the Twitter API for extracting data and the computational workload required for processing large social networks, I discarded potential subjects that had more than 5000 followers. In practice, this decision biases the sample against the bigger hubs in the social network. Other limitations in data collection include the impossibility to obtain information from private accounts, for technical and ethical reasons, and the presence of inactive users. These issues translate into a 23% of subjects in the contacts dataset with no data available.

This research work focuses on 92 subjects, and their egocentric networks, which is a small dataset that poses challenges in obtaining statistically significant results. The diversity of languages included, and small size of their respective samples, hindered any attempt to make comparisons between language groups.

Also, a challenge lies in automatically identifying the language of this type of short texts, and subsequently of nodes in the egocentric networks. In computer-mediated communication the text often has characteristics of both written and spoken language, with colloquial and regional dialect features, playful performance with orthography and typography [23], which adds difficulty for automatic language identification. As reviewed in section 3.3, other research works have encountered this problem and have reported high error rates in identifying romanized Arabic [9, 40].

Finally, during the data collection process, we did not collect geolocation information of posts. This type of information consists of GPS coordinates derived from the users' devices, or approximate area derived from the Internet Protocol (IP) address of the users' computer [40]. Only a small number of users publish geocoded posts, as a result, this condition would have reduced the number of subjects and biased the sample [83]. Instead, we collected the location information users provide in a specific field of the Twitter interface. However, this data is unreliable [40] and I decided not to take it into account during the analysis.

4.8 Reliability and Validity

Regarding the reliability of the data collected, this work focuses on messages and actions —like "following" someone— of multilingual users in Twitter, and is not using the data as a proxy for their behavior in other settings. The Twitter API provides access to this information. Spammers are the most compromising problem for the reliability of data. In the case of the 92 egos, I designed the selection steps

(section 4.2) to avoid spammers and, in relation to validity, I also discarded people that were not multilingual users as defined in this investigation.

For improving the reliability of language assignment, I used 30 posts per person, tested different language identification tools and the algorithm that assigns language labels to users based on their posts. I provide an estimated error rate below 10%.

In the social network analysis, I designed a two-step study to improve the reliability of the categories obtained in the qualitative phase with quantitative measures, and tested the results with a classification model.

In the factor analysis, I operationalized the multilingualism of a social network using the concept of entropy, which can be interpreted as the unpredictability of the language used in the network. The more people in the network using different languages, the higher the entropy. Unlike just counting the number of languages present, the entropy accounts for the weigh those diverse languages have on the network and provides a more accurate measure of multilingualism.

In the factor analysis, I used two regression models to compare the results and test their validity. Finally, the qualitative data conformed by the posts written by the 92 egos was categorized into public posts and replies. The comparison of categorical data requires the use of non-parametric statistics to obtain valid results. The theme analysis includes many examples from the data and compares the results with previous findings in related studies to support their validity.

Regarding external validity and generalizability of results, the small scale of the study limits the possibility for extrapolation to a wider multilingual population

in Twitter. On the positive side, the systematic documentation of steps in the grounded theory approach for classifying network types, complemented with network measures, enables replication. This replicability facilitates to scale up the study and potentially obtain more generalizable results.

4.9 Ethical Considerations

This type of research, which consists of collecting public content from the Internet with no aim of presenting subjects in a bad light, is considered a low risk research activity that does not require an approval process by the Institutional Review Board (IRB). In particular, I am not collecting any private information, like age, gender, or real names, neither I am collecting data from accounts made private. However, the study of these new social media environments with user-generated content is challenging the established ethical protocols. Despite this study following the current norms of the research community, many ethical issues are still being debated and protocols might change in the future.

In the article *Six Provocations for Big Data* [22], the authors warn that users posting publicly accessible messages online does not automatically make them consent for anyone to collect and use their data. They have an intended audience and purpose, and are unaware of their data being collected. Unfortunately, there is no practical way to obtain consent from users or to inform them of the data collection process.

The main concern is the user's privacy. The only identifiable personal information in the present datasets are public usernames, but I took the additional precaution of using anonymized identification codes for all the subjects. When presenting textual content, susceptible of including usernames, either I eliminated the user mention or I replaced it for a fake name. However, the vast majority of textual content was used only for automatic language identification.

I have encountered the problem of discovering one minor in the dataset by reading a post stating the age. At that point, I eliminated immediately the subject and corresponding data from the sample. This raises the question about how to detect minors and be able to discard their data. Twitter added in 2012 an age screening program, but it only works in the context of a minor trying to follow a brand intended for adults and registered in the program [101].

Chapter 5

Social Network Analysis

The main goal in this study is to develop a classification of egocentric networks based on the number of language groups that conform them and the patterns of connections between the groups. The types of egocentric networks constitute theoretical constructs to understand the ways in which multilingual users of Twitter are connecting language groups.

For that purpose, I visualized the 92 egocentric networks with the Gephi social network analysis tool. The visualizations represent people as colored nodes and the "follower of" relationship, as edges. The colors represent the single language they use in Twitter, if they are bilingual, or have no data available. The ego is taken out of the picture to avoid obscuring the display with too many edges; all members of the egocentric network are connected with the ego by definition. I chose the layout "Force Atlas", which is a force directed placement scheme developed by Mathieu Jacomy in 2007 for Gephi [35]. The Force Atlas layout follows a similar placement scheme as the commonly used Fuchterman-Reingold layout, where the algorithm replicates a hypothetical physical system trying to minimize the energy, balancing attraction between nodes connected by springs [38]. The force-directed placement schemes are particularly useful in revealing network structure [10], such as communities.

In summary, the visualizations convey structural information and language information about the social network, by separating the social groups or communities in the layout and distinguishing the language groups with colors.

As explained in section 4.1, this study follows an exploratory design with two phases. First, I use a grounded theory approach to identify emergent types of egocentric networks focusing on the structural relationships of language groups. I use grounded theory in the generic sense, to define theoretical constructs derived from qualitative analysis of data, following the principles of the book by Corbin and Strauss [17]. This approach consists of a sequence of coding stages, firstly establishing some basic properties observed in the social networks, secondly extracting codes from the visualizations as defined by their properties, and finally grouping codes into categories according to shared properties.

In the second phase, I complemented the qualitative study of visualizations with network statistics to provide a robust definition of network types. Also, I propose an application of these types using machine learning for classification, which also serves for testing the results. Finally, I discuss the findings in relation to the theoretical framework and related work.

5.1 Qualitative Approach

As an initial step based on visual differences, I separated the egocentric networks in three types: 25 monolingual or very small networks, 62 bilingual networks,

and 5 trilingual networks. Based on this initial classification, I established quantitative thresholds that define these types:

- Bilingual networks: have at least 7 nodes using a second language (L2), and the L2 group represents at least 2% of the graph nodes;
- Monolingual or very small networks: have fewer than 7 nodes using L2, or the L2 group represents less than 2% of the graph nodes;
- Trilingual networks: have at least 7 nodes using a third language (L3), and the L3 group does represents at least 7% of the graph nodes.

A higher threshold for trilingual networks enables to overcome the problem of noise in multilingual networks, where differentiating a third language among several others sometimes becomes challenging. This issue is less accentuated in bilingual networks, which will be the focus of the subsequent analysis. Before proceeding with the analysis of bilingual networks, I provide some insights about trilingual networks with three examples from the dataset.

The first example is the egocentric network of the user "Kepa"[1] (fig. 5.1). The Basque group on the upper side (dark blue) connects with the Spanish group (green in the middle) and in turn, the Spanish group connects with the English group at the bottom (pink). Basque is a minority language in Spain and a co-official language in the Basque Country region, where Spanish is also official language. This network illustrates the interesting intersections and overlaps of language groups in the context of a bilingual region, where English is taught as language for international

[1]Reported user names are changed for privacy protection.

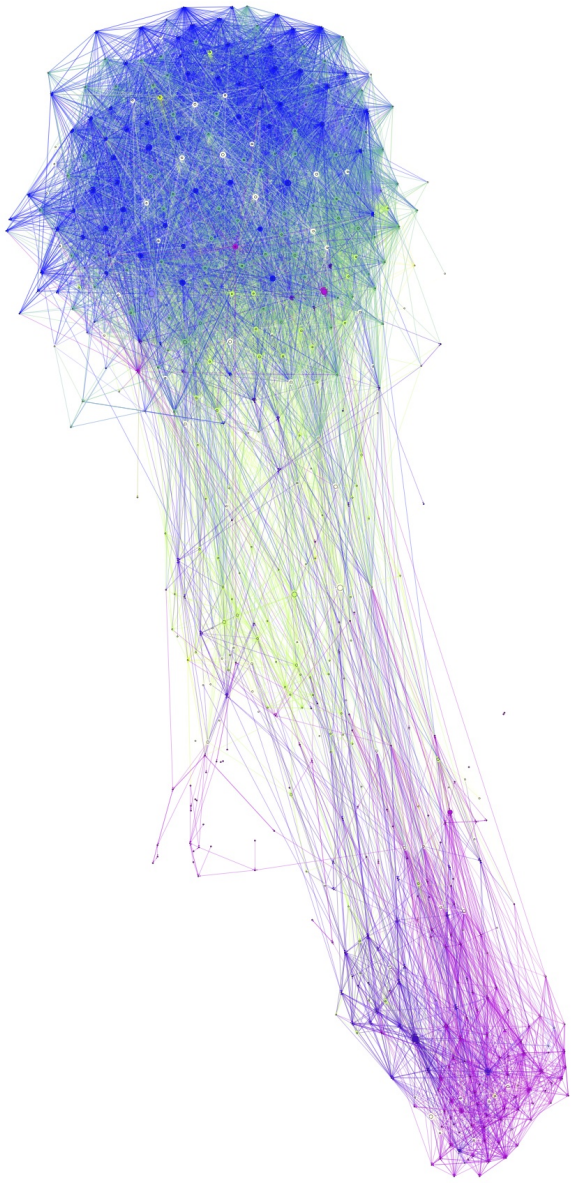


Figure 5.1: Basque group on the upper side (dark blue) connects with the Spanish group (green in the middle) and in turn, the Spanish group connects with the English group at the bottom (pink) and English-Spanish bilinguals (violet). Visualization made with Gephi.

communication. The Spanish-posting group seems to create a path of communication between English and the Basque community.

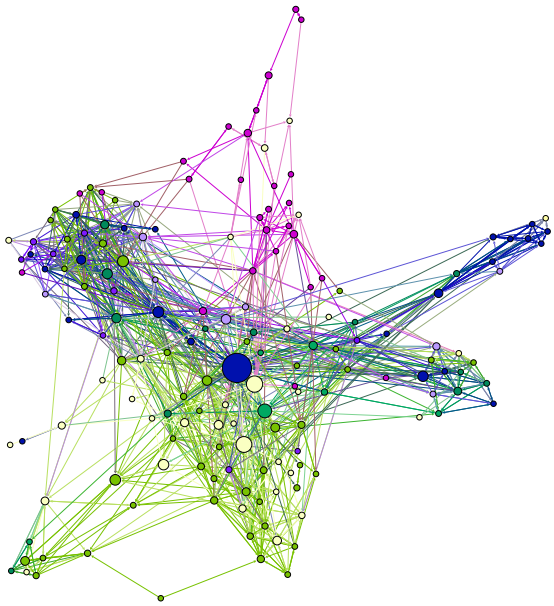


Figure 5.2: Catalan group in the center (dark blue) integrated in the Spanish group at the bottom (light green), connects with the English group at the top (pink). Bilinguals of Catalan-Spanish are represented in dark green, Catalan-English in light violet, and Spanish-English in dark violet. The nodes in light yellow represent nodes with no data. Visualization made with Gephi.

The second example is the egocentric network of the user "Montse" (fig. 5.2). The Catalan group in the central axis of the graph (dark blue) is completely integrated within the Spanish group on the lower side (light green), and there is a smaller English group on the upper side (pink). It is noteworthy the number of bilinguals of Catalan and Spanish (darker green), followed by Catalan and English (light violet) and Spanish and English (darker violet). Catalan is a minority language in Spain and a co-official language in the Catalonia region, where Spanish is also official language. This network illustrates a different flavor of language groups' overlaps in the context of a bilingual region, where English is taught as language for international communication.

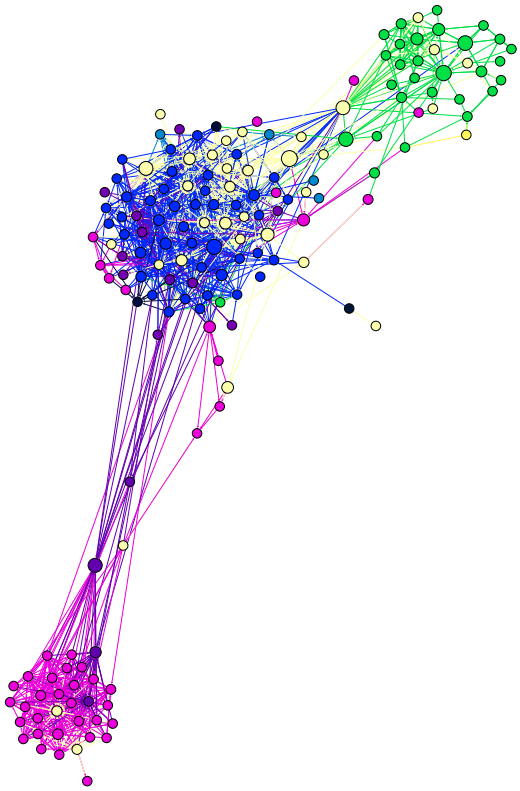


Figure 5.3: The Chinese group in the center (dark blue) connects through a few nodes with the Japanese group on the upper side (green), and with the English group on the lower side (pink), through some bilinguals (violet). The nodes in yellow represent nodes with no data. Visualization made with Gephi.

Finally, the third example is the egocentric network of the user "Wei" (fig. 5.3). The Chinese group in the center (dark blue) connects through a few nodes with the Japanese group on the upper side (green), and with the English group on the lower side (pink). Some of the users connecting the groups either post in English or both in English and Chinese. In this example, English seems to be playing the role of international *Lingua Franca*, connecting the Chinese-posting group with other language groups.

I focused the social network analysis on the bilingual networks, for simplifying the categorization to types of intersections between two language groups. During the initial coding, I created a list of properties observed in the visualizations concerning the structural relationships between the languages groups. These properties are shown in table 5.1:

Properties		
A) Degree of connection between language groups	A1	few connections
	A2	tightly connected
B) Degree of integration of one language group inside another	B1	separated
	B2	partial integration
	B3	complete integration
C) Relative size of one language group respect to the other	C1	similar size
	C2	very different size

Table 5.1: Properties of bilingual networks observed in the Gephi visualizations.

When combining the three types of properties, I deductively obtained 12 codes, for instance: **code 1** consisted of two language groups of similar size (C1), separated (B1), and connected by a few nodes (A1); **code 2** consisted of two language groups of very different size (C2), separated (B1), and connected by a few nodes (A1); **code 9** consisted of two language groups of similar size (C1), tightly connected (A2), and one language group has been partially penetrated by users of the other (B2) ; **code 12** consisted of two language groups of very different size (C2), the small one completely integrated within the big one (A2, B3), etc.

During the final iteration of the coding process, I observed some codes had no instances in the dataset or very few. Those codes that had very few instances could be grouped with codes of similar properties. For instance, regarding codes with a high degree of integration of one language group inside another (B3), there are few instances of language groups with similar size (C1), therefore I merged codes with properties A2 and B3, regardless of the differences in group size (either C1 or C2).

In relation to codes with no instances, some properties presume others, like B3 or B2 (some degree of integration of one language group inside another) require a high degree of connection between the groups (A2); in consequence, certain combinations of properties are not possible. For this reason, some codes were discarded.

- Code 1 (A1, B1, C1) with 12 networks;
- Code 2 (A1, B1, C2) and code 8 (A2, B1, C2) grouped together have 12 networks;
- Code 3 (A1, B2, C1), code 4 (A1, B2, C2), code 5 (A1, B3, C1), and code 6 (A1, B3, C2) have contradictory properties, because B2 and B3 require A2, and there are no instances in the dataset;
- Code 7 (A2, B1, C1) with 12 networks;
- Code 9 (A2, B2, C1) and code 10 (A2, B2, C2) grouped together have 9 networks;
- Code 11 (A2, B3, C1) and code 12 (A2, B3, C2) grouped together have 17 networks;

The resulting groups of codes constitute the five categories of bilingual networks obtained with a qualitative approach. Below, I define the categories of egocentric networks in relation to the patterns of intersection between language groups. Figure 5.4 illustrates them with examples from the data. The names of the categories are metaphorical; here *bridge* is not used as the graph theory term. See appendix A for all the visualizations and the categories they were assigned.

- *Gatekeeper* (Fig. 5.4.1): two language groups connected by a few nodes only, with properties A1, B1, and C1 (12 networks);
- *Language bridge* (Fig. 5.4.2): two tightly connected language groups, but still separated, with properties A2, B1, and C1 (12 networks);
- *Peripheral language* (Fig. 5.4.3): a dominant language group connected to a small or not cohesive language group, with properties A1 or A2, B1, and C2 (12);
- *Union* (Fig. 5.4.4): two tightly connected language groups, where one language group has been penetrated by the other, with properties A2, B2, and C1 or C2 (9 networks);
- *Integration* (Fig. 5.4.5): one language group inside another with properties A2, B3, and C1 or C2 (17 networks).

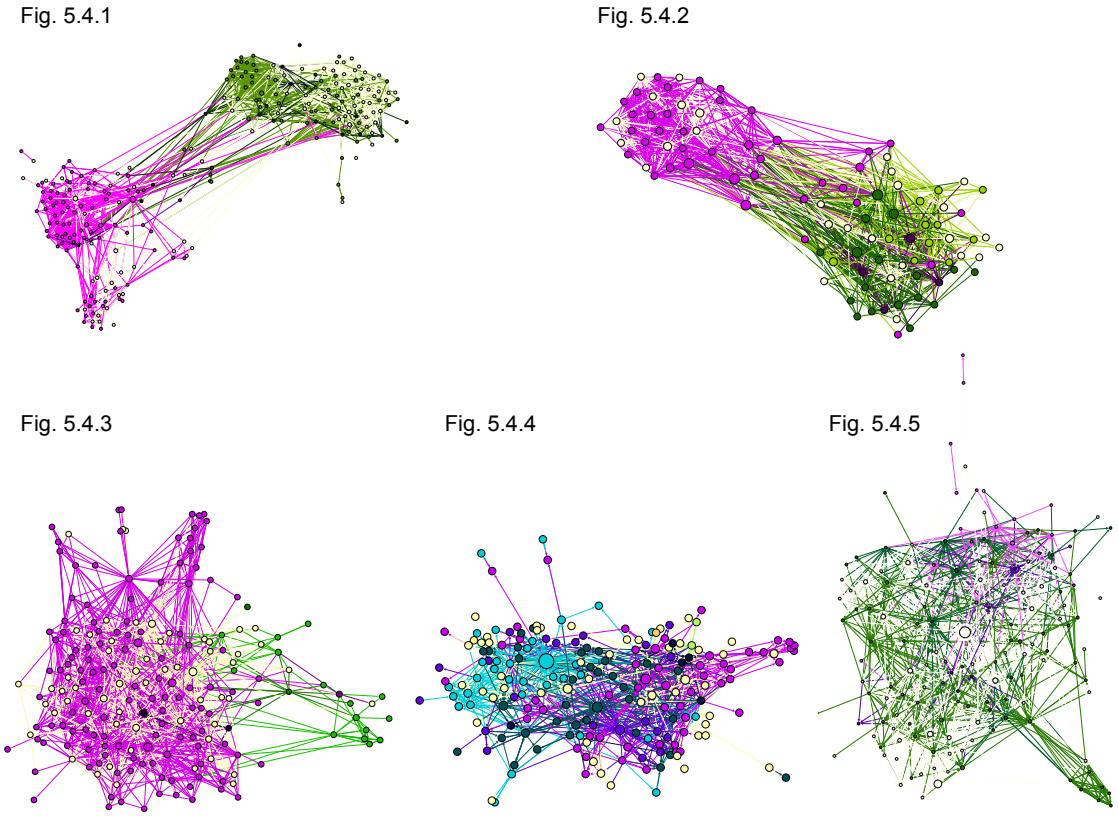


Figure 5.4: Networks of 5 multilingual Twitter users exemplifying the types of egocentric networks. The nodes are their contacts and the edges represent the "follower of" relationship. Pink nodes post in English and yellow/white is used for nodes with no data. Fig. 5.4.1 is the gatekeeper type; there is a French group on the right side (green) loosely connected with an English group on the left. Fig. 5.4.2 represents the language bridge type; in this network, the Japanese group on the right side (green) is tightly connected with the English group on the left, and intermingled with bilingual users (violet and dark green). Fig. 5.4.3 shows the peripheral language, Portuguese, on the right side (green) of the dominant English group. Fig. 5.4.4 exemplifies the union type, where the Greek group on the left (turquoise) is merging and mixing with the English group on the right, and there are many bilinguals (violet and dark green). Fig. 5.4.5 illustrates the integration type; the English group being inside the Arabic (green). Visualizations made with Gephi.

The categories *gatekeeper* and *language bridge* present a continuum of increasing connectivity between the two language groups, where extreme cases could potentially belong to the other category. Similarly, the *union* and *integration* categories present a continuum of increasing penetration of one language group within the other. The implication is that no statistic is going to divide these categories cleanly. However, the network statistics helped to refine which networks were in which categories in the extreme cases.

These different structures can potentially impact information diffusion [80] across languages and nations. In the case of the gatekeeper type, and peripheral language, cross-cultural awareness and information diffusion between the language groups depend on a small number of users. If we look at the proportion of links between the language groups, it seems that information will have higher chances of crossing the language barrier in the case of the union and integration types.

5.2 Network Statistics

Similarly to how user types were defined by network structure in [112], I explored different network statistics to provide a robust definition of the types of bilingual networks. The objective is to define a set of measures that, taken together, can differentiate each network type. Note that this analysis continues to focus on the set of 62 bilingual networks.

First, I tried to convey the qualitative property of degree of connection between language groups with the *cross-language edge ratio* ($XLangR$), as suggested by Prof.

Jennifer Golbeck. To compute this ratio, I used the total number of edges in the graph (T), except those linking to nodes with no data or a non-identifiable language, and the number of edges linking two nodes of different language (t):

$$\frac{t}{T} = XLangR \tag{5.1}$$

Additionally, the ratio between inner edges in the L2 group and the edges going out of the group could convey both the degree of connection of the L2 group with the rest of the graph and the relative size of the group with respect to the graph. Computing the *L2 inner/crossing edge ratio* ($XL2R$) requires: the number of edges connecting two nodes of L2 (τ_{L2}), and the number of edges connecting a L2 node with a node in a different language (t_{L2}).

$$\frac{\tau_{L2}}{t_{L2}} = XL2R \tag{5.2}$$

Another property that is related to the degree of connection and overlap between the language groups is the the bilingual ratio. After determining the two main languages, L1 and L2, computing the *bilingual ratio* (BR) requires the number of nodes in each group (n, m), and the number of bilinguals using both L1 and L2 (b):

$$\frac{b}{n+m} = BR \tag{5.3}$$

Finally, to account for the qualitative property of different size of the two main language groups, I use the proportion of nodes in the L2 group ($p(L2)$) with respect to the sum of nodes in L2 (n) and L1 (m):

$$\frac{n}{n+m} = p(L2) \tag{5.4}$$

As explained in section 5.1, the network categories present a continuum of increasing connectivity and overlap between two language groups, where extreme cases could potentially belong to another category. Even though no statistic is going to divide the categories cleanly, the figures below show how the five categories can be regrouped into three main types that are differentiated by the statistics.

The categories *gatekeeper* and *language bridge* present a continuum of increasing connectivity between the two language groups, but are different from the other types because the *L2 inner/crossing edge ratio* is higher, which implies more connections within the same language group than across language groups (figure 5.5). Also, the L2 proportion differentiates the *gatekeeper-bridge* from the *peripheral* type because the two language groups tend to be of similar size, whereas the different sizes of the language groups is a defining property of the peripheral type (figure 5.6).

Similarly, the *union* and *integration* categories present a continuum of increasing penetration of one language group within the other. The box plots in figure 5.7, representing the *cross-language edge ratio*, show that the integration and union types have higher ratios and are clearly differentiated from the other types. This pattern is consistent with the *bilingual ratio* (5.8), which reinforces the differentiation between integrated (union and integration) and separated (gatekeeper, language bridge, and peripheral language) types.

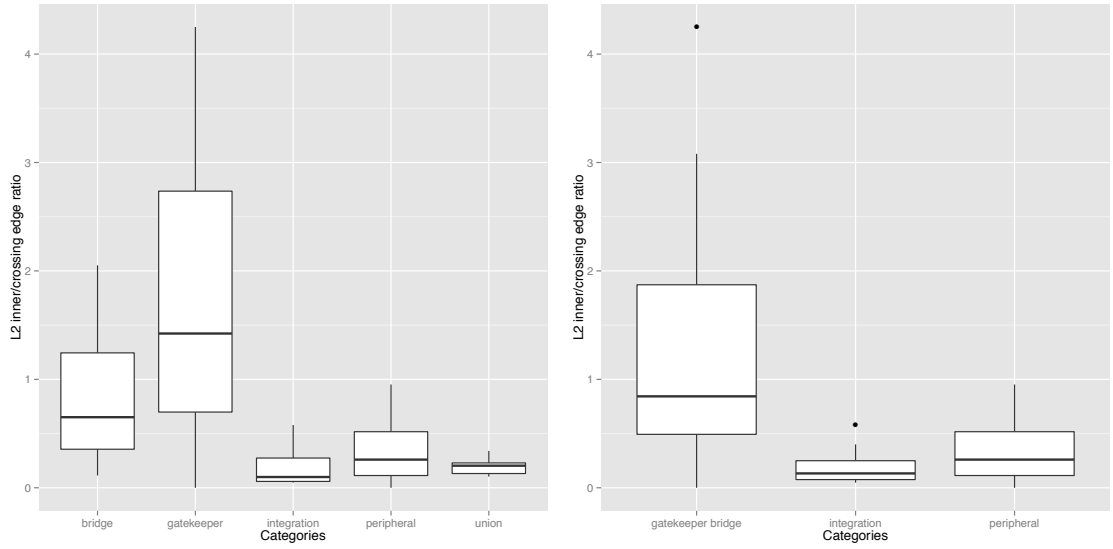


Figure 5.5: L2 inner/crossing edge ratio for five categories (left) and for three categories (right). This statistic differentiates the gatekeeper-bridge type.

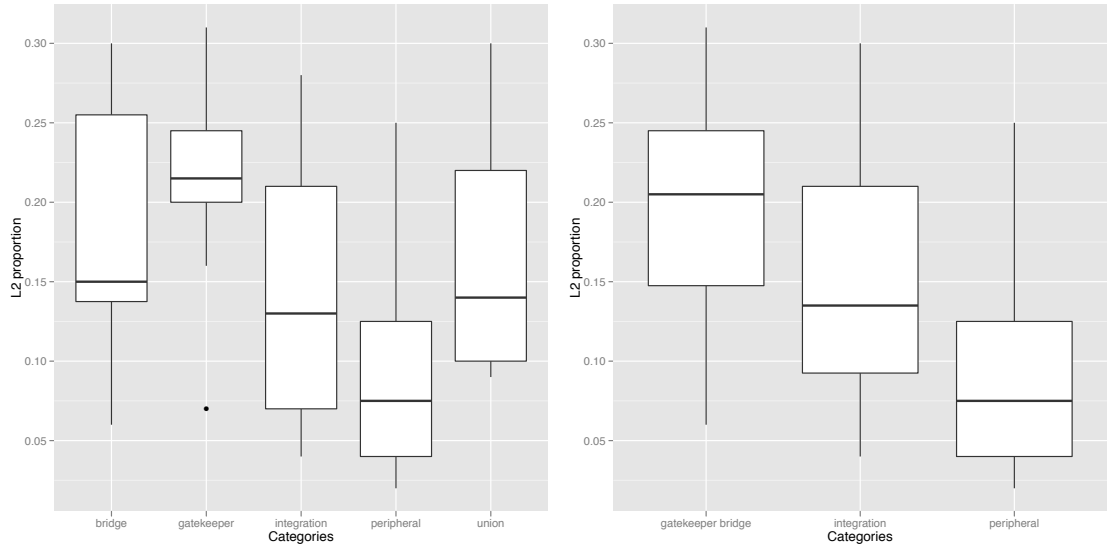


Figure 5.6: L2 group proportion for five categories (left) and for three categories (right). This statistic differentiates the peripheral language type.

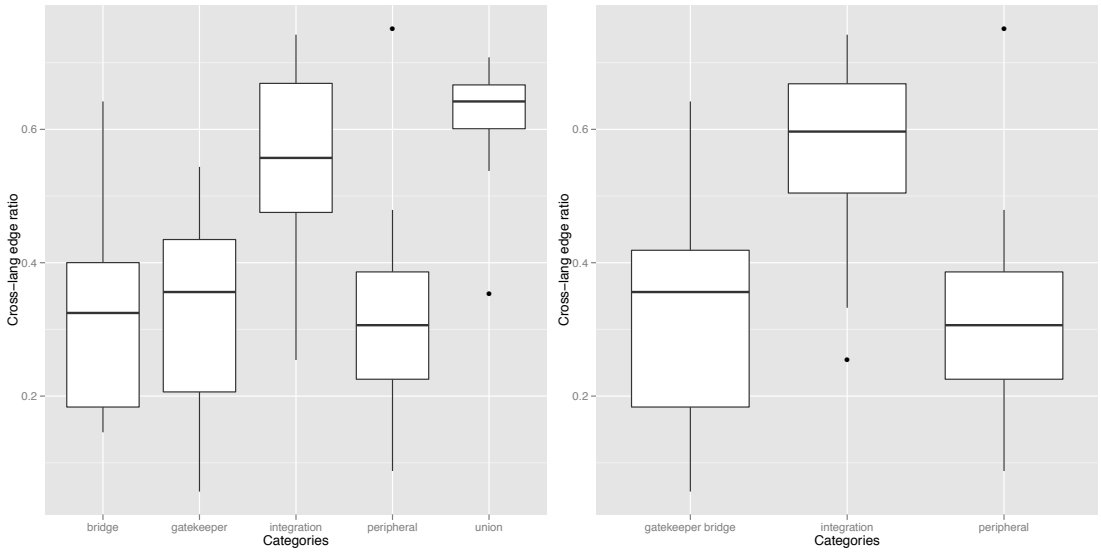


Figure 5.7: Cross-language edge ratio for five categories (left) and for three categories (right). This statistic differentiates the integration and union type.

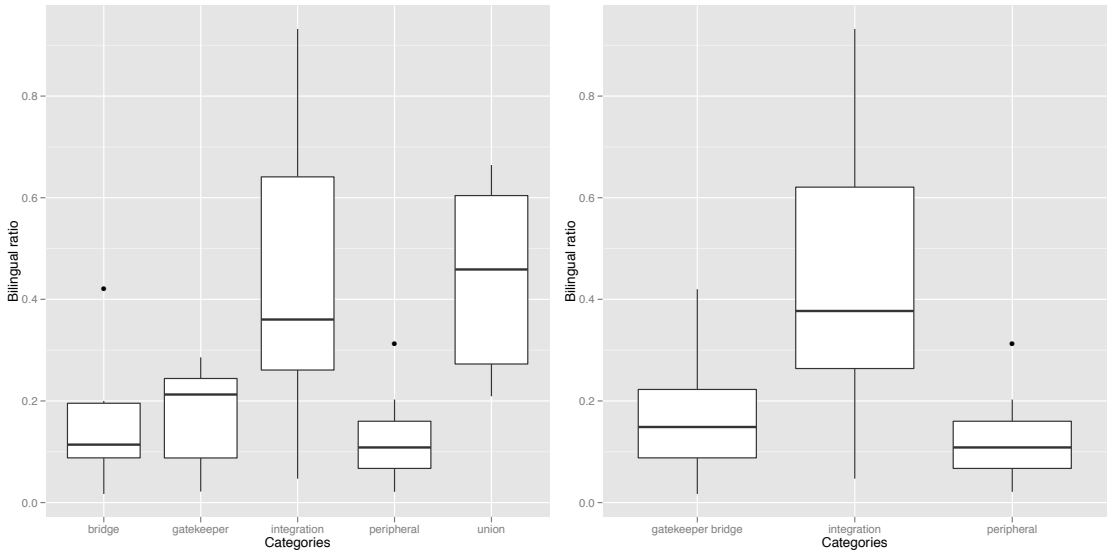


Figure 5.8: Bilingual ratio for five categories (left) and for three categories (right). This statistic differentiates the integration and union type.

In summary, from a quantitative approach three main types of intersection between two language groups in the social network can be defined:

- *Gatekeeper-Language bridge*: defined by a high L2 inner/crossing edge ratio, more connections within the same language group than across language groups, and language groups of similar size (24 networks);
- *Integration and union*: defined by high cross-language edge and bilingual ratios (26 networks);
- *Peripheral language*: the L2 group accounts for a small proportion of the graph, and it does not meet the defining properties of integration and union types (12 networks).

Following the reasoning in section 5.1, the *cross-language edge ratio* and the *bilingual ratio* can reflect the potential for information dissemination across language borders. If we are able to classify the types of intersection between language groups in a set of egocentric networks, we might be able to predict which networks have more potential for cross-lingual linking, translation, and cross-cultural awareness. However, the relationship between the types and the spread of information across languages requires further investigation and fall outside the scope of this work.

5.3 Application of Categories

One potential application of the categories, particularly the three types that are differentiated more clearly with the statistics, is the classification of bilingual

egocentric networks. If the linkage between the types and the potential for cross-language information dissemination is supported by further research, this classification could be fundamental in detecting nodes and their egocentric networks that can spread information across language and national borders more effectively.

In this section, I test the results of the social network analysis with a classification model using machine learning. I trained the classification model using support vector machines (sequential minimal-based implementation, SMO) and the dataset of 62 bilingual networks divided into three types. This dataset included the attributes type, L1, L2, cross-language edge ratio, L2 inner/crossing edge ratio, bilingual ratio, and proportion of L2. I used the Weka (Waikato Environment for Knowledge Analysis) free software for machine learning, developed at the University of Waikato, New Zealand.

Figure 5.9 shows the confusion matrix: all 26 networks of the integration type were classified correctly; 19 of 24 gatekeeper-bridge networks were classified correctly, while only half of the networks of peripheral language type were classified correctly. In general, 51 networks of 62 were classified correctly and 11 incorrectly. The F-measure for accuracy is 0.812 in average, but is particularly high for the integration type, 0.881. These results show a promising potential for prediction, even with this small dataset. As observed, these statistics are enabling differentiation between the types of bilingual networks.

```
=== Summary ===

Correctly Classified Instances          51               82.2581 %
Incorrectly Classified Instances        11               17.7419 %
Kappa statistic                          0.7127
Mean absolute error                      0.2688
Root mean squared error                  0.3474
Relative absolute error                 63.0008 %
Root relative squared error             75.1714 %
Coverage of cases (0.95 level)          96.7742 %
Mean rel. region size (0.95 level)      66.6667 %
Total Number of Instances               62

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure         Class
                 0.792    0.079    0.864      0.792   0.826             gatekeeper bridge
                 1        0.194    0.788      1       0.881             integration
                 0.5      0.02     0.857      0.5     0.632             peripheral
Weighted Avg.    0.823    0.116    0.831      0.823   0.812

=== Confusion Matrix ===

  a  b  c   <-- classified as
 19  4  1 |  a = gatekeeper bridge
  0 26  0 |  b = integration
  3  3  6 |  c = peripheral
```

Figure 5.9: Classification results using 10-fold cross-validation for the SVM model. This model was trained using the dataset of 62 bilingual networks with attributes type, L1, L2, cross-language edge ratio, L2 inner/crossing edge ratio, bilingual ratio, and proportion of L2.

5.4 Discussion

According to the Global Language System theory, polyglots provide cohesion to the system [25]. In section 2.3, I explain that the cohesion of a social graph depends on the edges that prevent the entire graph from breaking in isolated components [38]. In other words, multilingual users might be preventing the social graph of Twitter from breaking into isolated language groups, or "language bubbles" [46], where information is concealed and similar views reinforced. As motivated in section 1.2, instead of promoting isolated communities, social media sites should seek to expose their users to the unexpected [119] and foster cross-cultural awareness.

This social network analysis reveals how multilingual users are standing between language groups. Reusing the concept of "language bridges" applied by Etling et al. [32] on the blogosphere, multilingual users are forming part of a language bridge between communities in varying degrees. These varying degrees are presented in this chapter as a continuum of increasing connections between the language groups and a continuum of increasing penetration of one language group within the structure of the other. The classification of egocentric networks, or intersections of language groups, could serve to distinguish those egos who might be playing a role as gatekeepers [82] or language brokers [55], and also unveils that not all multilingual users are necessarily in such position. For instance, in the case of the union and integration types many users are connecting both language groups aside from the ego itself.

As a result of this analysis other questions arise: what are the profiles and social contexts of these multilingual users and how they relate to the type of network? For instance, does the integration type reflect a minority or immigrant community in a country? Do small English-posting groups integrated in a non-English group reflect an elite in a country? An example related to the later question can be found in section 3.1, where I reviewed a study on email and Internet chat in Egypt documenting the use of English by a professional elite [108].

The relationship between these types of egocentric networks with the potential impact on information flows remains an open question. It seems reasonable to hypothesize in future research that certain types —like the union and integration categories— might favor cross-lingual linking and dissemination, while other types

—like the gatekeeper category— might be interesting for those seeking purposeful concealment of information.

Methodologically, social network analysis enables going beyond survey information about multilingualism, like the large-scale survey *The Twitter of Babel* [83], and facilitates a deeper understanding about the structural relations between language communities, potentially shedding light into the dynamics of international communication. In this respect, the present study takes a similar approach as *Language Networks on LiveJournal* [53], but enhances the descriptive analysis with the creation and definition of theoretical constructs: the types of intersections between language groups in egocentric networks. Also, this study conceives the egocentric network as a language ecology where the ego is immersed at the micro-scale level, influencing its communication strategy and language choices. This is relevant to the next chapter on social network factors for language choice.

Chapter 6

Factor Analysis

The main goal of this study is to explore how the social network influences the language choices of the multilingual Twitter user. In particular, I tested if we can model the number of times this person (the ego) chooses one language over the other using some characteristics of the egocentric network as predictors. The dependent variables considered are the frequency of English use and non-English language use within the 50 posts of the ego. The frequencies represent the language choices of the multilingual user.

As explained in section 2.5, I consider inter-sentential code-switching when the language changes from one post to the next, while bilingual posts would be cases of intra-sentential code-switching. In this study, I only take into account inter-sentential code-switching and each post represents one interaction. Every post was assigned a language label by the automatic language identification tool and, subsequently, frequency counts of posts in each language were calculated for every ego. Finally, the two most frequent languages of a user were selected to represent his or her options for language choice. English was always one of them due to the sampling conditions.

The factors —independent variables or predictors— are the proportion of English and non-English language users in the social network, and the degree of multi-

lingualism of the social network. The relative importance of factors, or their weight, is represented by the coefficients obtained by fitting two different generalized linear models to the dataset: linear regression, and logistic regression. The main hypotheses are: higher proportions of English users in the network will be a good predictor for more frequent English use by the ego; inversely, higher proportions of non-English language users in the network will be a good predictor of more frequent use of a non-English language by the ego; and the multilingualism of the network will be also a predictor of English use, reflecting its role as a *lingua franca*.

6.1 Operationalization of Variables

The language choices of the multilingual user are defined by two dependent variables: (1) the number of English posts within the 50 (or fewer) posts extracted for each multilingual ego, (2) and the number of posts in other language, called L2. The dependent variables reflect the aggregation of posts at user level, not individual posts. In other words, the models do not consider if one particular post is written in English or L2, but how much or little English a person will tend to use in interactions via Twitter. The factors considered are:

- proportion of English users in the network, represented by the number of speakers labelled as English users and divided by the total number of nodes in the network;

- proportion of users of the most frequent non-English language in the network (L2), represented by the number of speakers labelled as L2 users, and divided by the total number of nodes in the network;

- the *multilingual index* of the network, which represents the degree of multilingualism of the social network.

As suggested by Prof. Jordan Boyd-Graber, the multilingual index can be operationalized as the entropy of a multinomial distribution (formula 6.1). This idea borrows the concept of entropy from Information Theory [94].

In this context, the entropy can be interpreted as the unpredictability of the language used in the network. An entropy close to 0 means that most people in the network are writing in one language, hence the language of the network is more predictable. The more people in the network using different languages, the higher the entropy, reflecting the uncertainty about the language of the network. Unlike providing just the number of languages as a measure of multilingualism, the entropy accounts for the weight those diverse languages have on the network in the form of probabilities.

$$H = -\sum_{i=1}^{n} p_i log(p_i) \tag{6.1}$$

Equation 6.1 for calculating the multilingual index of a social network is borrowed from Shannon's entropy theorem [94]. In this dissertation study, n is the number of languages in the network and p_i is the number of nodes in language i divided by the total number of nodes.

Ego en use	Ego L2 use	N of posts	entropy	net en use	net L2 use
28	19	50	0.79230492	0.51793722	0.45515695
11	38	50	0.67065588	0.63473054	0.35329341
35	15	50	0.42538128	0.84931507	0.10958904
18	6	25	0.69040118	0.47368421	0.52631579
15	35	50	0.57881166	0.7202381	0.26785714

Figure 6.1: Input data file for factor analysis with a reduced number of lines for illustration purposes. The columns represent, from left to right, dependent variables *English use by the ego* and *L2 use by the ego*, number of available posts for the ego, and factors *multilingual index or entropy*, *proportion of English users* and *proportion of L2 users* in the egocentric network.

6.2 Regression Models and Analysis

In this study, I used two different generalized linear regression models to build a probabilistic model that relates a dependent variable y to more than one independent or predictor variable [29].

Formula 6.2 is the *multiple linear regression* equation for three predictor variables, x_1, x_2, x_3: proportion of English users, proportion of L2 users, and multilingual index of the network.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{6.2}$$

In formula 6.2, $\beta_1, \beta_2, \beta_3$ are the regression coefficients. β_1 is interpreted as the expected change in y associated with a 1-unit increase in x_1, while x_2 and x_3 are held fixed [29]. Analogous interpretations hold for β_2 and β_3. The intercept of the fitted line is β_0, which is the predicted value of y when all factors have value 0 [29].

I used the linear regression model for two dependent variables y_{en} and y_{l2}, which are operationalized as the normalized count of posts written in English by the ego (y_{en}) and the normalized count of posts written in L2 by the ego (y_{l2}). Given that not all egos have 50 posts available, the normalization consists of dividing a particular count by the total number of posts available for the ego. However, the output of the linear regression model are numbers from 0 to infinity. For this reason, the linear regression model might not be the best option for this dataset.

Alternatively, *logistic regression* can be used to get probability scores (between 0 and 1) as the predicted values of the dependent variable y [79]. Formula 6.3 is the transformation equation from a linear regression output to logistic regression probabilities, with three predictor variables x_1, x_2, x_3 [79].

$$\log \frac{y}{1-y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{6.3}$$

Like in the previous model, I used the logistic regression model for two dependent variables y_{en} and y_{l2}, which are operationalized as pairs of counts: (y_{en}, N$-y_{en}$) and (y_{l2}, N$-y_{l2}$), where N is the total number of posts available for a user.

I used the R language for the statistical analysis. R is an open programming language and software environment for statistical computing. As a result of fitting these generalized linear models, R outputs the regression coefficients for the independent variables or factors, including the intercept, and indicating positive or negative correlation. In addition, R provides the specific p-value scores for each of the regression coefficients.

Firstly, I used the linear regression function in R to model the use of English by the ego (model 6.4) and the use of L2 by the ego (model 6.5).

$$-lm(Ego.en.use/N.of.posts \sim entropy + net.en.use + net.L2.use) \tag{6.4}$$

$$-lm(Ego.L2.use/N.of.posts \sim entropy + net.en.use + net.L2.use) \tag{6.5}$$

Secondly, I used the logistic regression function in R to model the use of English by the ego (model 6.6) and the use of L2 by the ego (model 6.7).

$$-glm(response \sim entropy + net.en.use + net.L2.use, family = binomial('logit')) \tag{6.6}$$

$$-glm(responseL2 \sim entropy + net.en.use + net.L2.use, family = binomial('logit')) \tag{6.7}$$

In the logistic regression model, the depended variables are operationalized as pairs of counts:

$response < -cbind(Ego.en.use, N.of.posts - Ego.en.use)$

$responseL2 < -cbind(Ego.L2.use, N.of.posts - Ego.L2.use)$

Predictors of Ego.en.use	Estimate	Std. Error	p-value
(Intercept)	0.0002	0.554	1.000
entropy	0.0414	0.160	0.797
net.en.use	0.796	0.502	0.117
net.L2.use	0.075	0.454	0.868

Table 6.1: Linear regression coefficients for modeling the use of English by the ego. None of the coefficients are statistically significant. The proportion of English users in the network is the most important predictor of English use by the ego.

6.3 Results

The results of the linear regression model in table 6.1 do not provide statistically significant coefficients for predictors of English use by the ego. I established the level of significance for the coefficients at a p-value of 0.05. The proportion of English users in the network has the greatest coefficient, indicating this factor is more important for predicting the use of English by the ego. Both the proportion of English users in the network and the multilingual index have positive correlation with the use of English by the ego, as stated in the hypothesis.

Similarly, the results of the linear regression model in table 6.2 do not provide statistically significant coefficients for predictors of L2 use by the ego. The proportion of English users in the network is the factor with the greatest coefficient in absolute value, but is negatively correlated with L2 use by the ego. The proportion of L2 users in the network is positively correlated with L2 use, as stated in the

Predictors of Ego.L2.use	Estimate	Std. Error	p-value
(Intercept)	0.482	0.548	0.382
entropy	0.0335	0.159	0.833
net.en.use	-0.354	0.497	0.479
net.L2.use	0.263	0.449	0.559

Table 6.2: Linear regression coefficients for modeling the use of L2 by the ego. None of the coefficients are statistically significant. The factor proportion of L2 users in the network has a positive correlation with the use of L2 by the ego, whereas the factor proportion of English users has a negative correlation.

hypothesis. However, in disagreement with the hypothesis, the entropy is positively correlated with L2 use. In summary, the proportion of English users and the proportion of L2 users in the network are better predictors of L2 use by the ego than the entropy.

The results of the logistic regression model in table 6.3 include the coefficients for predictors of English use by the ego. The proportion of English users in the network is a statistically significant predictor. Like in the previous model, the proportion of English users in the network has the greatest coefficient, indicating this is the best predictor of English use by the ego. In this model, all the positive and negative correlations of the coefficients are in agreement with the hypothesis, i.e. both the proportion of English users in the network and the multilingual index correlate positively with the use of English by the ego, while the proportion of L2 users in the network correlates negatively.

Predictors of Ego.en.use	Estimate	Std. Error	p-value
(Intercept)	-1.718	0.918	0.061
entropy	0.114	0.265	0.668
net.en.use	2.981	0.832	0.0003 *
net.L2.use	-0.086	0.739	0.907

Table 6.3: Logistic regression coefficients for modeling the use of English by the ego. The proportion of English users in the network is a statistically significant predictor and the most important for predicting English use by the ego.

Predictors of Ego.L2.use	Estimate	Std. Error	p-value
(Intercept)	-0.563	0.899	0.531
entropy	0.302	0.251	0.245
net.en.use	-1.109	0.801	0.170
net.L2.use	1.551	0.737	0.035 *

Table 6.4: Logistic regression coefficients for modeling the use of L2 by the ego. The proportion of L2 users in the network is a statistically significant predictor. Both the proportion of L2 users and English users in the network are important predictors of L2 use by the ego.

Finally, the results of the logistic regression model in table 6.4 include the coefficients for predictors of L2 use by the ego. Using this model, the proportion of L2 users in the network is a statistically significant predictor if establishing the level of significance at $p = 0.05$. Also, the proportion of L2 users in the network has the greatest coefficient, indicating this is the most important predictor of L2 use by the ego. The proportion of English users in the network correlates negatively with the use of L2 by the ego, and the proportion of L2 users in the network correlates positively, as stated in the hypothesis. However, in disagreement with the hypothesis, the entropy correlates positively with the use of L2 by the ego.

6.4 Discussion

In conclusion, the two generalized linear regression models consistently show that the proportion of English users in the network constitutes a key influencing factor in the frequency of English use by the multilingual individual, as stated in the hypothesis. This result was statistically significant in the logistic regression model. Also, the coefficient of this factor was the greatest in the two models.

Similarly, the proportion of L2 users in the network is a very important factor influencing the frequency of L2 use by the multilingual person, in agreement with the hypothesis. This result was statistically significant in the logistic regression model. Also, the coefficient of this factor was the greatest in the logistic regression model of L2 use. Interestingly, the factor proportion of English use in the network is also an important predictor for L2 use by the ego, but is negatively correlated.

Regarding the multilingual index (or entropy), the results are inconclusive about the relation to the language choice of multilingual users. The hypothesis that the entropy could be a good predictor of English use is not confirmed. A future study could deepen into the question of English being used as a *lingua franca* by focusing on multilingual egocentric networks with no monolingual users of English. Controlling this variable can eliminate the confounding influence of users writing in English only.

In essence, these results suggest that the multilingual Twitter users perceive the language composition of their egocentric network and interact accordingly. Or, on the contrary, the language choices of multilingual users might attract followers of a specific language profile. Most probably, the relation goes both ways, in a self-feeding cycle. Social networks evolve over time and users may adapt their language choices in a dynamic relationship with their egocentric network. As Marwick and boyd theorized: "[...] identity on Twitter is constructed through conversations with others. Tweets are formulated based partially on a social context constructed from the tweets of people one follows" [81](11).

Other factors that I initially tested in this analysis were the most frequently used non-English language of the network and the type of network. However, these factors posed specific challenges. In the 92 networks, there were a total of 16 languages that appeared as the most frequently used non-English language. As a consequence, the data were too sparse for any one language to be operationalized as a factor. The type of network structure resulting from the social network analysis was challenging to use as a factor because there are not clear-cut divisions between

types, but stages in a continuum. In future work, it will be interesting to study users' awareness or intuition about their social network type and how this might affect their language choices.

There are other factors influencing language choice that do not fall under the scope of this work: cultural and linguistic context in a particular region, the perceived availability of online resources in a language, social context, language competence, geographic location of the ego, time zone, etc.

Chapter 7

Exploring Textual Features

In this chapter, I shift attention from the social network to the content of the posts written by the egos. There is a convention in Twitter for addressing a message to a particular user or referencing a person, the "mentions", which consists of an @ sign and a username [81]. Honeycutt and Herring [54] studied the @ sign as a marker of addressivity in Twitter; they found that more than 90% of messages with the @ sign were addressed to a user, 5% were referencing a person, and the rest were indicating location or other functions. Surprisingly, they did not comment on the key factor of the mention location within the message to differentiate the posts intended for a specific user: mentions in the beginning of the post are typically used to reply to someone's message.

In the first part of this study, I look at the textual feature of the @ sign at the beginning of a post as an indicator of addressivity, in particular, to distinguish the posts that are replies to an individual. In other words, this indicator can be used to differentiate the type of exchange: sending a public post, including repostings with a comment, and replying to an individual. The objective is to test the hypothesis that the type of exchange is a factor that affects language choice.

The second part of this study takes a qualitative approach to detect the themes that might help in creating cross-cultural awareness, where the multilingual users

might be trying to reach an international audience, acting as mediators from the point of view of their messages. I identify themes related to non-English speaking countries or communities in English posts and, also, I identify English hashtags (keywords preceded by the # sign) inserted in non-English posts. Using a generic theme analysis, this study serves as an explorative qualitative phase to inform the design of future studies after this dissertation work.

7.1 Description of the Data

The dataset used in this study was called "ego dataset" in chapter 4 and includes the last 50 posts (or fewer) of the main 92 subjects, who are multilingual users. In total, the dataset contains 4,423 Twitter posts, associated to their respective authors. Note that these posts are never automatic repostings (using a "retweet" button), due to the requirements during data collection, as explained in chapter 4. The majority of posts have a language label, obtained with automatic language identification. In some cases, there was no language label because the post contained only symbols or URLs, and those were removed before proceeding to the automatic identification of the language. The precise number of posts with a language label is 4,374.

In preparation for this analysis, I revised the language labels with two objectives:

1. eliminating from frequency counts of English automatic posts sent by applications, for example, "Posted a picture on Facebook", "liked a photo on Face-

book", "favorited a Youtube video", "I am at something @ name of place" (foursquare);

2. identifying bilingual posts with the appropriate label. I used the criteria described in section 4.4.2 to classify a post as bilingual, in particular, the post has to meet one of these conditions:

   – one word is in a second language in a post with fewer than five words;

   – two words are in a second language in a post that has between five and ten words included, except if those two words are a named entity;

   – at least three words are in a second language in a post with more than ten words, except when they are a named entity.

Also, the posts were classified in three types of exchange: public post (ToAll), reposting with a comment (RT), or replying to an individual (reply). I designed a simple algorithm for the automatic classification of the posts, using regular expressions, i.e. when a post starts with the @ sign followed by a username is a reply to that person.

Initially, I posed the question whether I will find more *bilingual posts* of the type *reposting with a comment*, thinking of potential translations, which triggered the revision of language labels and posts to identify them. Also, this initial question justified the differentiation between general public posts and repostings with a comment, using the convention of "RT" or "rt" [66]. However, the resulting number of bilingual posts in the ego dataset was very low, 37, and none of them were repostings

with a comment. For this reason, I discarded any hypotheses related to bilingual posts.

To sum up, the dataset used has 4,374 posts, all of which have a language category (English, other), and a category of exchange type (ToAll, RT, reply).

7.2 Hypothesis Testing: Fisher's Exact Test

In this section, I propose to test the following hypothesis: *English is used more when addressing a post to the public in general (ToAll and RT) than when replying to individuals*. This hypothesis is based on the empirical observations of previous research studies on email and mailing lists [31, 65], focusing on the impact on language choice of addressing a message to one person or to a multilingual audience; in the later case English was preferred for its function as a lingua franca.

If a is the number of replies in English, and b is the number of ToAll and RT posts in English, the null hypothesis can be stated as:

$$H_0 : a \geq b \tag{7.1}$$

And the alternative hypothesis is:

$$H_a : a < b \tag{7.2}$$

I set the commonly accepted value of p = 0.05 as the level of significance associated with the null hypothesis.

Honeycutt and Herring [54] estimated that roughly 30% of all Twitter posts contained an @ sign regardless of language. However, they recognized that English

was by far the most frequent language in their sample. More recently, Weerkamp et al. [110] looked specifically at the different proportion of replies depending on the language in Twitter, which varies from 36% of posts being replies in Dutch and 34% in Spanish, to 25% of replies in English, and 13% of replies both in Portuguese and Indonesian. Similarly, in the ego dataset the number of replies to specific users is lower in general than posts addressed to a wider audience. In particular, 22% of English posts are replies, while 35% of posts in other languages are replies.

Given the lower number of replies versus public posts in this dataset, I normalize the counts. Therefore, I compare the proportion of replies that are written in English, $\frac{a}{a+c}$, with the proportion of ToAll and RT written in English, $\frac{b}{b+d}$, where c is the number of replies in other languages, and d is the number of ToAll and RT posts in other languages.

The results show that the null hypothesis seems to be untrue: the proportion of replies in English is $\frac{a}{a+c} = 0.3810$, while the proportion of public posts written in English is $\frac{b}{b+d} = 0.5368$. To reject the null hypothesis, I have to apply a statistical test of significance. Since I am comparing the categories of the posts, the most appropriate non-parametric test is the Fisher's Exact Test [36].

In the Fisher's Exact Test, the data can be displayed in a 2x2 contingency table (table 7.1). The probability of obtaining any such set of values is given by the hypergeometric distribution in equation 7.3.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!} \tag{7.3}$$

	Replies	ToAll+RT	Language total
English	a=482	b=1669	a+b=2151
Other	c=783	d=1440	c+d=2223
Type of exchange total	a+c=1265	b+d=3109	n=4374

Table 7.1: 2x2 contingency table for the Fisher's Exact Test.

The resulting p-value is 2 e^{-21}, which is much lower than the set value. In conclusion, we can reject the null hypothesis in favor of the alternative hypothesis: in the case of multilingual Twitter users, they use English more frequently in public posts than in replies to individuals.

7.3 Discussion: Addressivity as a Factor

This result reinforces the idea that addressing an online message to a perceived multilingual audience encourages the use of English. In Twitter, the only previous study that has looked at addressivity as a factor for language choice focused on Welsh-English bilinguals [61]. The study by Johnson [61] also found that they use proportionally more English in public posts (53%) than in replies to individuals (44%) in a sample of 500 posts. The author speculates that the use of English is encouraged in Twitter for its potential to reach a wider audience [61]. The work by Johnson reported very few cases of bilingual posts [61], which is consistent with the description in section 7.1.

In section 3.3, I reviewed two works that suggest Twitter is used more for conversational purposes in some languages, with higher frequency of @ signs, while in other languages is more common to use it for sharing resources, as the higher frequency of URLs and repostings might indicate [110, 55]. Taking these previous findings into consideration, future work should study the combination of language profile of the user and addressivity of the message to understand language choices in Twitter.

The lesson for system designers is that different types of exchange between people, with the corresponding number of sources and receivers (one or many), could prompt *code-switching*, a user changing the language.

7.4 Theme Analysis

In chapter 5, I studied the social connections between language groups, but ultimately, I am interested in understanding the complementary roles of social connections and topic-based linkages in creating cross-cultural bridges.

This qualitative analysis constitutes a first exploration of themes that could potentially connect communities speaking different languages. I use the ego dataset, which was not initially collected with the idea of detecting communities around a topic, instead the purpose was to gather multilingual users regardless of their interests. With this limitation in mind, as well as the numerous languages present in the dataset, I decided to focus on identifying keywords related to non-English speak-

ing countries or communities in English posts, and English "hashtags" (keywords prefixed with the pound sign) inserted in non-English posts.

Previous works [24] have attempted to develop various classification schemes of content in Twitter, but they are too broad for the purposes of this study. Dann [24] focuses his review on five classification schemes and complements them with his own scheme of six general categories, namely conversational, pass along, news, status, phatic, spam, and 24 subcategories, like response, location, endorsement, headlines, sports, events, etc. In this section, I do not base my analysis on these existing classification schemes, instead I read through the data without prior preconceptions in search of answers to the questions: why is the user mentioning a non-English speaking place or a non-English language? What is she/he saying about it? Why is the user inserting an English hashtag in a non-English post?

The descriptive nature of this generic theme analysis intends to stimulate more systematic research questions and classification schemes to discover the topics and feelings that bring multilingual online communities together.

7.4.1 International themes in the English language set

Firstly, while reading the 2,151 posts in English, I identified the names of places, cities, countries, or languages that are not English speaking. At the same time, I annotated the context in which they were mentioned. After a second revision, I used the annotations about the context as the basis for the following emergent themes:

- international news, which include links to media sources and reactions (i.e. “Arab Revolution power”, “The Ukrainian experience for Arab world by Lionel Beehner ”, “Hope that everyone in Japan is fine... ”);

- people’s travel plans or an accomplished travel (i.e. “can’t wait to fly to #Barcelona”, “Now considering Martinique & Guadeloupe for my next holiday”, “Hi back from germany”, “off to Geneva for a show at a fancy birthday party”);

- people’s location (i.e. “ It’s 3AM here in France so I’m kind of tired”, “[...]you are at the Schokoladenmuseum”, “[...]we need you in Oberhausen! [...]How long will you stay in Hamburg?”);

- event’s location (i.e. “Buskers Festival in Bern was fantastic!”, “annual Swiss Congress of Radiology”, “Global Performing Arts Exchange Singapore”, “Tomorrow is our concert in Dresden”);

- an opinion, of political kind or not (i.e. “miserable greek reality”, “I am for communism. Swedish communism”, “bahrain tonight should burn...”, “Given the birth rate & z population figures in Egypt, I can’t understand how sex is a taboo [...]”);

- internationalization of technology (i.e. “[...]software penetrating Angola”, “Portuguese RTS game for PS3 [...]”, “I use it in China to redirect my website”, “@adobe should organise an event for nord africa just like @google”);

- Sports (i.e. “German Rugby”, “Forza Milan!”);

- Culture (i.e. “Nigerian Clubbing etiquette”, “Victor Manuelle is a Puerto Rican artist!”, “Israeli band Orphaned Land rocks Turkey”);
- Language, including remarks about the language of a resource linked when is different from English (i.e. “New Review 9/10 Points (German)”, “Afrikaans Video”);
- gastronomy and restaurants (i.e. “Turkish scrambled eggs...” , “all you can eat at thai-jap rest...”);
- travel recommendations (i.e. “If a Lille visit is on the agenda [...] I’d highly recommend the LAM museum”, “see my latest thoughts on ‘Where To Stay’ in Istanbul”, “You should come to Israel during the summer, you’ll have a blast!”);

There are 227 posts in which a non-English speaking place or language is mentioned. Table 7.2 displays the frequency of the themes. The complete list of places and languages, with an extract of the textual context and associated theme can be found in Appendix B.

Looking at this list of themes found in English posts, I wondered which ones are drawing attention to countries and communities that are not English-speaking and providing some information. Generally, posts about people’s locations (32 instances) or travel plans and accomplished travels (33 instances) do not provide information about those places, and I speculate the function of these posts relates more to coordinating with friends than to draw attention to a culture.

Theme	*Instances*
International news	31
Reaction to international news	4
Travel plans	29
Accomplished travel	4
People's location	32
Events' location	27
Opinion	24
Tech internationalization	14
Sports	13
Culture	12
Language	8
Language of resource	4
Gastronomy and restaurants	10
Travel recommendations	9
A country's policy	2
Location and language	1
Culture and humor	1
Travel plans and opinion	1
Movies	1

Table 7.2: Frequencies of themes related to non-English speaking places and languages mentioned in English posts.

Likewise, the posts related to events' locations (27 instances) might have a primary coordinating function, but in some cases (eg."oktoberfest") the post is drawing attention to events of cultural interest in a country and could be considered as fostering cross-cultural awareness.

The theme *international news*, which includes also the reactions to the news, is the theme that most frequently fosters cross-cultural awareness in the ego dataset (35 instances). These posts show the interest of the author in a non-English speaking community and provides information to the followers. *Opinions, of political kind or not* (24 instances) is the subsequent most frequent theme in raising cross-cultural awareness. In this case, the author also shows interest in a non-English speaking community and provides some piece of information, albeit the impressions are sometimes negative. More scarce, there are posts referring to *Sports* (13 cases), *Culture* (12 cases), *gastronomy and restaurants* (10 cases), and *travel recommendations* (9 instances) of countries that are non-English speaking, which also provide some information about cultural aspects.

Aside, the theme *internationalization of technology* is interesting in its own right because it reflects on technology adoption in different areas of the world. Although the instances are low in this dataset (14 cases), it points to a promising application for tracking the adoption of technology products and services at an international scale.

Finally, it is worth noting that among the 12 mentions of a language different from English, there are 4 instances in this dataset where users specify the language of a resource they are linking. They are creating a cross-language link, a phenomenon

that has been studied on the blogosphere [53, 47]. Cross-language linking, like the cross-language social connections studied in chapter 5, contribute to building paths between communication spheres online that are separated by language, enabling the flow of information across national boundaries.

In summary, even though the mention of a non-English place or language does not always come with information about it, some themes might be facilitating cross-cultural awareness, like international news, sports, and culture. Unfortunately, sharing of resources in languages other than English into the English language sphere on Twitter seems to be infrequent, or at least the language notices preceding the link. Alternative data collection techniques could shed light into this phenomenon in future research.

7.4.2 English hashtags in the non-English language set

In a second phase, I read the 2,223 posts that were left out from the English language set. However, among those, there are posts in English that were classified as automatic (see section 7.1). I ignored such posts for the purposes of this analysis. Aside from English, there are 18 languages in the ego dataset. Given the challenge that so many languages posed for the analysis, I focused on identifying English "hashtags" only, keywords or phrases prefixed with the pound sign (#).

The purpose is to explore the reasons why a user writing in a language different from English might want to add an English word or phrase. Are they examples of

code-switching? Are they a mechanism to draw international attention? Or are there other motivations behind?

The hashtag was a convention of the Internet Relay Chat (IRC) introduced and accepted by users as the Twitter tagging feature in 2007 [15]. Hashtags are used to "funnel related tweets into common streams" [57], by aggregating posts with a common hashtag. It is important to bare in mind that the hashtag can be a means to classify a message, but also provides the user with visibility in a "many-to-many" conversation about a topic, potentially enabling the user to reach an audience beyond their followers. The Twitter system leverages hashtags for creating lists of popular topics in real time, called "trending topics". These trending topics are visible by all users, thus potentially drawing worldwide attention. As a side note, the proportion of Twitter posts with hashtags seems to vary with language, for example, a study reports that in German they account for 25% of posts, 14% in English, and as low as 4% in Japanese [110].

I created a list with the identified hashtags in English, and also I included hashtags of brand names and products, names of places in English or transliterated into latin script, and many acronyms. Acronyms posed a particular challenge, given that many of them were just informal ad-hoc abbreviations and required search and documentation to understand the meaning. Often, they were abbreviations of conference names, music festivals, and other events with a potential international audience. Occasionally, I could not determine the meaning of the acronym, in consequence I did not include those on the list of selected hashtags. Also, I discarded acronyms that were referring to local events or institutions (eg. German institution

#rbb, Portuguese political debate #e2011pt), when they were specific of the culture and language in which the author was writing and did not constitute a change of language or script.

Subsequently, I classified the hashtags using the annotations about their meaning. I tried to classify them in topics, like *Information and Communication Technologies*, and *political topic and campaigns*. However, some hashtags have a primary conversational function instead of referring to a topic, as studied by Huang et al. [57]. They argue that these types of tags "serve as a prompt for user comment" and "the resulting content is an asynchronous massively-multi-person conversation"; also, they provide a few examples, like #igrewupon, #liesmentell [57]. The *conversational tags* of Huang et al. [57] correspond to what Laniado and Mika [70] called "tags characterized by a common sentiment", which they illustrate with examples such as #thankfulfor and #youknowyouareuglyif. They estimated that 20% of Twitter hashtags belong to this type. Also, many hashtags refer to named entities. Laniado and Mika [70] estimated that 39% of hashtags are named entities, most commonly referring to organizations, products, and events.

As a result of this diversity, first I classified hashtags in two main groups: conversational tags, and the rest. Secondly, I divided the conversational tags in three types: emergent discourse conventions in Twitter (eg. #fail, #wtf), reflecting on Twitter use (eg. #1000followers, #odd_trend), and informal or ad-hoc Twitter genres (eg.#a_thought, #kindlyAdvice, #roadrage). Finally, I classified the rest of the hashtags into groups of topics, and grouped aside the brands, devices, events,

locations, and dates. The tables of hashtags, including frequency of appearance and the categories in which they are classified, can be found at the end of this section.

A prominent group of English hashtags relates to *Information and Communication Technologies* (eg. #mobileusers, #Android, #Cloud Computing, #hyperlinking), which also spans brands and devices, such as #microsoft, #skype, #Ipod (see table 7.5). Among these hashtags, there are named entities, but also cases of intrasentential code-switching, where the user switches from one language to another within the same sentence [62]; following Joshi's definition [62], English is the embedded language within a non-English matrix language in these examples.

While the topic of Technology and Internet seems to trigger the use of English terms, international news (eg. #bahrain, #egypt), and campaigns, such as #1billionhungry, might be biased towards English hashtags to draw global attention (see tables 7.7 and 7.8). International news could potentially lead to a trending topic and draw attention about events unfolding in real time in some part of the world, like it was the case during the popular uprising in Egypt during 2011 [16].

There are also numerous hashtags referring to conferences and music festivals. If the organizers of such events announce or promote a hashtag, they avoid the problem of fragmentation of message streams related to the event due to variations of keywords [15]. However, in this sample there are variants of a hashtag for the same event, such as #caexpoitaly and #caexpo, #gmaghreb and #gmaghreb11, #sepIn11 and #sepIn (see table 7.6). Letierce et al. [72] studied the use of Twitter and hashtags in conferences of Semantic Web researchers and revealed that users have a desire to participate in the discussion around the conference and see their

messages included in the conference stream, while hoping to increase their network. They concluded that Twitter is used as a background communication channel during those events.

Most interestingly, international conferences and music festivals attract multilingual and multicultural audiences, who can conform to the same Twitter hashtag and generate multilingual conversations around the event taking place. These international events might be key in promoting cross-language sharing of resources and creating social ties across language and national boundaries in the online communication sphere.

There are still other reasons to use an English hashtag when writing in a different language: as studied by Huang et al. [57], some hashtags are prompts for user comment and a way to participate in a multi-person conversation. A conversation that can be multilingual.

A few of these conversational tags constitute emergent discourse conventions in Twitter, even adopted from other online sites, such as the commonly used #like in the social networking site Facebook. Kooti et al. [66] studied the emergence and evolution of this type of conventions looking at the specific example of reposting in Twitter. They provided data showing how the user community was choosing to include "RT" in their repostings more often than other alternatives over time. In other words, Twitter users have progressively agreed in the use of certain codes, such as adding #fail to their post when they talk about disappointing or deceiving news (7 times in the non-English sample), #wtf to express disbelief, #FF or "follow friday" to recommend other users, etc (see table 7.3). Even though these conventions

come from the English language, they have been adopted in other languages for communicating in Twitter.

Similarly, users in the non-English sample sometimes categorized their posts in an "informal or ad-hoc genre" by adding an English hashtag, like #a_thought, #kindlyAdvice, #thisislife, #roadrage (see table 7.4). These examples are along the lines of those presented by Huang et al. [57], and Laniado and Mika [70], referring to common sentiments, but also less persistent over time than the previously discussed emergent discourse conventions. These informal or ad-hoc genres in English seem to have the potential to spread internationally and be adopted across languages in Twitter, but this phenomenon is still not well documented.

In summary, some English hashtags reflect *code-switching* in relation to certain topics, like Information and Communication Technologies. A question for future research could be if this topic affects language choice, favoring the use of English. Also, international news and campaigns might tend to trigger the use of English hashtags to draw global attention. Finally, international events organizing back-channel comments around a common hashtag, as well as certain conversational tags, could be the focus of further research for their potential to foster multilingual conversations.

Hashtag	*Frequency*	*Meaning/Context*
#fail	7	Sharing a bad experience or deceiving news
#wtf	2	Expressing surprise or disbelief
#FF	1	"follow friday": recommending a person or organization to follow
#Like	1	
#np	1	now playing or no problem

Table 7.3: Conversational tags: emergent discourse conventions in Twitter, social networks or online chat.

Hashtag	*Frequency*
Reflecting on Twitter use	
#TweepsMidName	5
#1000followers	1
Twitter #addicted	1
#odd_trend	1
#summer trends	1
Informal or ad-hoc Twitter genres	
#a_thought	1
#kindlyAdvice	1
#thisislife	1
#roadrage	1
#supportedby	1
#UNeverKnow	1

Table 7.4: Conversational tags: reflecting on Twitter use and informal or ad-hoc Twitter genres.

Topic	*Hashtag*	*Frequency*	*Meaning/Context*
ICT	#mobileusers	6	
	#cisa	2	Online live broadcasting
	#OVH	2	Online virtual hosting
	#Android	1	
	#AR	1	Augmented Reality
	#Cloud Computing	1	
	#Honeycomb	1	Android version
	#hyperlinking	1	
	#launch	1	a website
	#opendata	1	
Devices	#fb_funerals	2	Facebook
	#n900	1	Nokia model
ICT brands	#Ipod	1	
	#microsoft	1	
	#samsungcheerdance	1	Samsung
	#skype	1	
Vehicle brand	#Audi	1	

Table 7.5: Hashtags: ICT topic, brands and devices.

Topic	*Hashtag*	*Frequency*	*Meaning/Context*
Conferences/events	#caexpoitaly	9	CA expo 2011
	#mcdd10	7	Mobile Camp Dresden
	#caexpo	5	CA expo 2011
	#gmaghreb	2	Google event in Maghreb
	#gmaghreb11	1	Google event in Maghreb
	#innovlab2011	1	
	#pycon4	1	
	#SMW11	1	Social Media Week 2011
	#sepIn11	1	
	#sepIn	1	
Music festivals	#212RMX	2	
	#fib2011	2	
	#readingandleeds	1	
Music	#arcticmonkeys	1	Band
	#FearFactory	1	Album
TV and sports	#BlueWolves	1	Mongolian soccer team
	#Comedystreet	1	German TV program
	#tv	1	

Table 7.6: Hashtags: events, music, TV, and sports.

Topic	*Hashtag*	*Frequency*	*Meaning/Context*
Location	#bahrain	12	
	#berlin	2	Germany
	#egypt	2	
	#greece	2	
	#germany	1	
	#korinthos	1	Greece
	#Mongolia	1	
	#liveyourmythingreece	1	Greece
	#Tahrir	1	Egypt
	#uniineurope	1	Europe
Dates/time	#14feb	1	events in Bahrain
	#september	1	
	#winter	1	
Celebrations	#Jerusalemday	1	
	#Ramadan	1	

Table 7.7: Hashtags: location, time, and other named entities.

Topic	*Hashtag*	*Frequency*	*Meaning/Context*
Project management	#marketing	1	
	#scrum	1	
Political/Campaign	#debtocracy	13	
	#1billionhungry	1	
	#endSH	1	End Street Harassment
	#sexquota	1	
Not classified	#networking	1	
	#selective_default	1	

Table 7.8: Hashtags: other topics.

Chapter 8

Discussion and Future Work

I have encountered many challenges in this explorative research that require bringing together multiple fields and diverse methods in ways that have not been established previously. For instance, one of these challenges was detecting multilingual users in Twitter and, more broadly, determining language profiles of users (if they are monolingual or multilingual), and assigning language labels accordingly. An example of the multidisciplinary character of this dissertation is the application of social network analysis to sociolinguistic questions.

As a result of the decision process for resolving the challenges as they became apparent, a concomitant contribution of this research are the methodology considerations. Namely, the process of testing natural language identification tools, the relationship between number of posts per user analyzed and estimated error rates in language profiling, can serve as a guide to approach the problems arising in the study of languages in Twitter.

Another challenge of doing research about Twitter (and other Internet platforms) is that the findings on particular aspects of this fast evolving environment can easily become outdated. This dissertation aimed at answering the research questions without requiring an exhaustive revision of the Twitter interface, which has changed several times in the past years, and at obtaining conclusions that could

be informative for the study of other social networks and communication environments that share the key characteristics of Twitter, like being public, reposting and replying capabilities, etc.

This chapter discusses the results of the studies that compose this dissertation, and highlights the key contributions and future directions of this research. Ultimately, I hope this discussion provokes questions for new studies.

8.1 Of Links, Social Ties, and Gravitational Forces

The vision of a cosmopolitan Internet with vibrant communities, enabling contact with the unfamiliar, discovery, and the serendipity that propitiates learning [119] is challenged by the existence of language frontiers online.

In the view of the Global Language System theory (section 2.1), multilingual people constitute the gravitational force that provides cohesion to the system, by connecting different language groups. There is empirical evidence of this language bridging in the blogosphere [53, 32].

The language ecology approach (section 2.2) connects these macro-scale dimension of languages with the micro-scale level of interactions between individuals. Social network analysis provides an analytic tool for studying these language ecologies that emerge from the interactions of the multilingual users with their social connections.

The main contribution of this dissertation is going beyond survey information about multilingualism in Twitter [83, 55], and providing a deeper understanding

about the structural relations between language communities in a social network online. Although inspired by previous studies on the blogosphere [53], this research enhances the descriptive analysis with the creation and definition of theoretical constructs: the types of bilingual networks.

Focusing on the networks of multilingual users, the social network analysis revealed three types of bilingual networks: the *Gatekeeper-Language bridge*, representing a continuum of increasing connections between two separate language groups; the *Integration and union* type, representing a continuum of increasing penetration of one language group within the structure of the other; and the *Peripheral language* type, where one language group is smaller or less cohesive, and lies at the periphery of the social graph.

This research conceives of the social network of multilingual users as a micro-scale language ecology, influencing their communication strategies and language choices. This conceptualization leads to a second key contribution, which is the novel idea of modeling the influence of social network factors in the language choices of the user.

In the factor analysis, the dependent variables considered are the proportion of English use and non-English use within the posts of the user. The factors included are the proportion of English and non-English language users in the social network of the multilingual subject, and the degree of multilingualism of the social network. The relative importance of factors is represented by the coefficients obtained by fitting two generalized linear models to the dataset (linear and logistic regression).

The proportion of English users in the network constitutes a key influencing factor in the frequency of English use by the multilingual individual. Similarly, the proportion of non-English language (L2) users in the network is a very important factor influencing the frequency of L2 use by the multilingual person. The results suggest that multilingual Twitter users perceive the language composition of their network and interact accordingly. Or on the contrary, the language choices of multilingual users might attract followers of a specific language profile. Most probably, the relation goes both ways, in a self-feeding cycle.

Shifting attention from the social network to the content of the posts written by the multilingual users, I tested the hypothesis that the type of exchange (public post versus reply to an individual) influences the choice between English and other languages. The result reinforces previous empirical findings suggesting that sending public messages to a seemingly multilingual audience encourages the use of English [31, 65].

Finally, there is another gravitational force that could connect language groups and affect language choice: topics [65, 5]. Common interest in certain topics attract people from different cultures, and encourages the creation of cross-language links to resources and news [47].

As a step toward future studies on international topics, this dissertation explores what themes might be raising cross-cultural awareness. I identified themes related to non-English speaking countries or communities in English posts, and I concluded that international news was the most popular theme.

Also, I identified English hashtags (keywords preceded by the # sign) inserted in non-English posts and related contexts that could encourage multilingual conversations. International conferences and music festivals attract multilingual and multicultural audiences, who conform to the same Twitter hashtag and generate multilingual conversations around the event taking place. These international events might be key in promoting cross-language sharing of resources and creating social ties across geographic regions.

If we embrace the idea of a vibrant language ecology on the Internet, we should challenge the existing structure of the network of hyperlinks and social ties. For instance, empowering multilingual users to leverage their social ties across language groups, facilitating translation, and recommending links to resources in different languages.

8.2 The Road Ahead...

Future directions for this research include scaling up the social network analysis to account for multilingual users with larger social networks. This will require improving the methods for analysis of larger collections of data, e.g. training natural language identification tools to detect transliterated text, and using spam detection algorithms.

Also, I envision expanding the theme analysis to include methods of automatic topic detection in multiple languages, or crowdsourcing annotations using platforms such as Mechanical Turk or CrowdFlower (i.e. sending micro-tasks to large numbers

of people for specifying the topics of Twitter posts). Further research could focus on topic-based networks, targeting the sampling to specific language pairs and topics for enabling comparisons across languages and a more complex factor analysis.

Finally, studying the evolution of social networks over time could unveil the relationship between the language composition of the social network, audience perception, and language choice. In relation to this, other questions arise: wether multilingual users are aware of the type of social network they have; and if they are, how this affects their language choices.

8.2.1 Translation and Mediation in Twitter

In section 7.1, I describe my unsuccessful attempt to generate a hypothesis in relation to *bilingual posts* and *repostings with a comment*, as a previous step to identify translation behaviors. However, the resulting number of bilingual posts in the ego dataset was only 37. It seems that the limited number of characters allowed poses a problem for including a translation together with the original comment in the same post. Alternatively, translations might be found in separate posts but, unlike reposting, there is no way to connect the translation to the original message. Also, some people create separate accounts for each language to address different audiences. Future research could use automatic topic analysis to identify translations.

Although Twitter enables the use of many languages and writing systems thanks to Unicode, it does not offer support for translation or features for strengthen-

ing connections between language groups. Regarding support for translation, there are not embedded linguistic resources on the interface, such as machine translation, dictionaries or transliteration tools. The *meedan* project [113] organizes volunteers for translating Twitter posts and has encountered a number of challenges: engaging users in translation, linking and representation of translations in relation to the original post, authorship, validation, etc.

Instead of relying on volunteer translators, we could seek ways to encourage translation, cross-language linking and connection behaviors that are happening already. However, in Robert Munro's words, there is not a "unified resource that links people by languages spoken" in social media [96], which would be a helpful starting point.

Recommendation mechanisms could foster the creation of cross-language links. For example, AlMeshary and Abhari [4] propose a strategy for recommending people to follow on Twitter with the purpose of obtaining local information in the context of a user relocating from a different country. They use machine translation to match the users' interests found in their posts with the local offers.

Finally, by studying the dynamic language preferences of multilingual users, not only we will be in a better position to design a satisfying experience for those users, but also we are learning how to help them in their mediation tasks. This dissertation advances in that direction by modeling the influence of factors in the language choices of the multilingual users.

8.2.2 Who Are the Multilingual Users?

This dissertation focuses on multilingual users because of their role in connecting different language groups. But who are they? Are they expatriates? Members of minority communities? Language learners?

Twitter posts and the languages in which they are written represent just a limited language profile of the user, and they barely provide any social context. Androutsopoulos recommends to take into account the digital surroundings when analyzing written text, for instance, looking at the pictures and videos that are linked [6]. Also, adding detailed geographic information could help in building a more complete profile.

Understanding more about the context of multilingual users could help in the identification of roles and motivations for mediating between language groups and in finding the relationship of these roles with network types.

The next step after this dissertation is adding consideration of geolocation information and content analysis of the resources linked in the posts to provide more attributes for nodes and edges in the social network analysis. Additionally, ethnographic methods could shed light on who are the people and what are the reasons that connect different language groups.

Chapter 9

Conclusion

Social media is international: users from different cultures and language backgrounds are communicating, generating and sharing content. However, language barriers emerge in the communication landscape online. The aspiration of an Internet that constitutes a cosmopolitan space and fosters language diversity has stumbled over the language frontier.

In the microblogging site Twitter, information spreads across languages and countries. But how are the news traveling across borders? Expatriates, migrants, minorities, diasporic communities, and language learners play an important role in forming transnational networks and cultural bridges between nations and communities. They are multicultural and multilingual.

This dissertation studied how multilingual users of Twitter mediate between language groups in their social network, looking at social connections and language choices. The overarching goal that motivates this research is to advance our understanding of the network structures and communication strategies that enable intercultural dialog, cross-language sharing of information, and awareness of global problems. The implication for the design of social media platforms is that, instead of constraining multilingual users to only one language option, technology should support their language-switching and mediating role between cultures.

The **objectives** of this dissertation were: (1) to explore the ways in which multilingual users of Twitter are connecting different language groups in their social network; (2) to model how the network influences their language choices; (3) and to explore what the textual features of their posts can elicit about language choices and mediation between groups.

RQ 1: *In what ways are multilingual users of Twitter connecting language groups?* Focusing on the social network of 92 multilingual users, the methodology combined a qualitative approach to social network analysis and network statistics to present a classification of network types based on the patterns of connections between language groups. The study followed an exploratory design, with a first qualitative phase that took a grounded theory approach to classify the network visualizations, and a second quantitative phase that complemented the qualitative study with network statistics specifically created to provide a robust definition of network types. Finally, I used machine learning for testing the results.

The social network analysis revealed three differentiated types of bilingual networks: the *Gatekeeper-Language bridge*, representing a continuum of increasing connections between two separate language groups; the *Integration and union* type, representing a continuum of increasing penetration of one language group within the structure of the other; and the *Peripheral language* type, where one language group is smaller or less cohesive, and lies at the periphery of the social graph.

RQ 2: *How is the social network of multilingual users in Twitter influencing their choice of language?* The factor analysis modeled the influence of a set of factors related to the social network in the language choices of multilingual users. The

dependent variables considered are the proportion of English use and non-English use within the 50 posts of the user. The factors included are the proportion of English and non-English language users in the social network of the multilingual subject, and the degree of multilingualism of the social network. The relative importance of factors, or their weight, is represented by the coefficients obtained by fitting two generalized linear models to the dataset (linear and logistic regression).

The proportion of English users in the network constitutes a key influencing factor in the frequency of English use by the multilingual individual. Similarly, the proportion of non-English language (L2) users in the network is a very important factor influencing the frequency of L2 use by the multilingual person. Interestingly, the influence of the factor proportion of English users in the network is also important when modeling L2 use, and negatively correlated to it. Regarding the multilingual index, the results were inconclusive about its influence in the language choice of multilingual users.

RQ 3: *Does the type of exchange in Twitter influence the language choice of multilingual users?* I shifted the attention from the social network to the content of the posts written by the multilingual users. First, I looked at the textual feature of the @ sign at the beginning of a post as an indicator of addressivity. Based on this indicator, I tested the hypothesis that the type of exchange (public post versus reply to an individual) influences the choice between English and other languages. The result reinforces previous empirical findings suggesting that sending public messages to a seemingly multilingual audience encourages the use of English.

RQ 4: *What the themes and textual features in the posts of multilingual users reveal about cross-cultural awareness or international dialogue?* Finally, I looked at content with the objective of detecting themes that might help in creating cross-cultural awareness, where the multilingual users could be acting as mediators from the point of view of their messages. I identified themes related to non-English speaking countries or communities in English posts and, also, I identified English hashtags (keywords preceded by the # sign) inserted in non-English posts. Using a generic theme analysis, I concluded that international news was the most popular theme when mentioning a non-English speaking place. This study serves as an explorative qualitative phase to inform the design of future studies after this dissertation work.

The main **contribution** of this dissertation is going beyond survey information about multilingualism and providing a deeper understanding about the structural relations between language communities in a social network online. This research work is one of the few that apply social network analysis to the study of sociolinguistic questions on the Internet. In particular, it contributes an original classification of network types based on the patterns of connections between language groups, complemented with new network statistics that enhance the definitions of these theoretical constructs.

Adapting the *Ecology of Language* approach from Sociolinguistics to the social network context, this research conceived of the social network of multilingual users as a micro-scale language ecology, influencing their communication strategies and language choices. This conceptualization led to the novel idea of modeling the influence of social network factors in the language choices of the user.

This dissertation can benefit the study of information diffusion regarding the potential impact of these types of network structures on cross-language flows. Also, it contributes to understanding users' behavior and informing the design of social media platforms.

Future directions for this research include: scaling up the social network analysis to account for multilingual users with larger social networks; studying topic-based networks and detecting cases of translation; targeting the sampling to specific language pairs and topics for enabling comparisons across languages and a more complex factor analysis; studying the evolution of social networks over time to explore how this affects audience perception and language choice.

The next step to this dissertation research is adding geolocation information and content analysis of the resources linked in the posts to provide more attributes for nodes and edges in the social network analysis. Finally, ethnographic methods could shed light on who are the people and what are the reasons that connect different cultural and linguistic groups.

Appendix A

Visualizations of Social Networks

This appendix contains the visualizations of the 92 egocentric networks, with the qualitative category assigned, the language codes and corresponding colors. Language labels can have one language code, following the ISO standard codes for names of languages (eg. "en" for English, "es" for Spanish, "de" for German), two language codes joined by the + sign in the case of bilinguals, the word "empty" for nodes with no data available, the number 0 for nodes where the language could not be identified. All visualizations were made with the Gephi social network analysis tool, using the Force Atlas layout. The size of the nodes represents betweenness centrality.

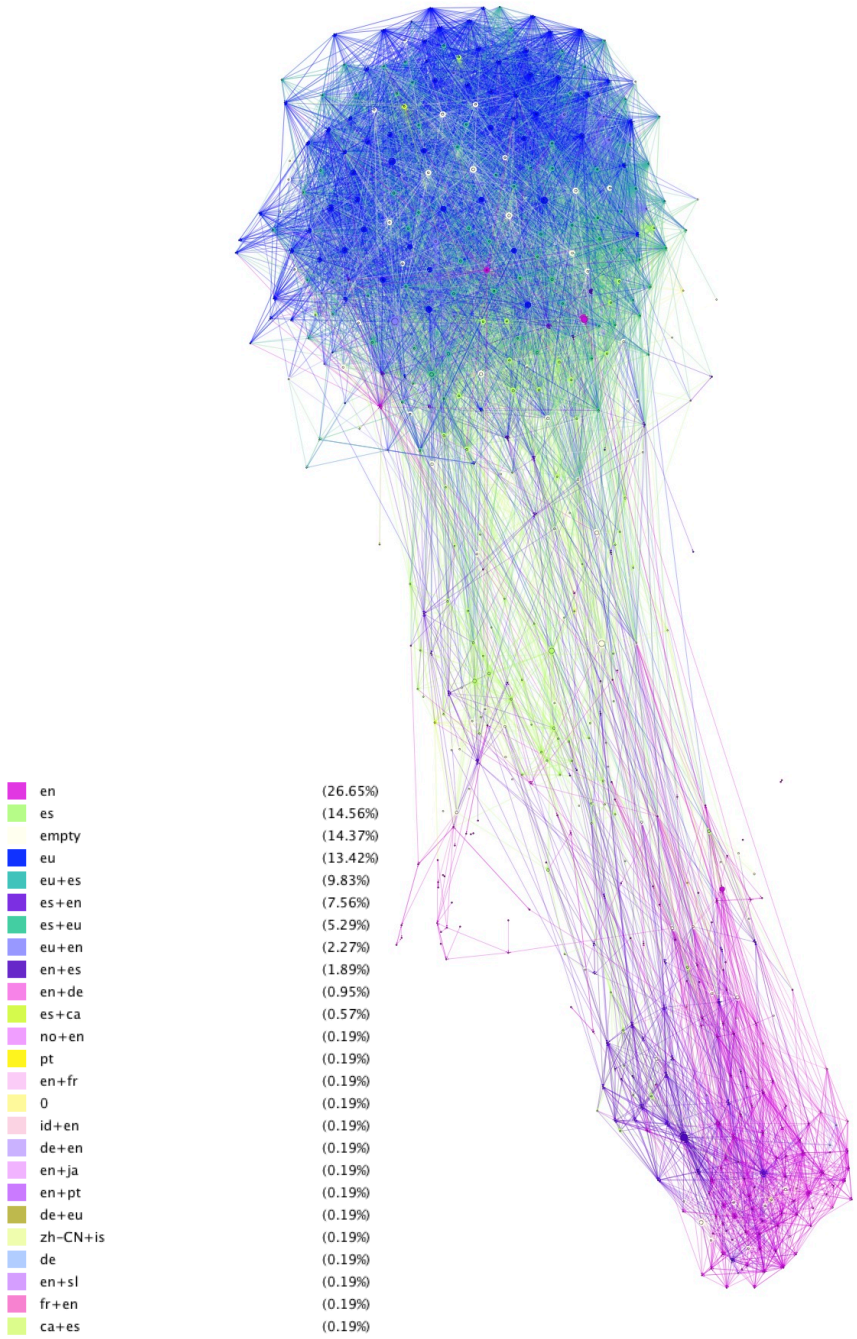


Figure A.1: Trilingual networks (1).

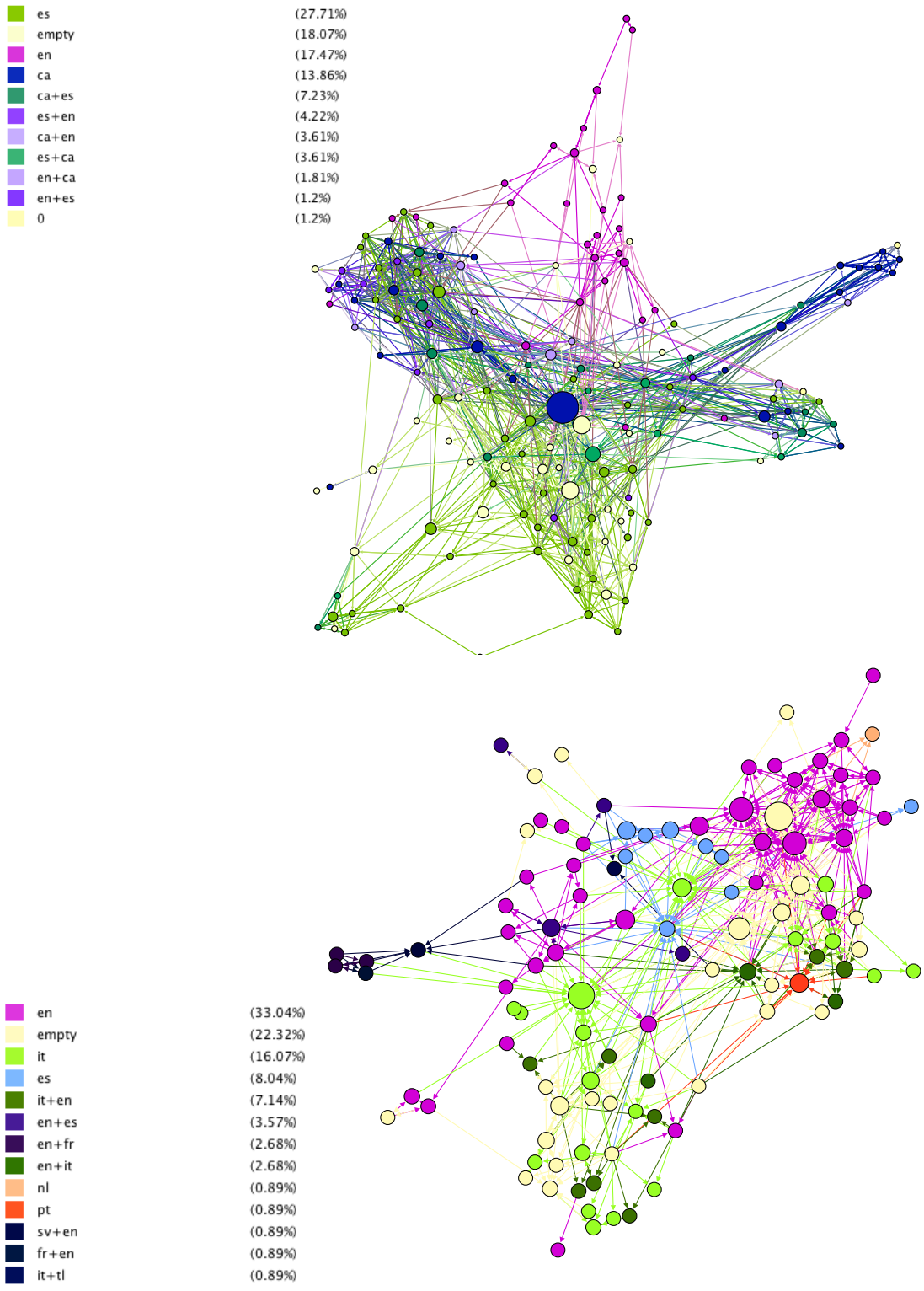


Figure A.2: Trilingual networks (2).

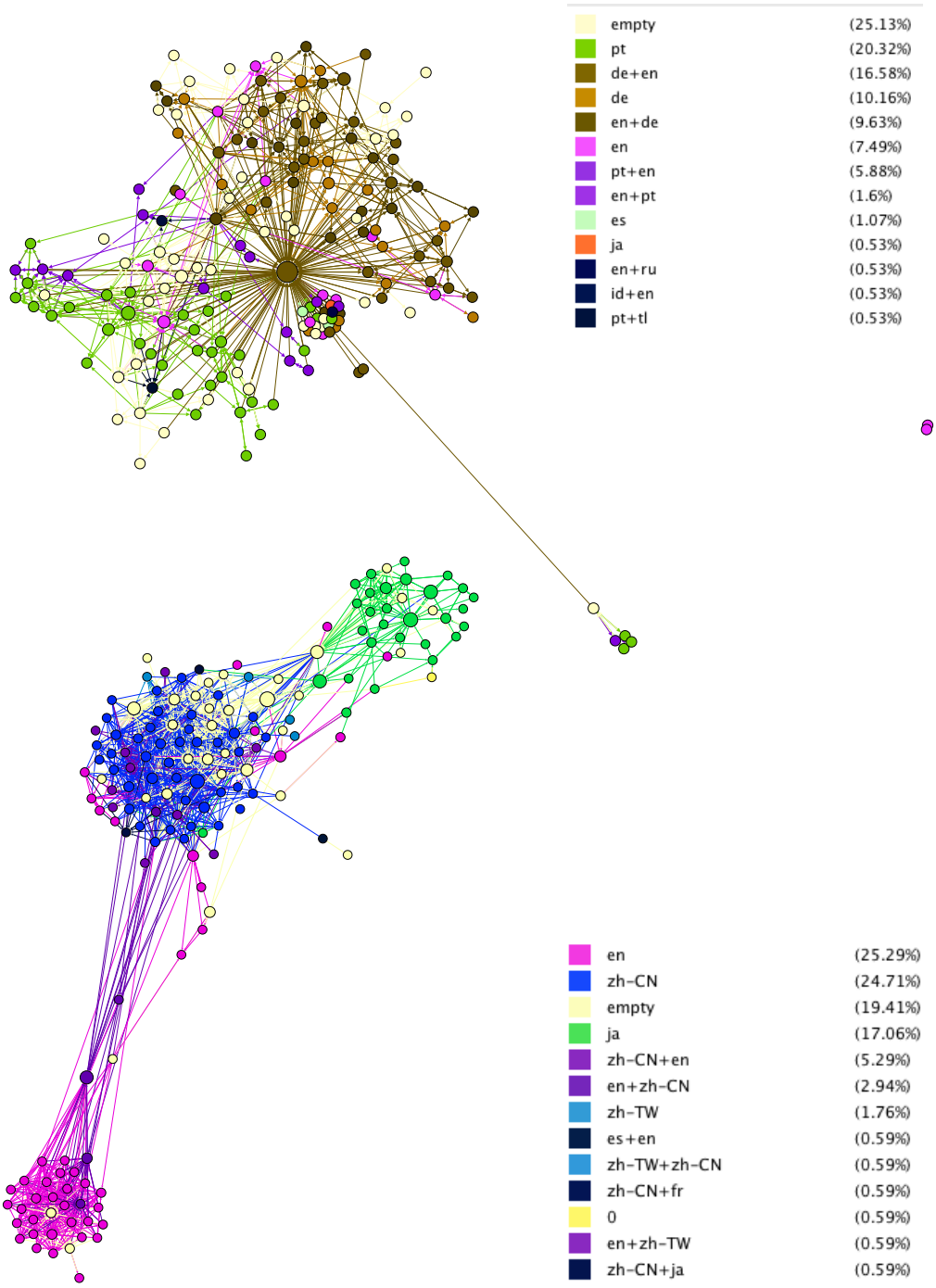


Figure A.3: Trilingual networks (3).

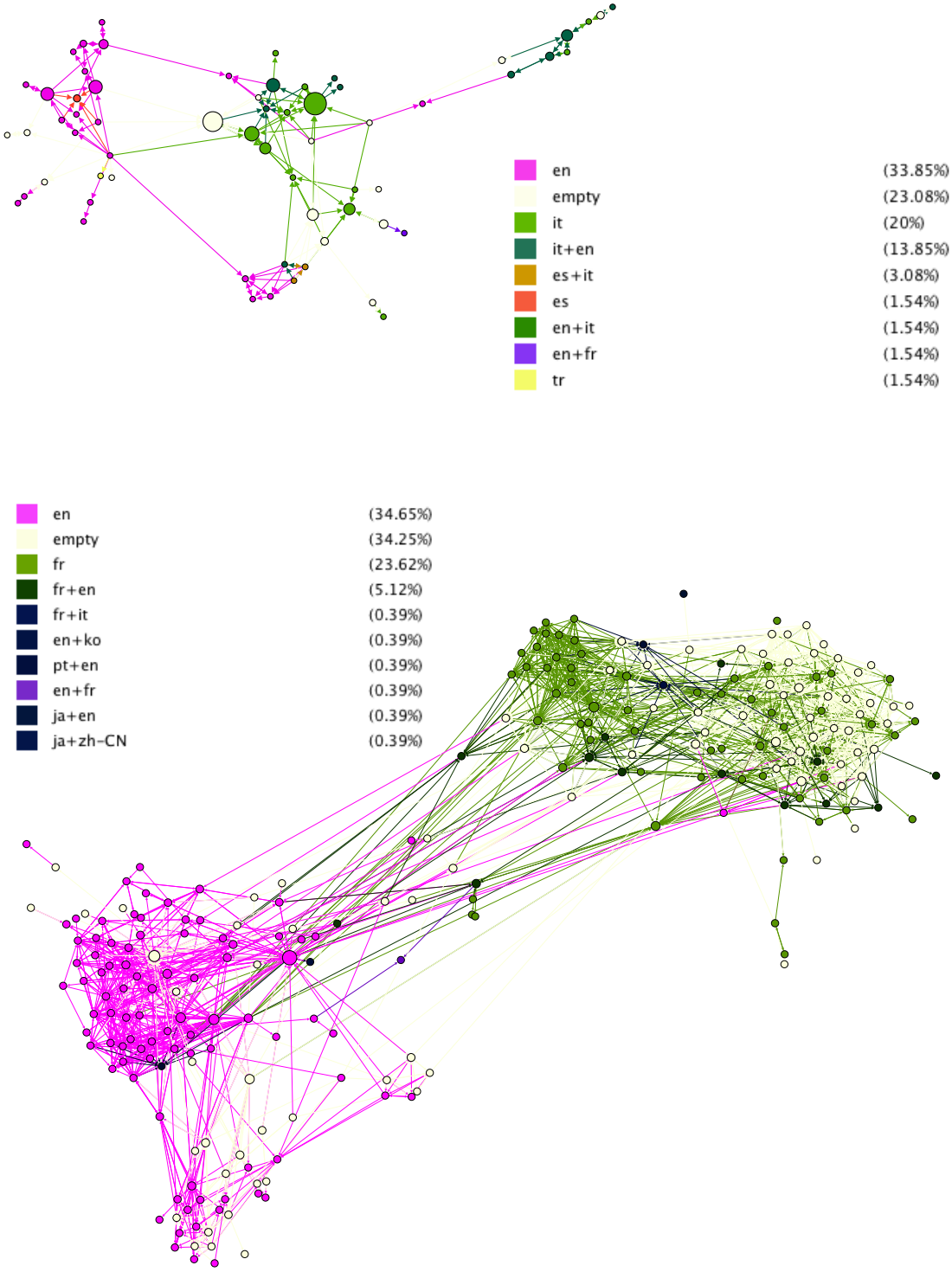


Figure A.4: Bilingual networks: gatekeeper type (1).

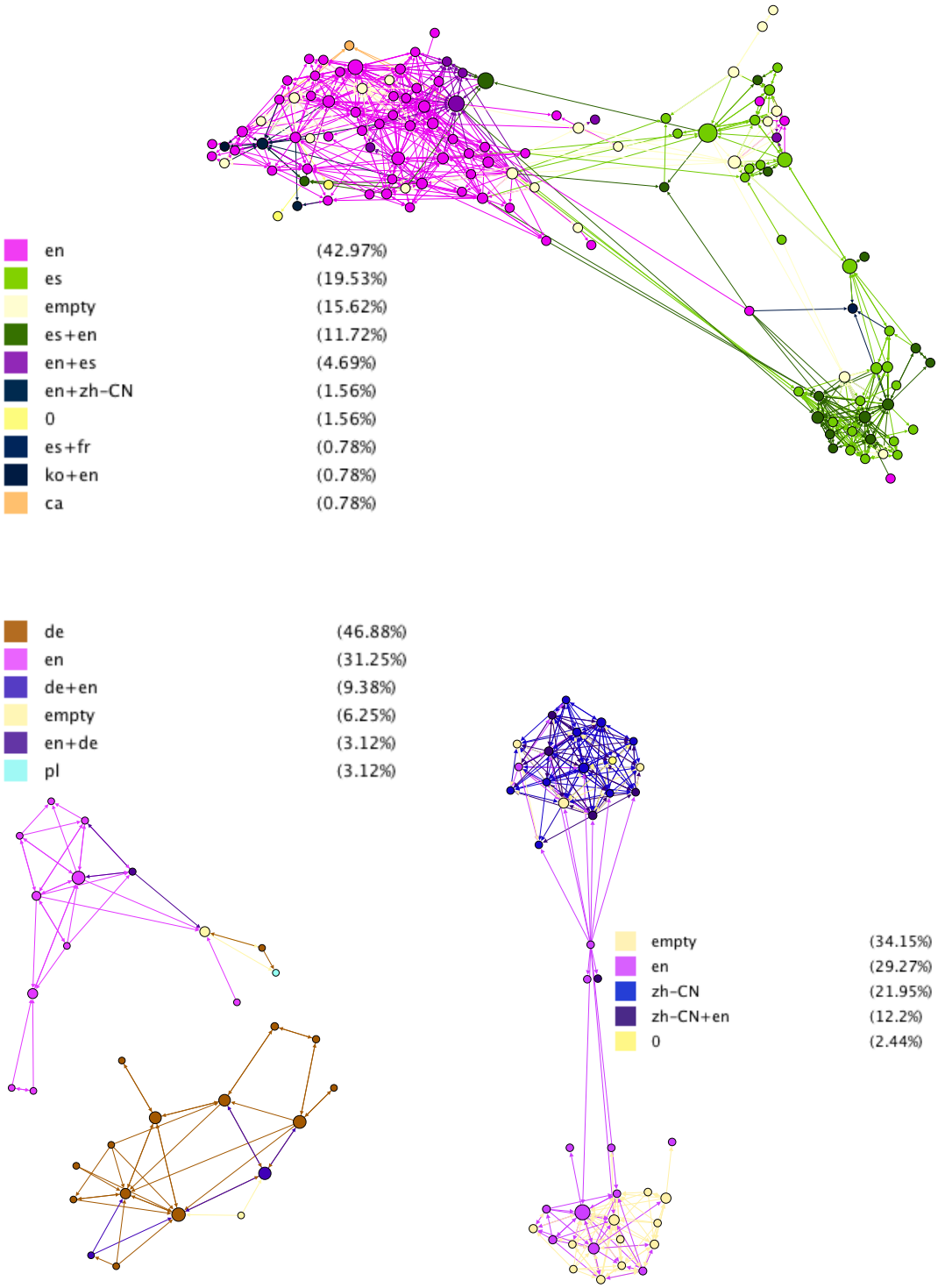


Figure A.5: Bilingual networks: gatekeeper type (2).

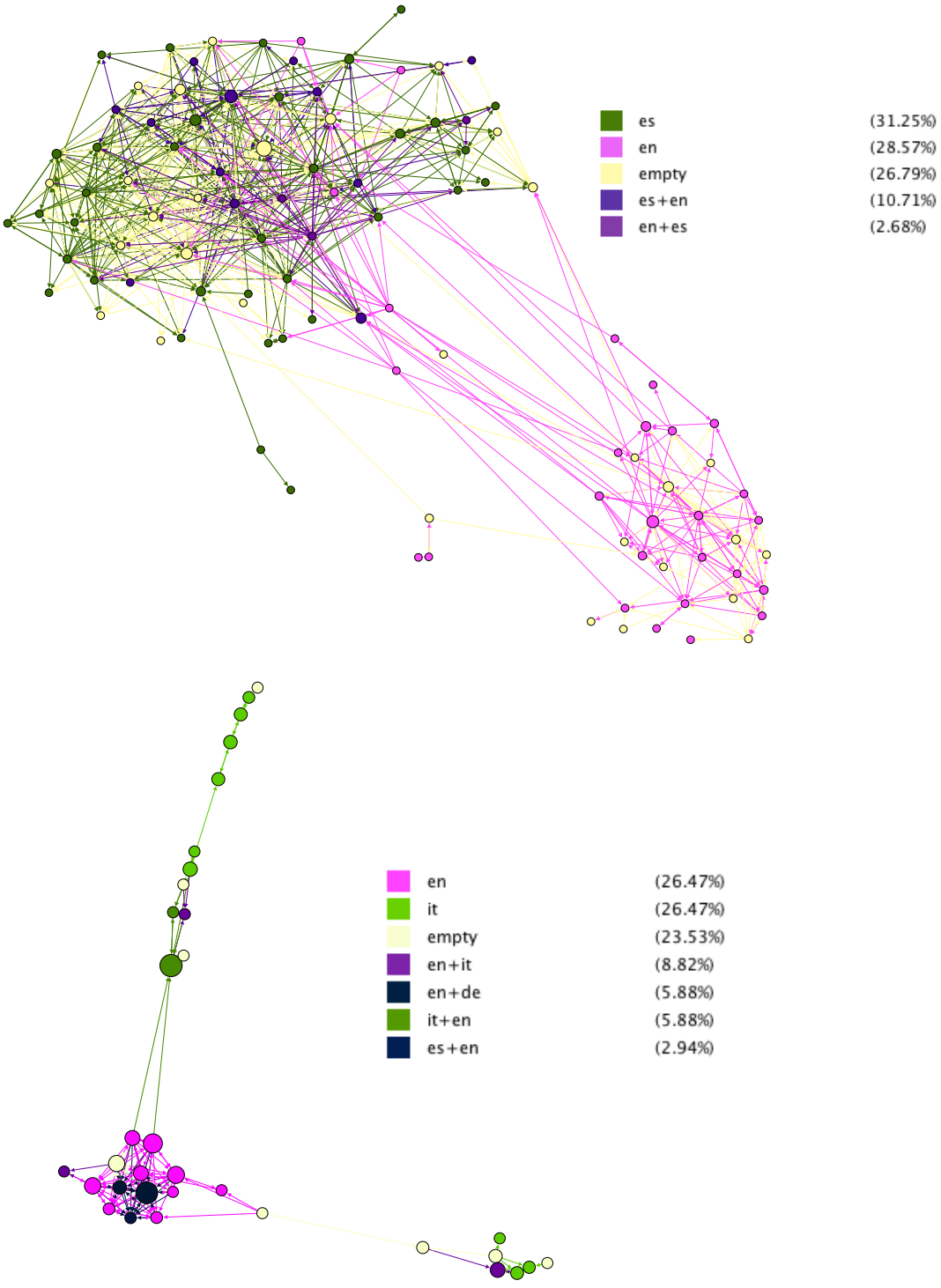


Figure A.6: Bilingual networks: gatekeeper type (3).

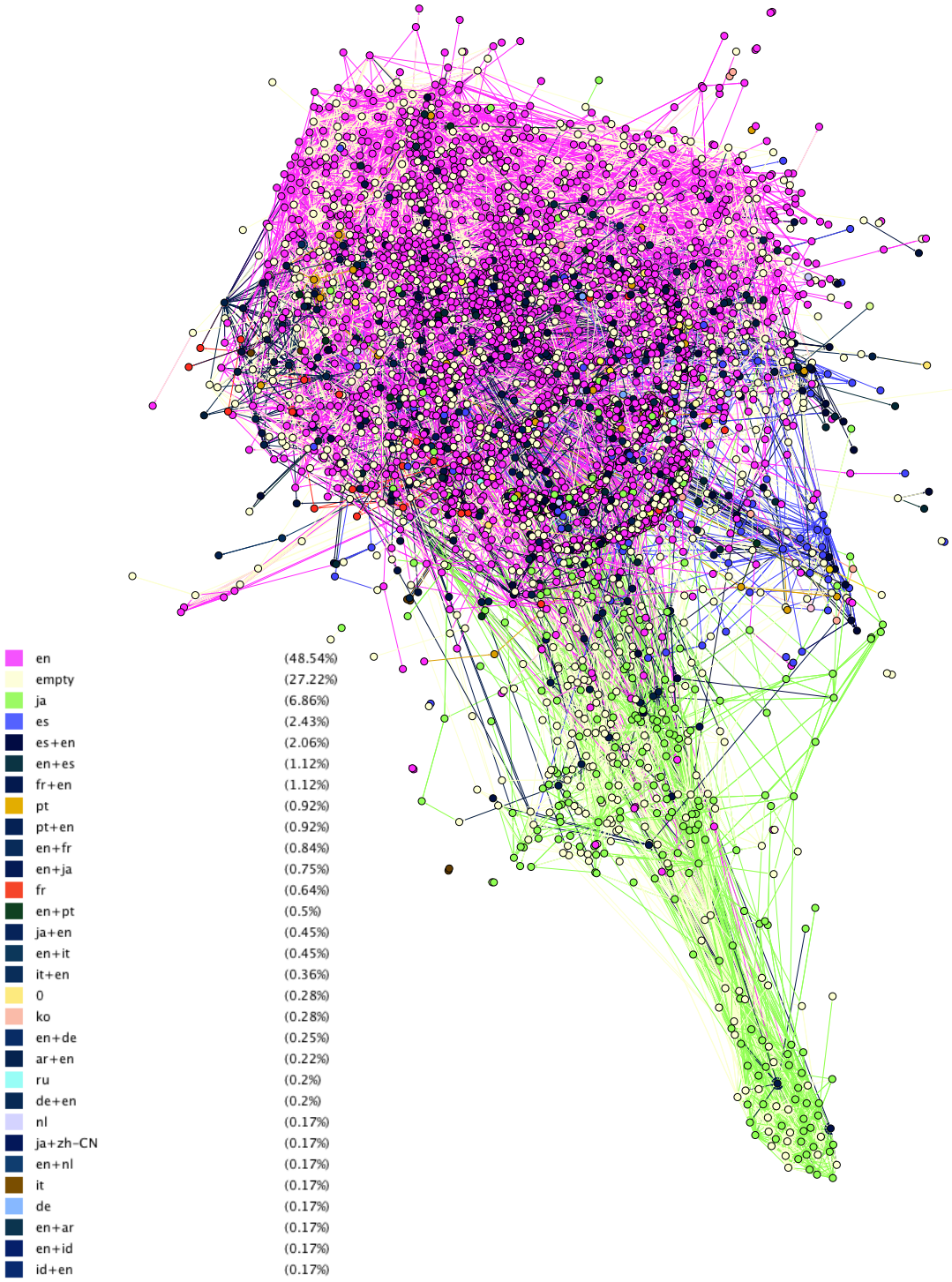


Figure A.7: Bilingual networks: gatekeeper type (4).

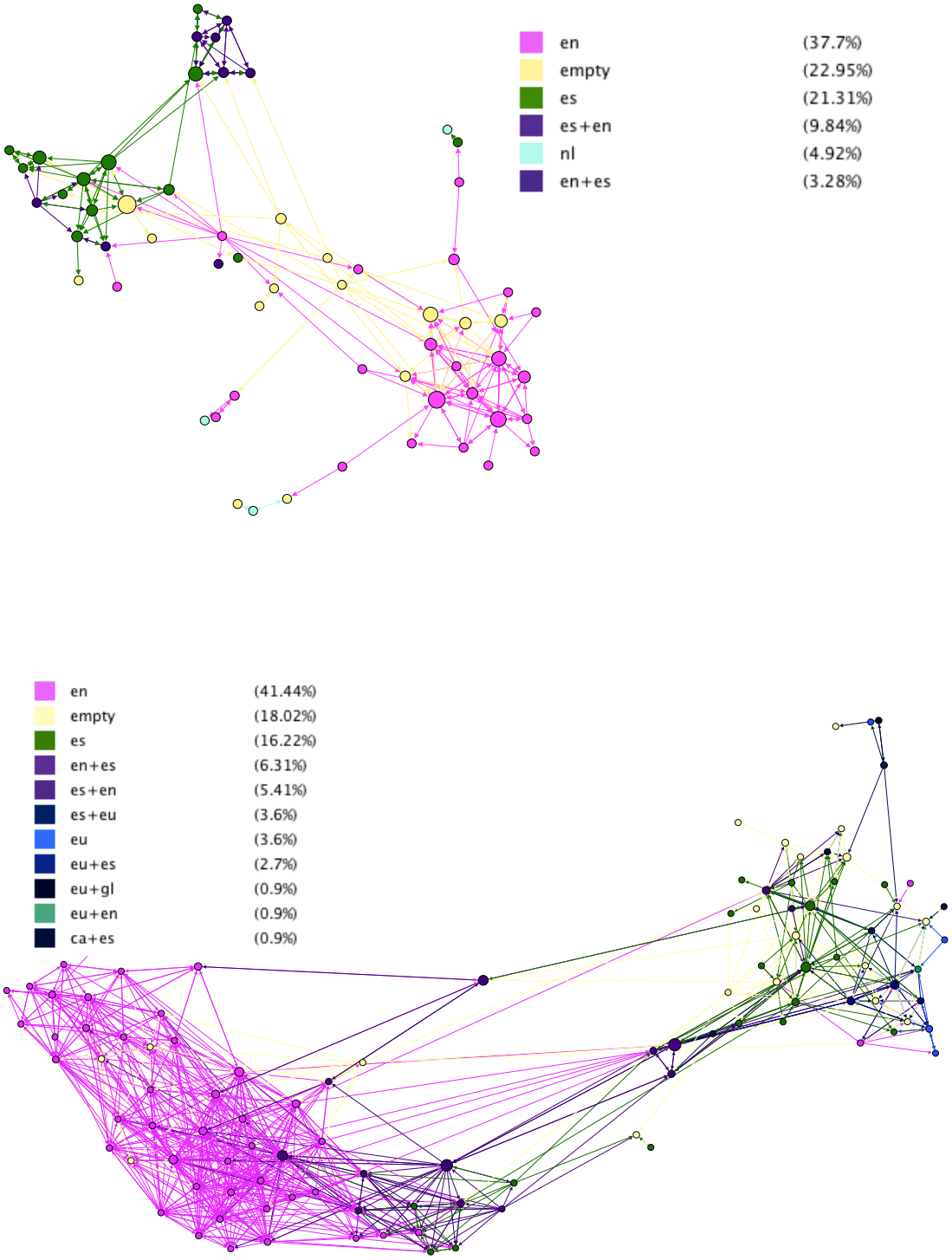


Figure A.8: Bilingual networks: gatekeeper type (5).

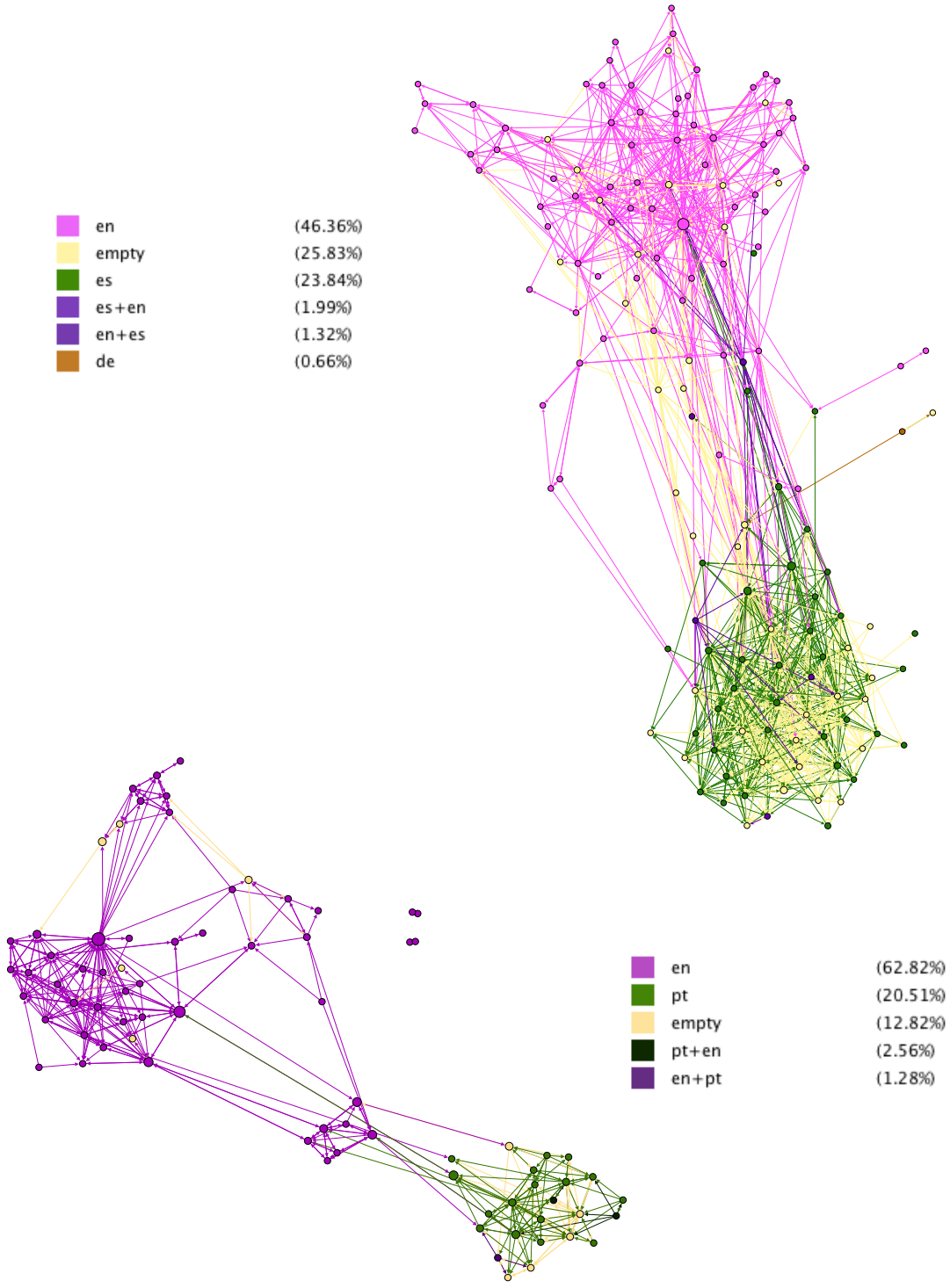


Figure A.9: Bilingual networks: gatekeeper type (6).

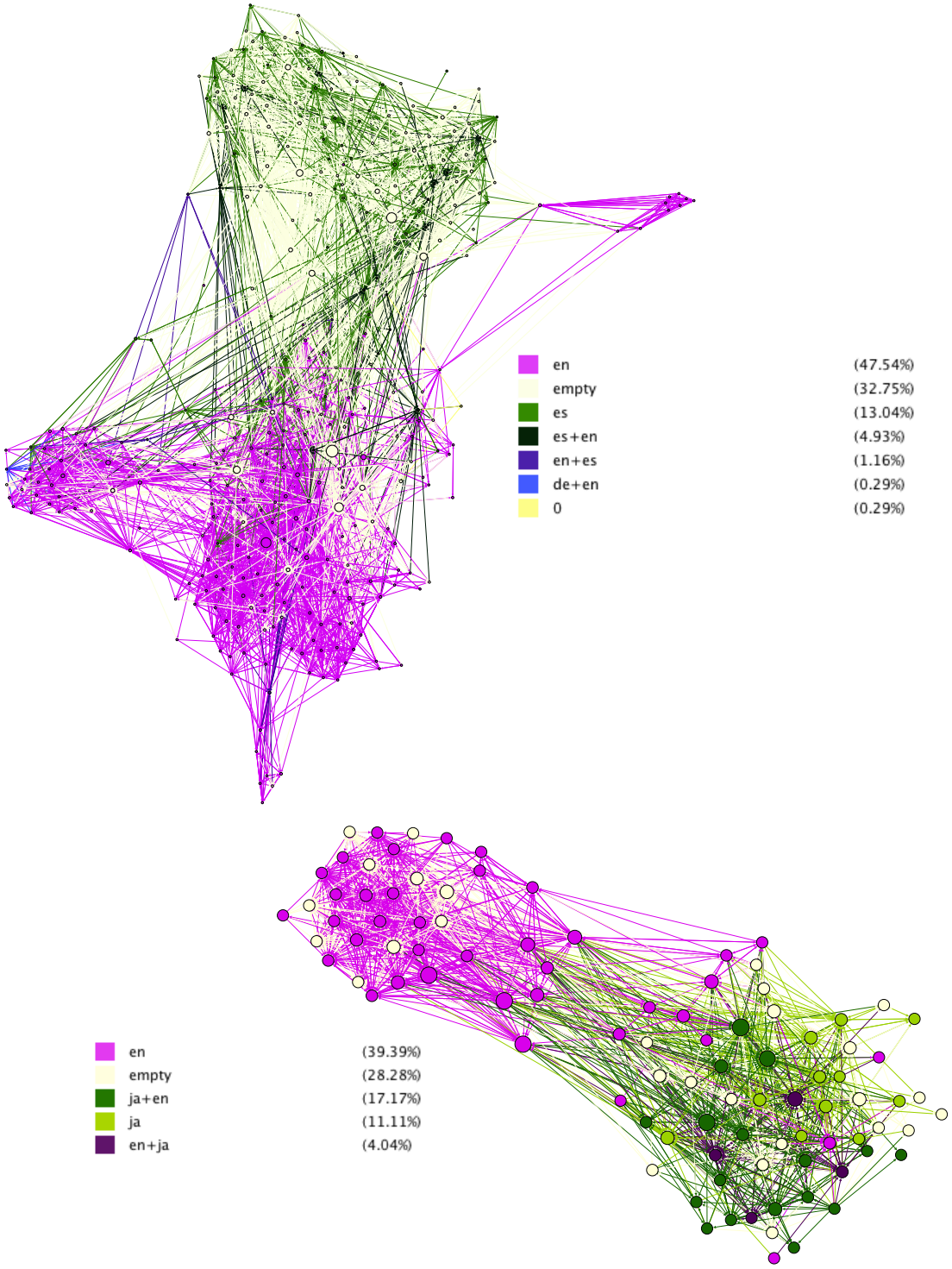


Figure A.10: Bilingual networks: language bridge type (1).

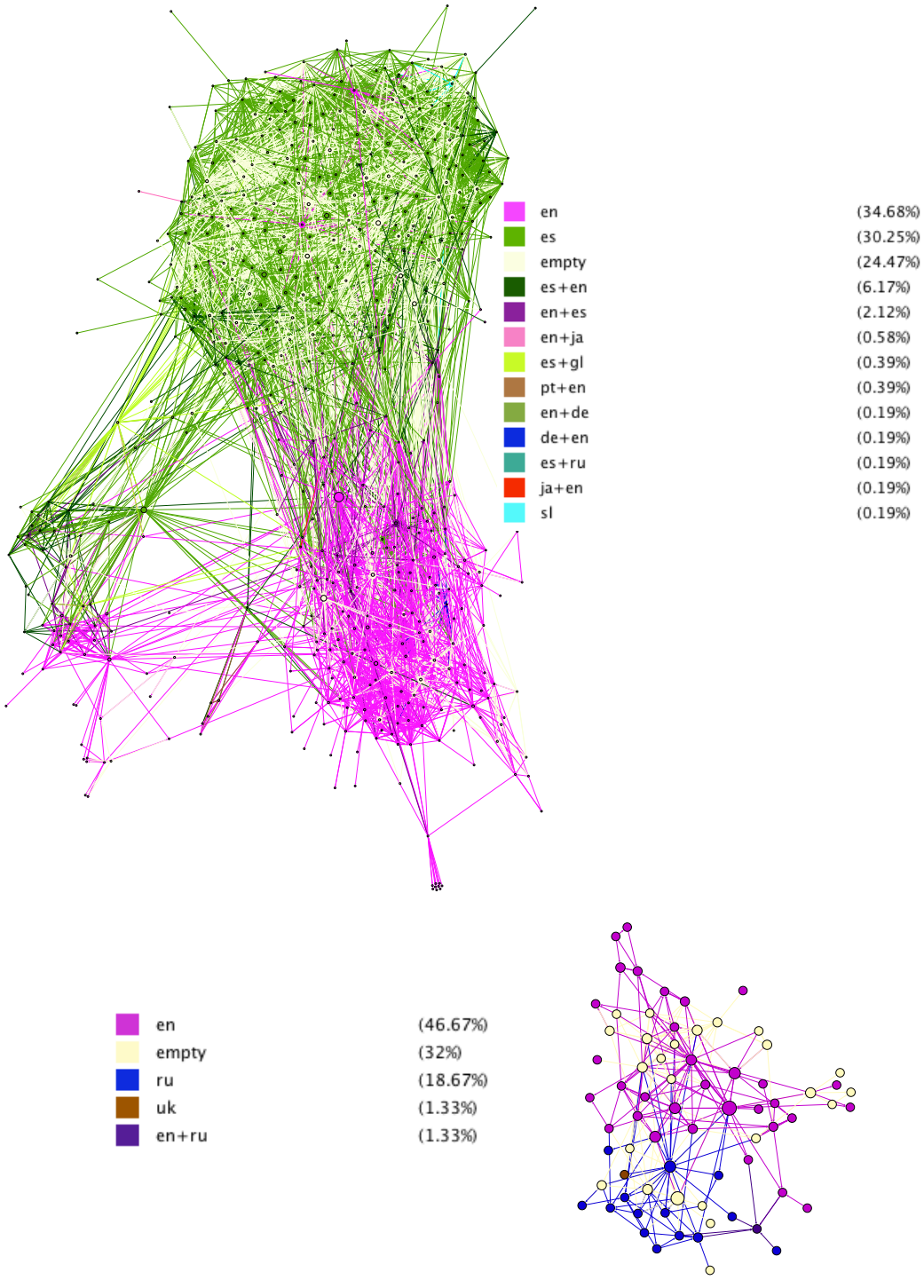


Figure A.11: Bilingual networks: language bridge type (2).

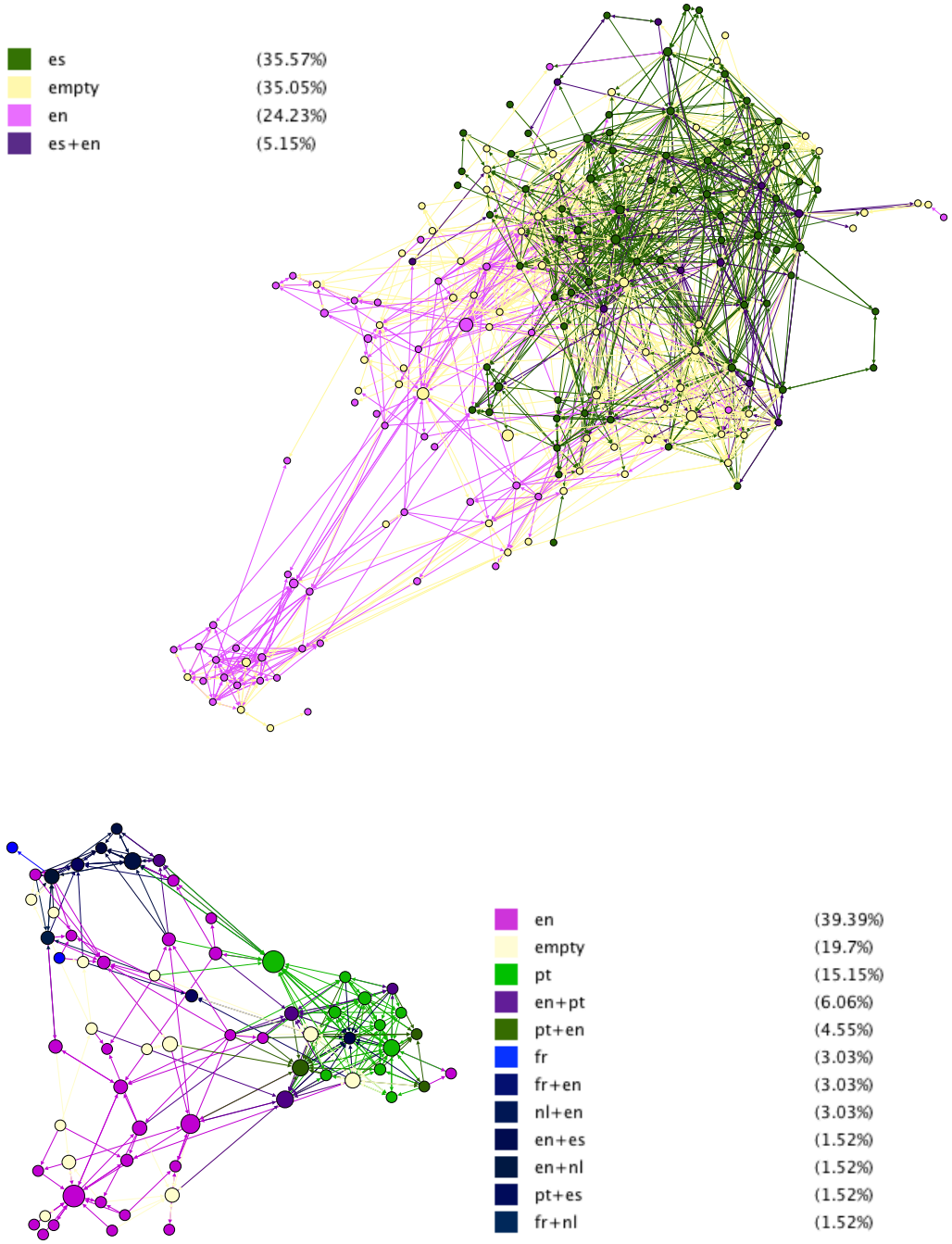


Figure A.12: Bilingual networks: language bridge type (3).

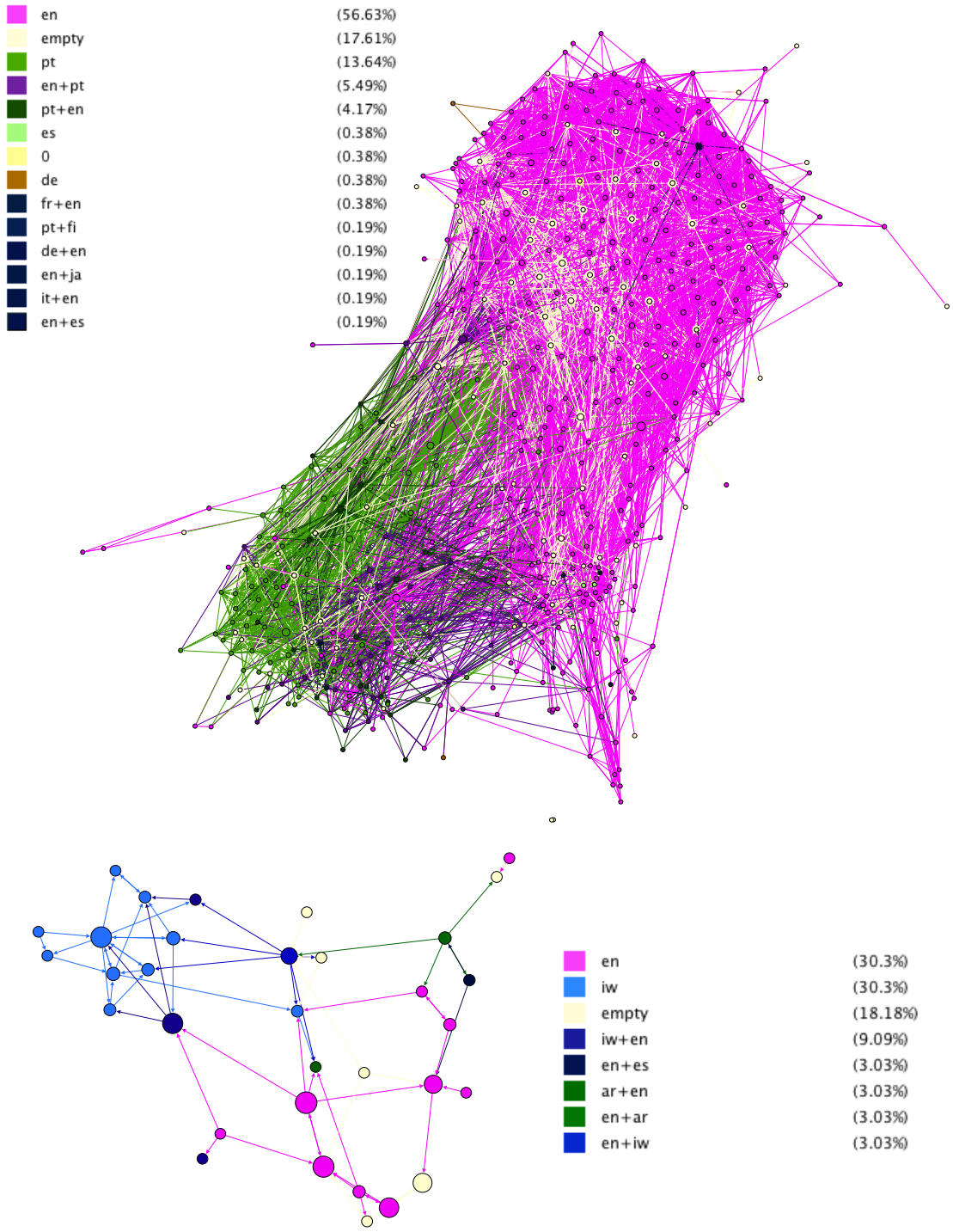


Figure A.13: Bilingual networks: language bridge type (4).

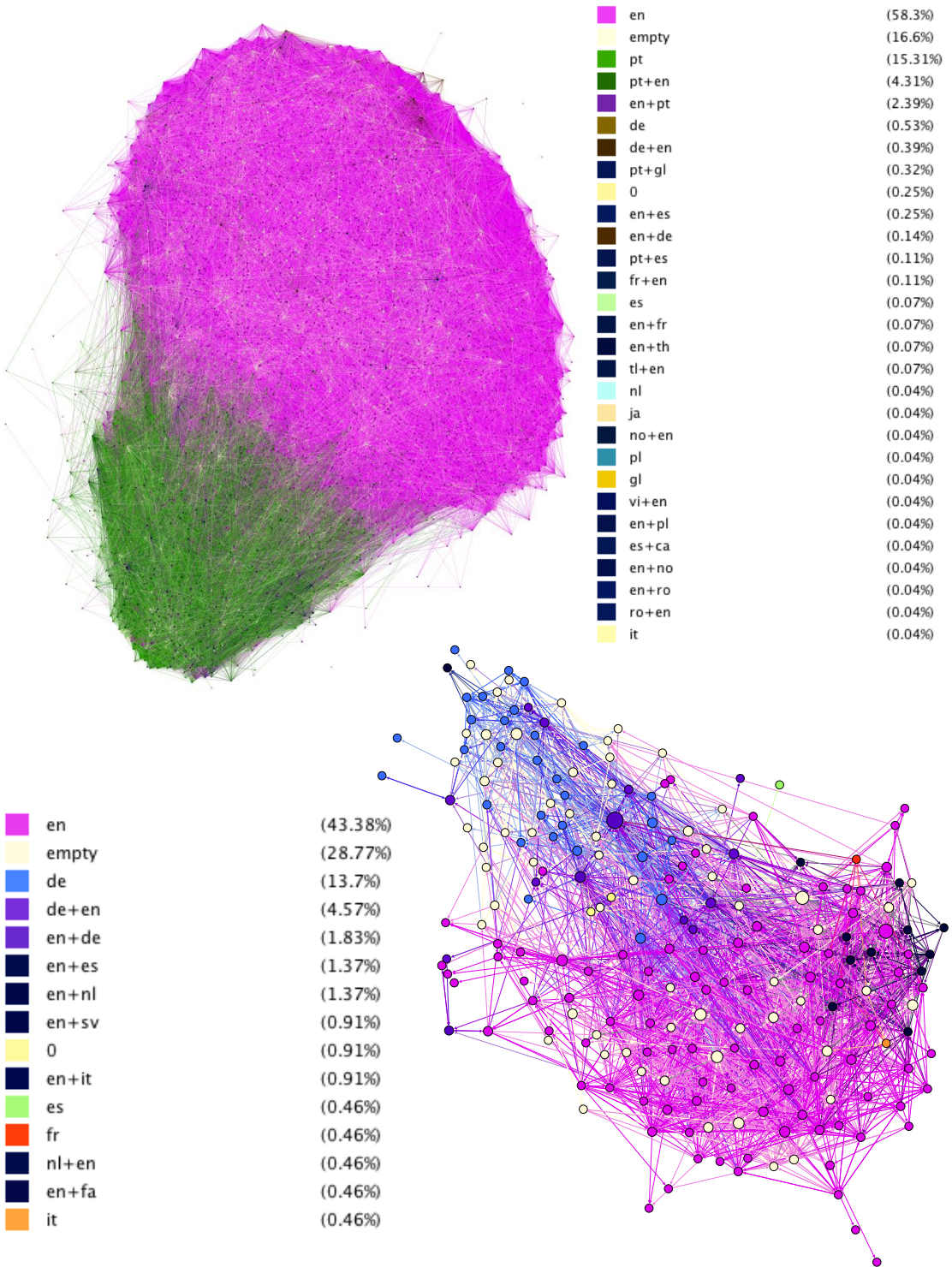


Figure A.14: Bilingual networks: language bridge type (5).

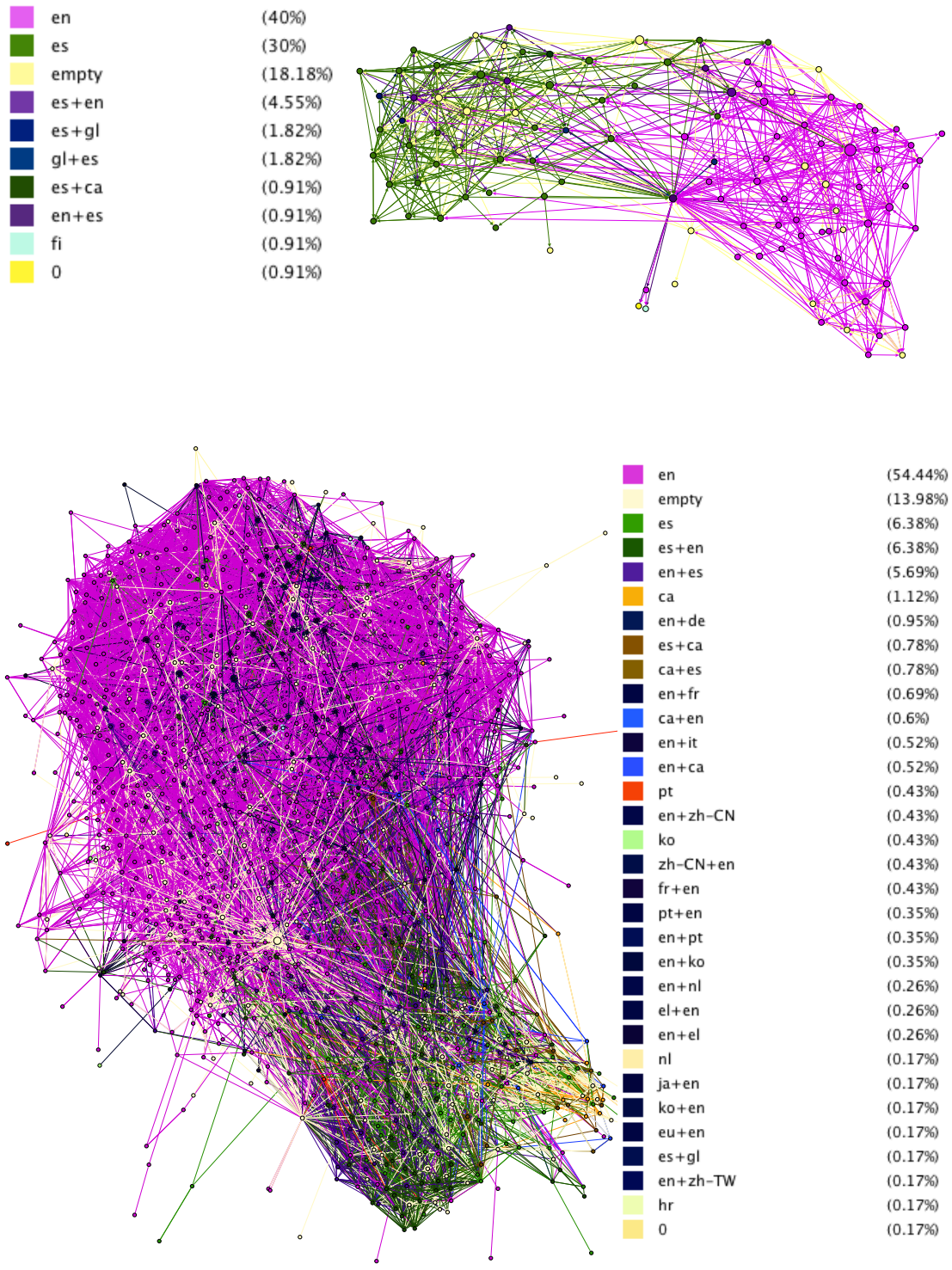


Figure A.15: Bilingual networks: language bridge type (6).

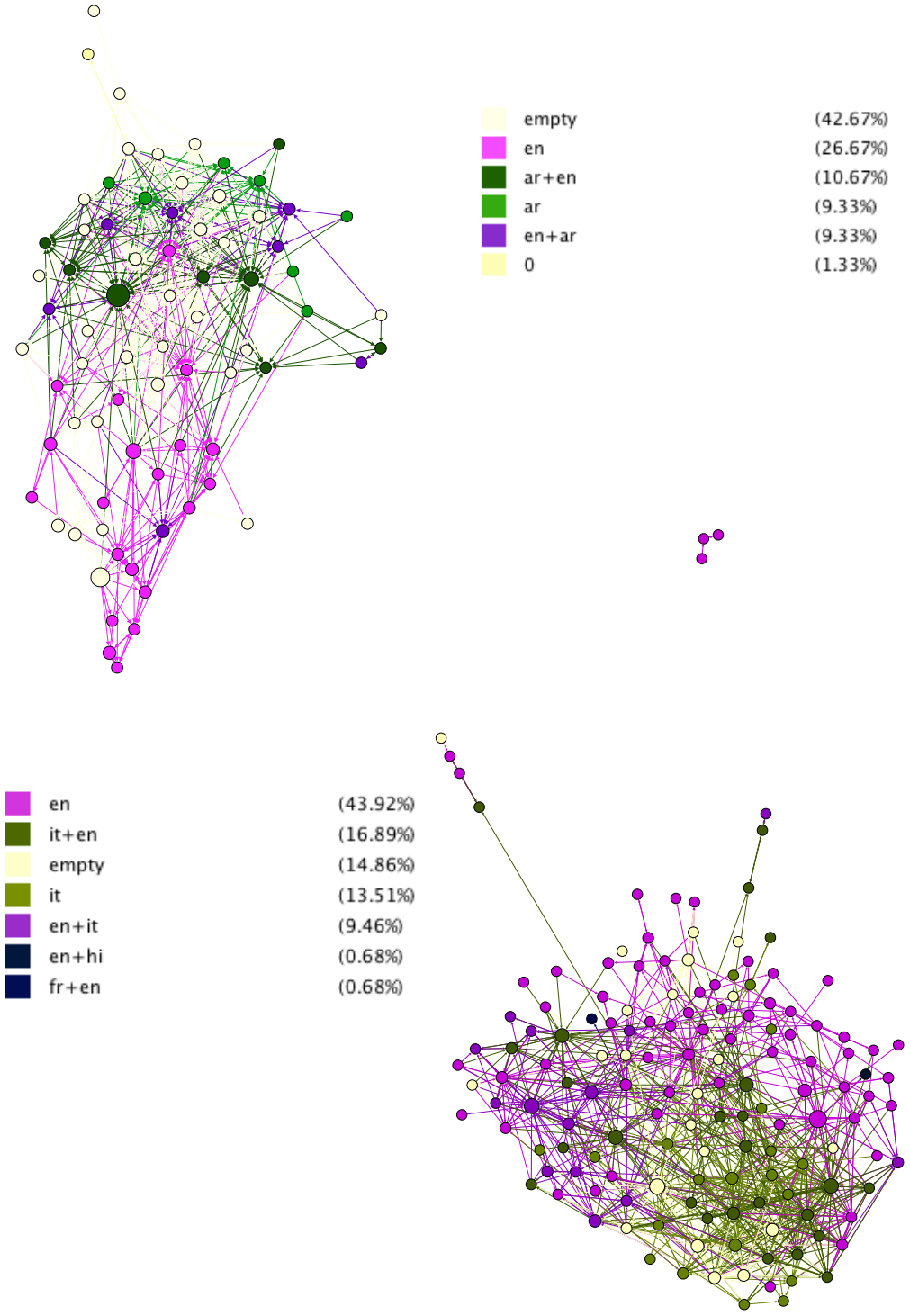


Figure A.16: Bilingual networks: union type (1).

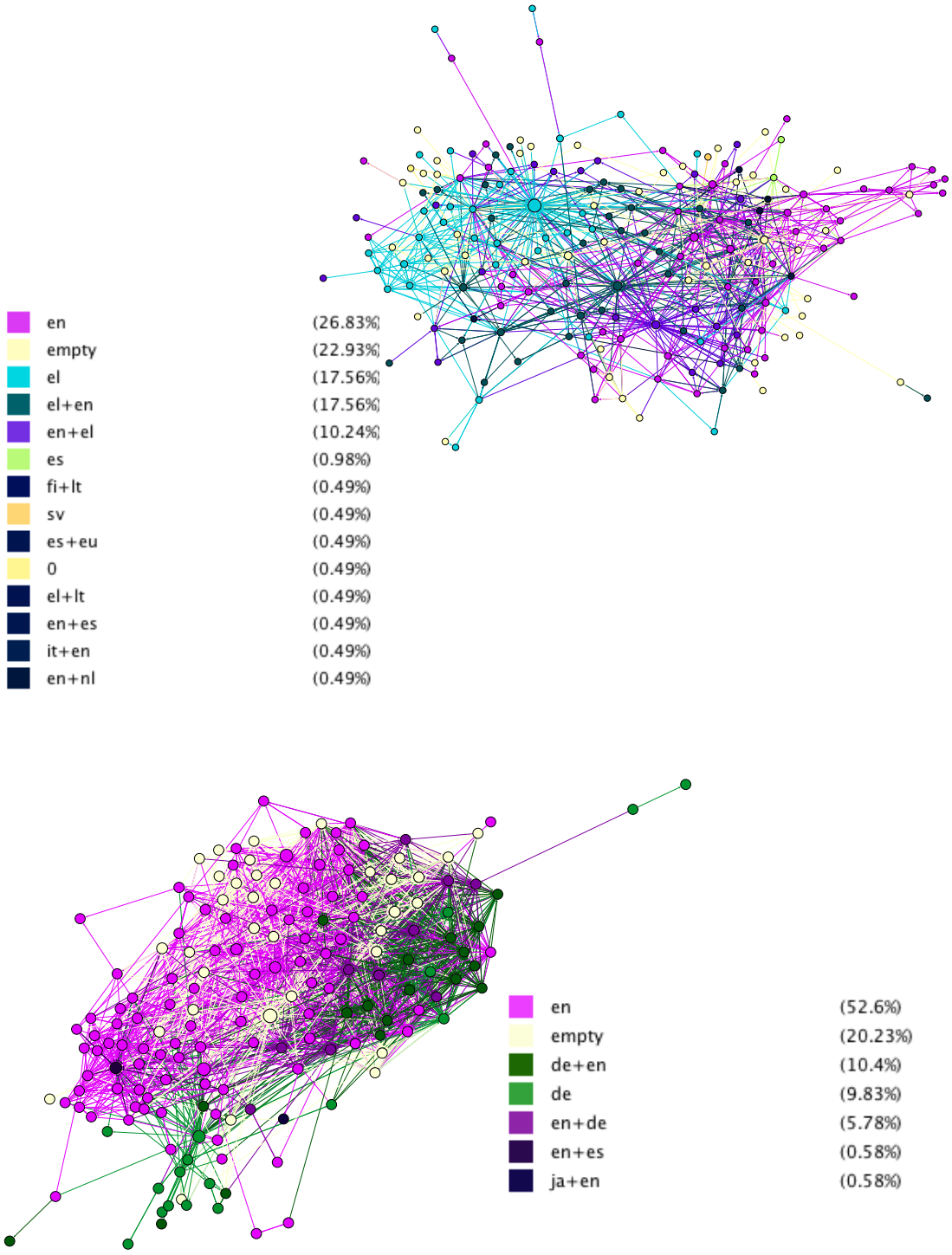


Figure A.17: Bilingual networks: union type (2).

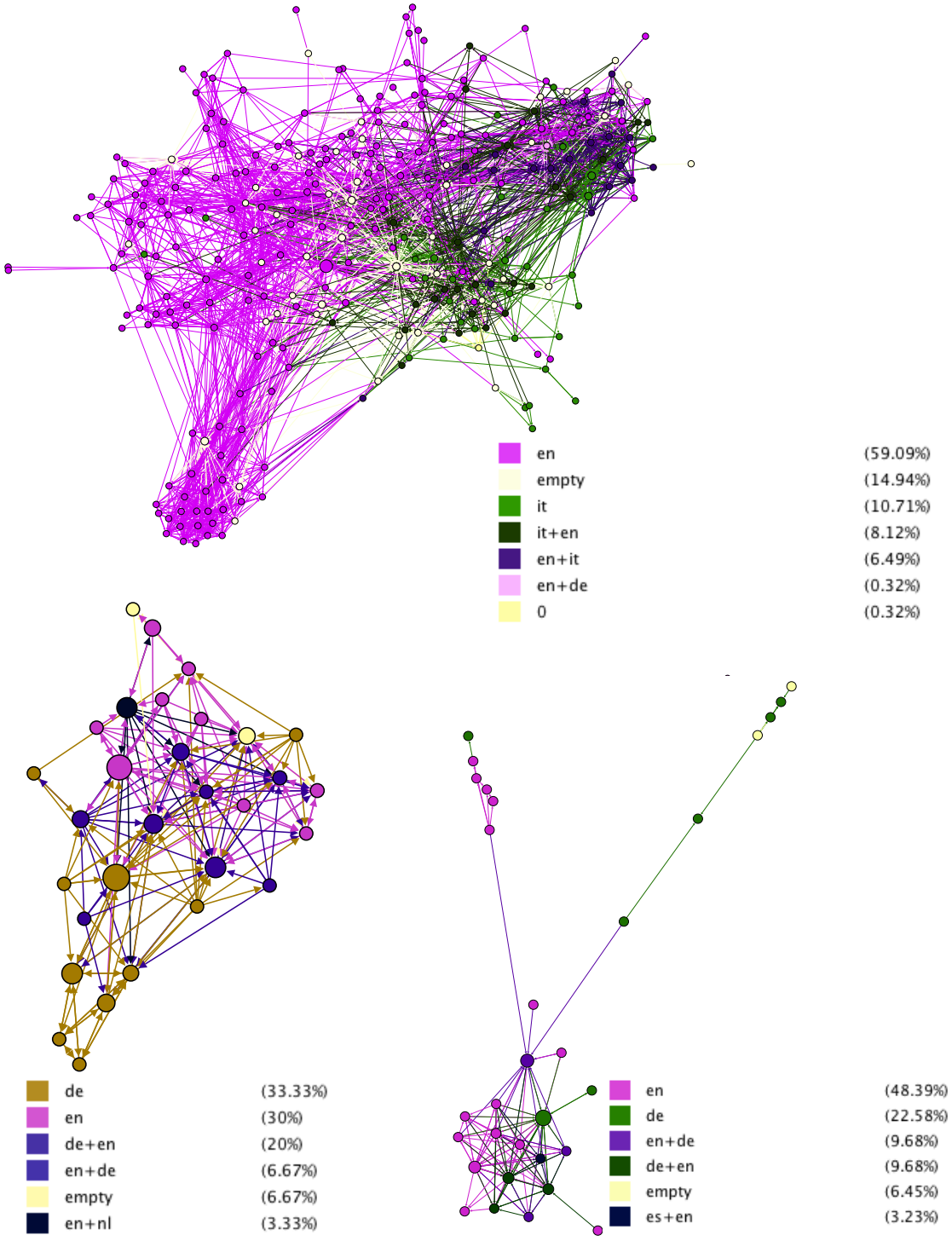


Figure A.18: Bilingual networks: union type (3).

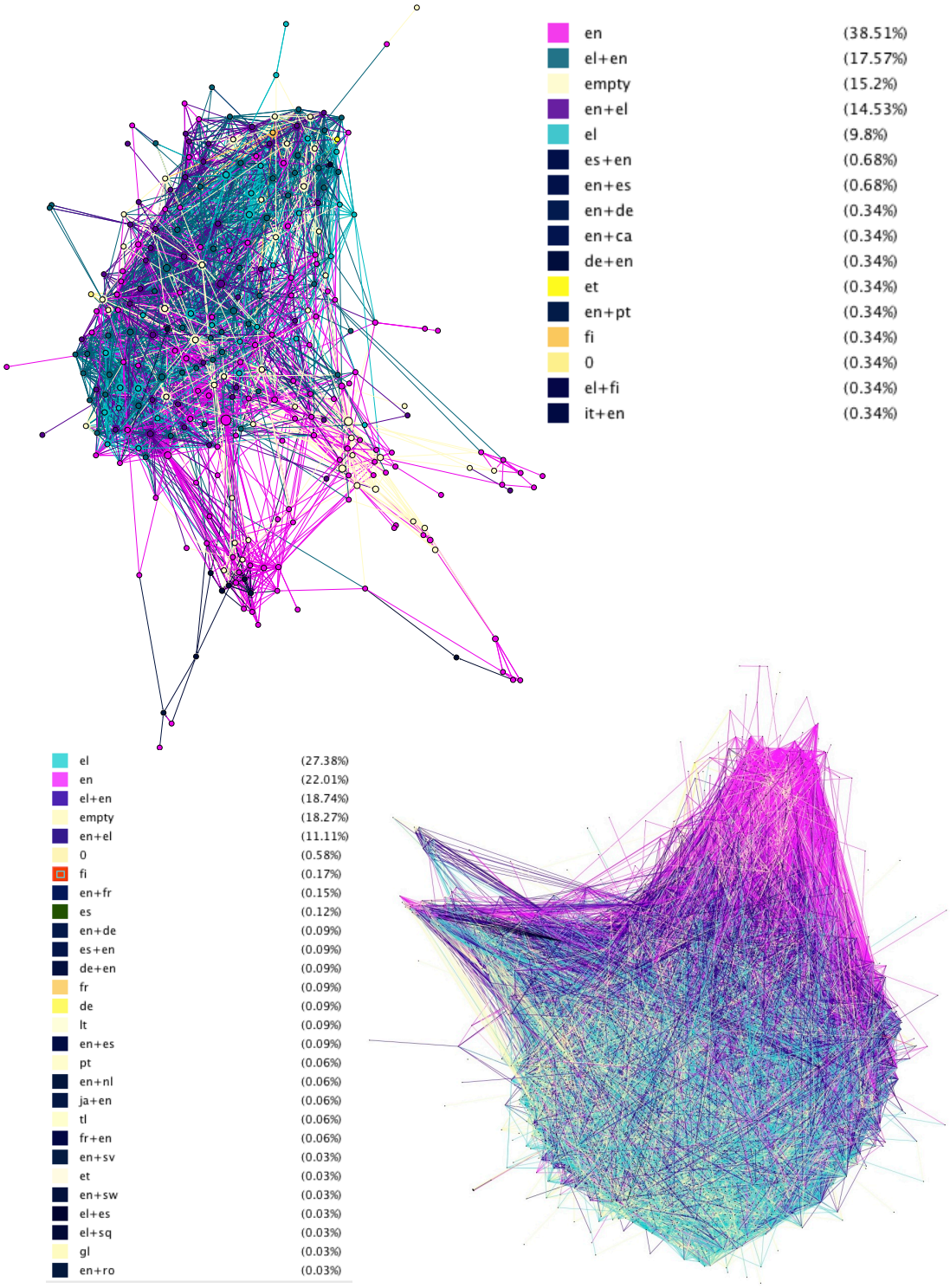


Figure A.19: Bilingual networks: union type (4).

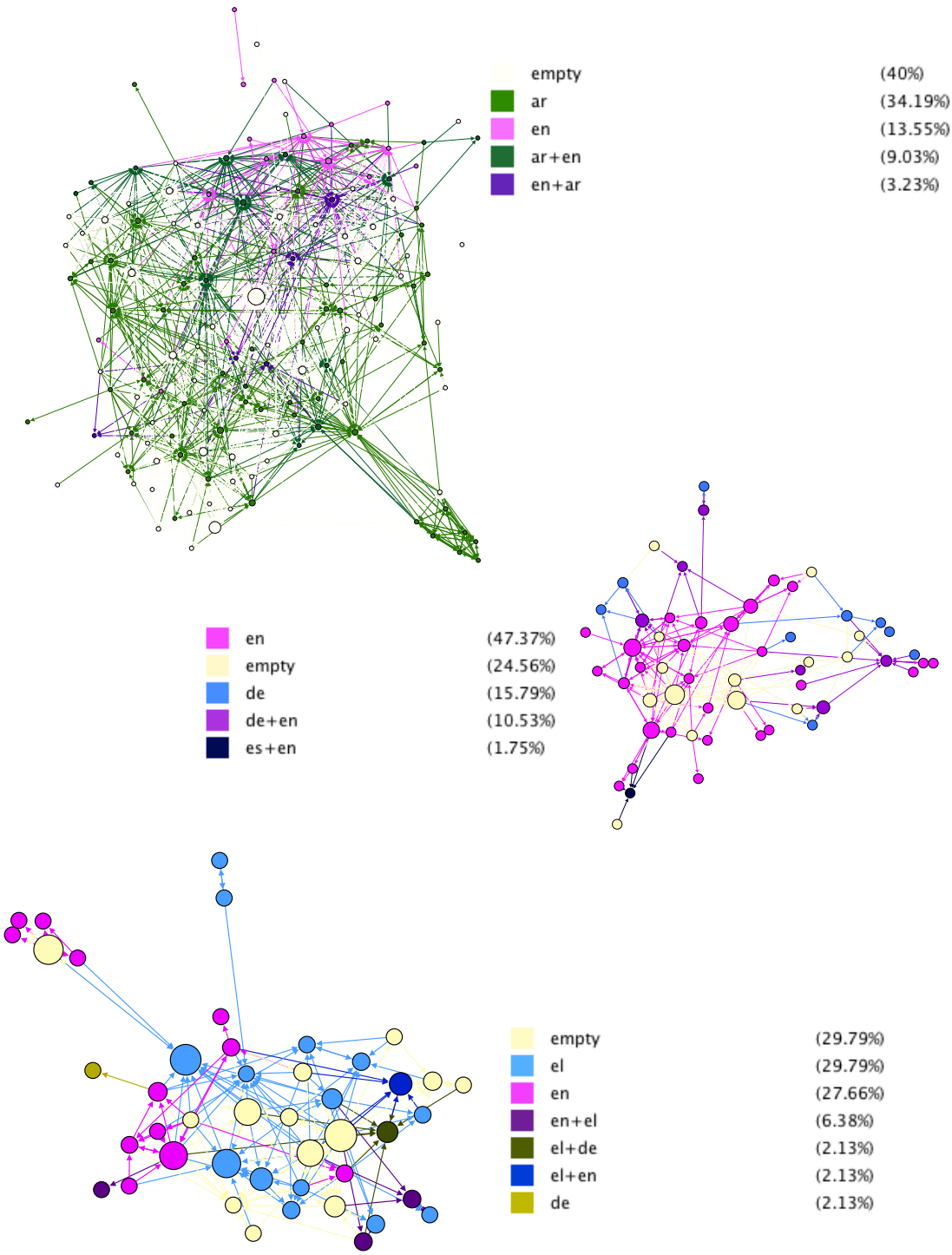


Figure A.20: Bilingual networks: integration type (1).

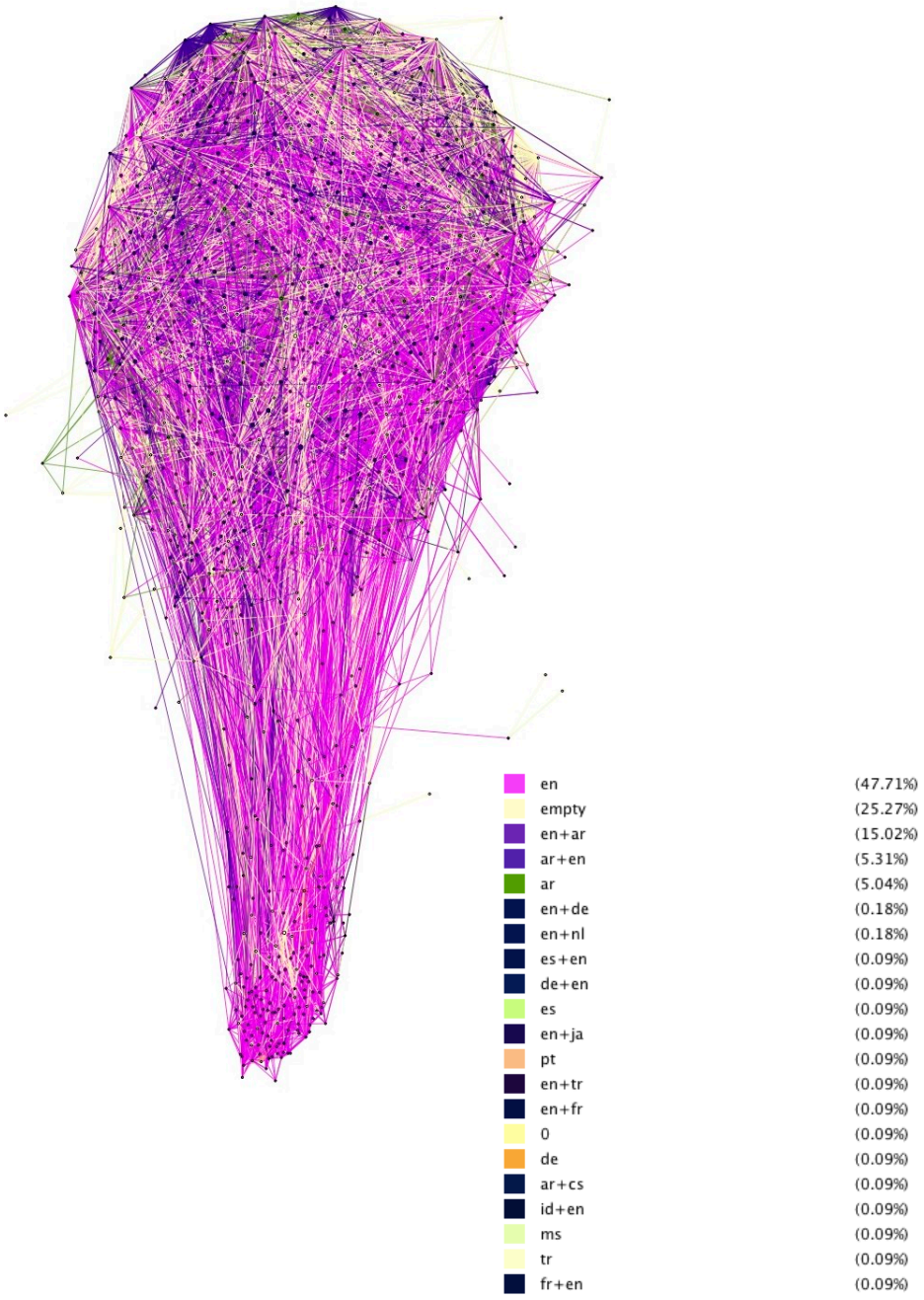


Figure A.21: Bilingual networks: integration type (2).

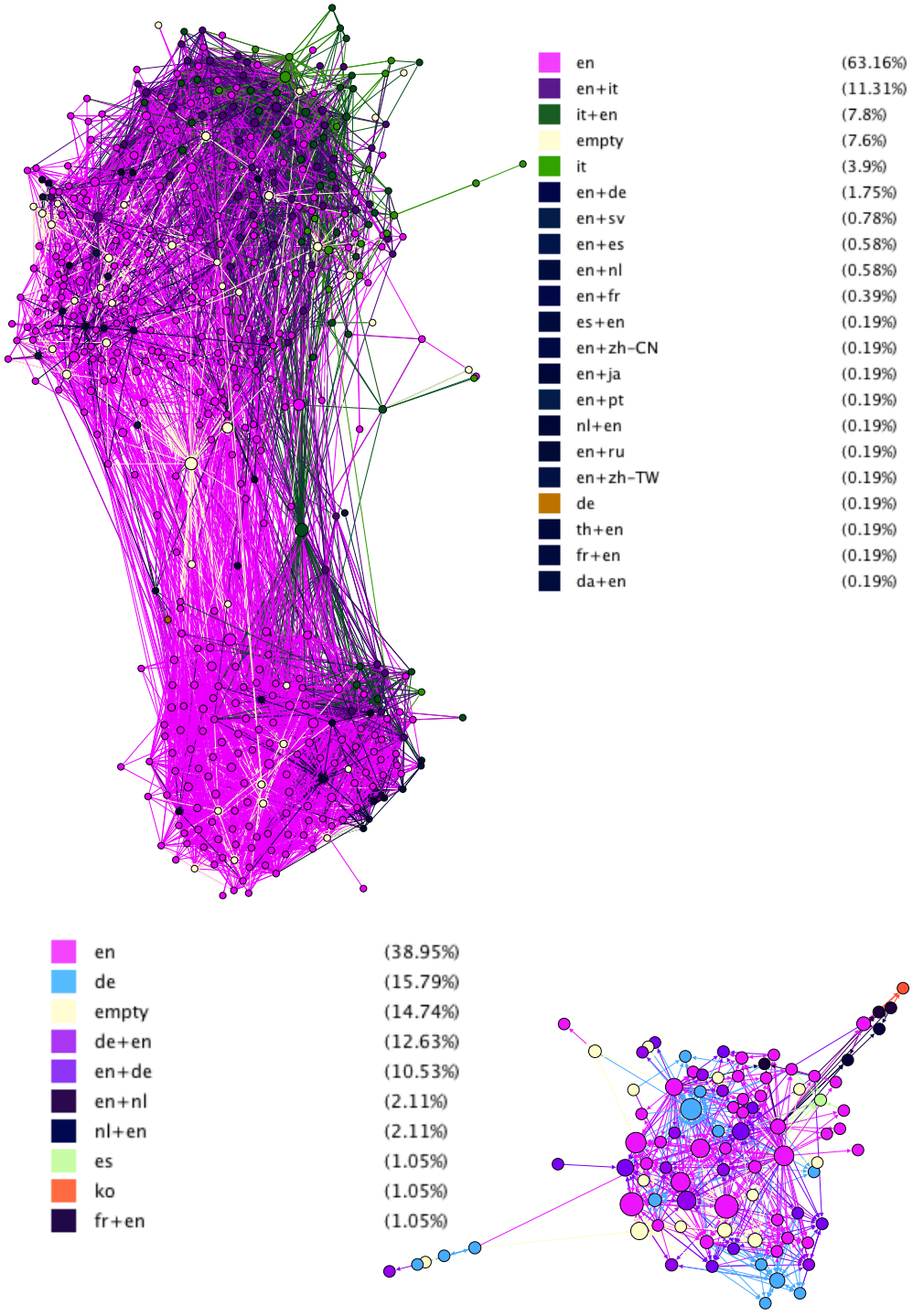


Figure A.22: Bilingual networks: integration type (3).

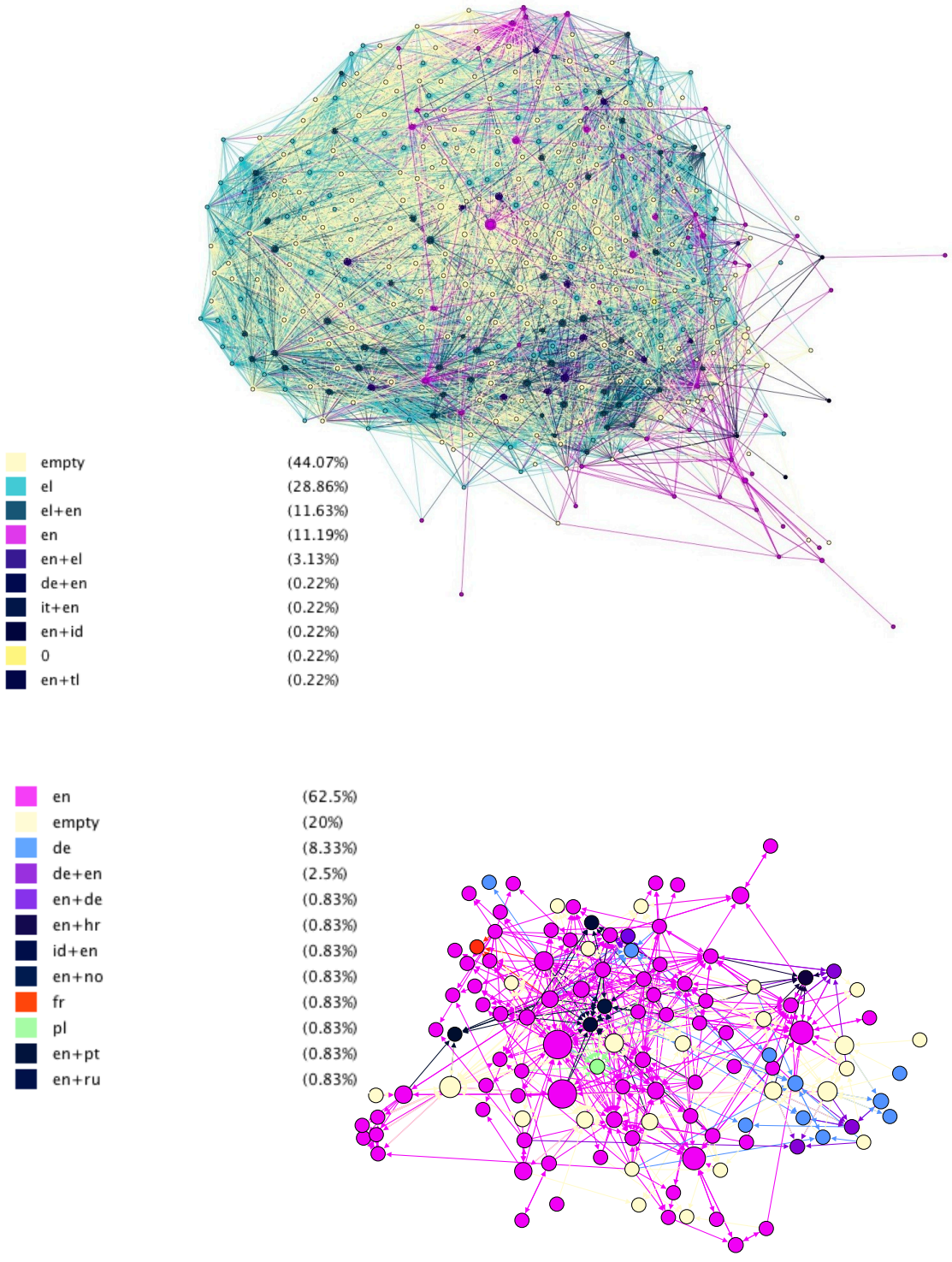


Figure A.23: Bilingual networks: integration type (4).

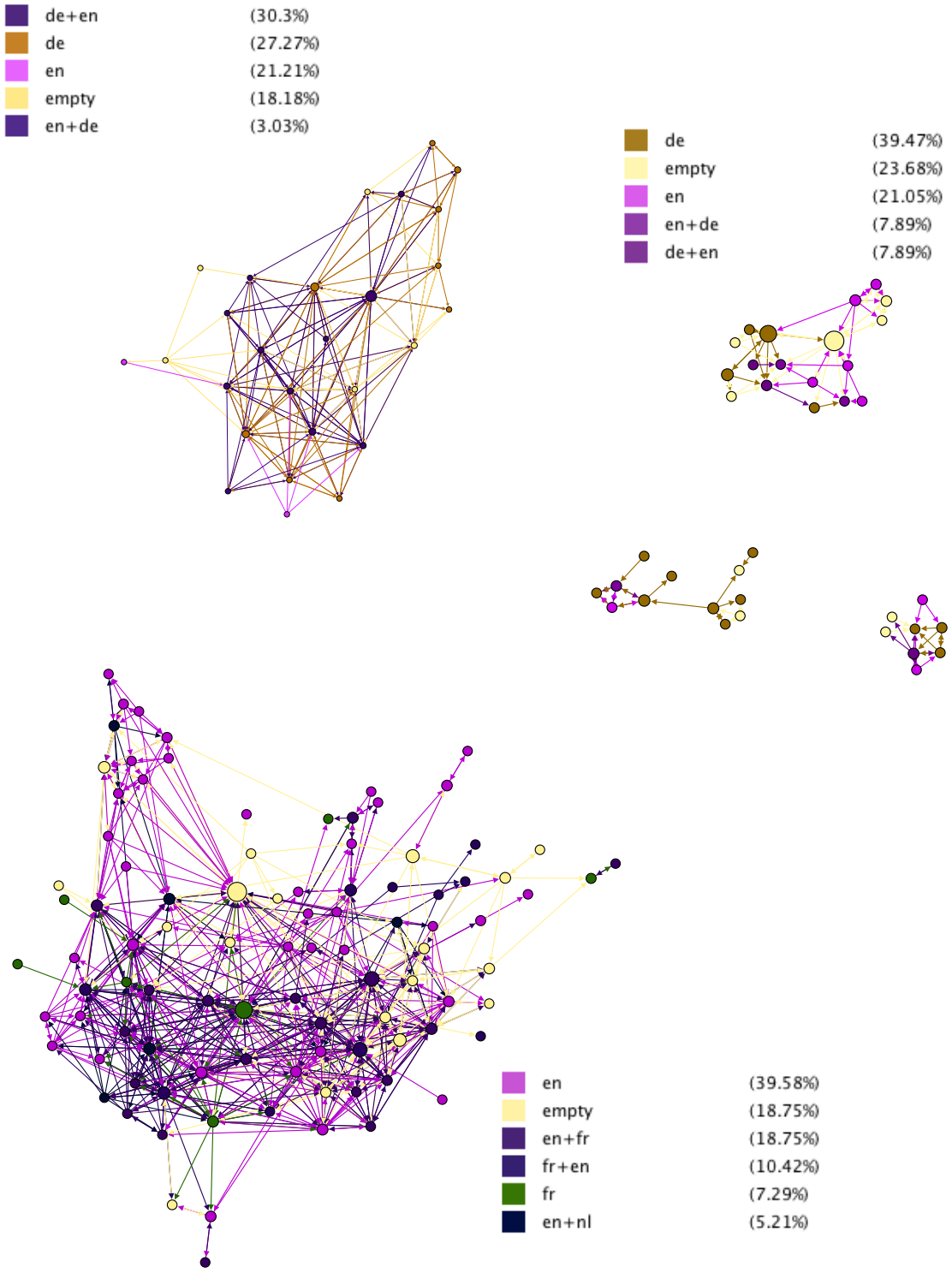


Figure A.24: Bilingual networks: integration type (5).

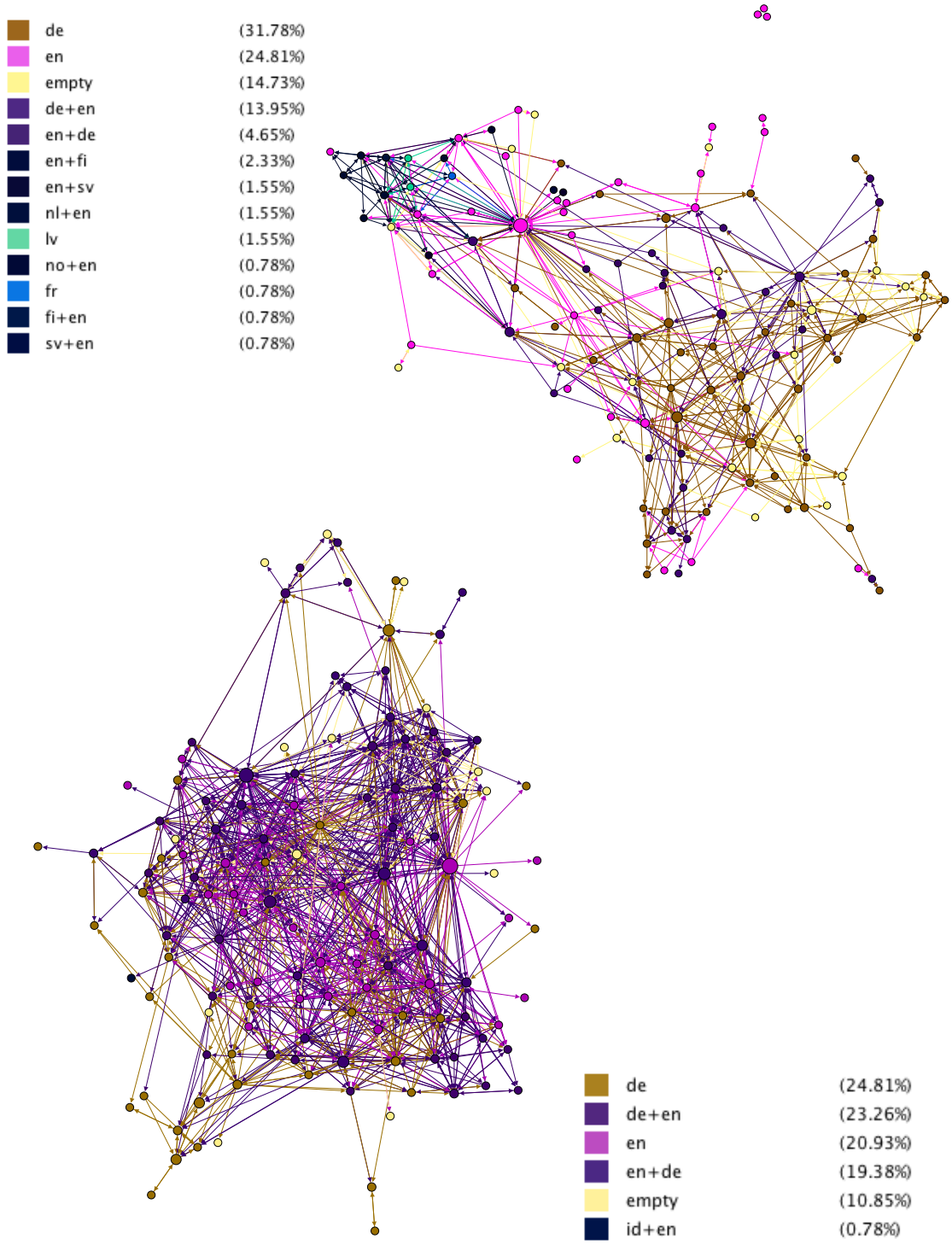


Figure A.25: Bilingual networks: integration type (6).

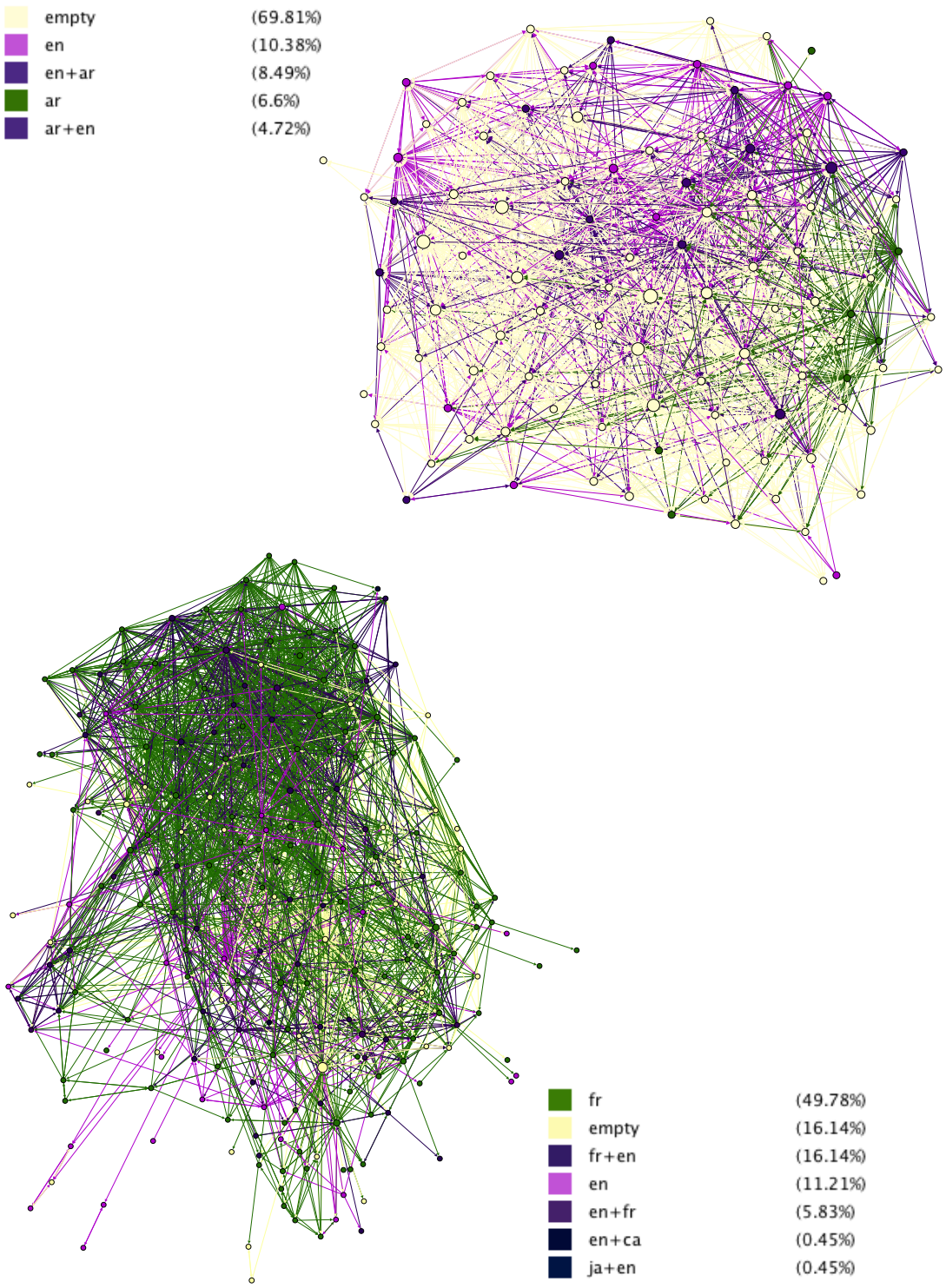


Figure A.26: Bilingual networks: integration type (7).

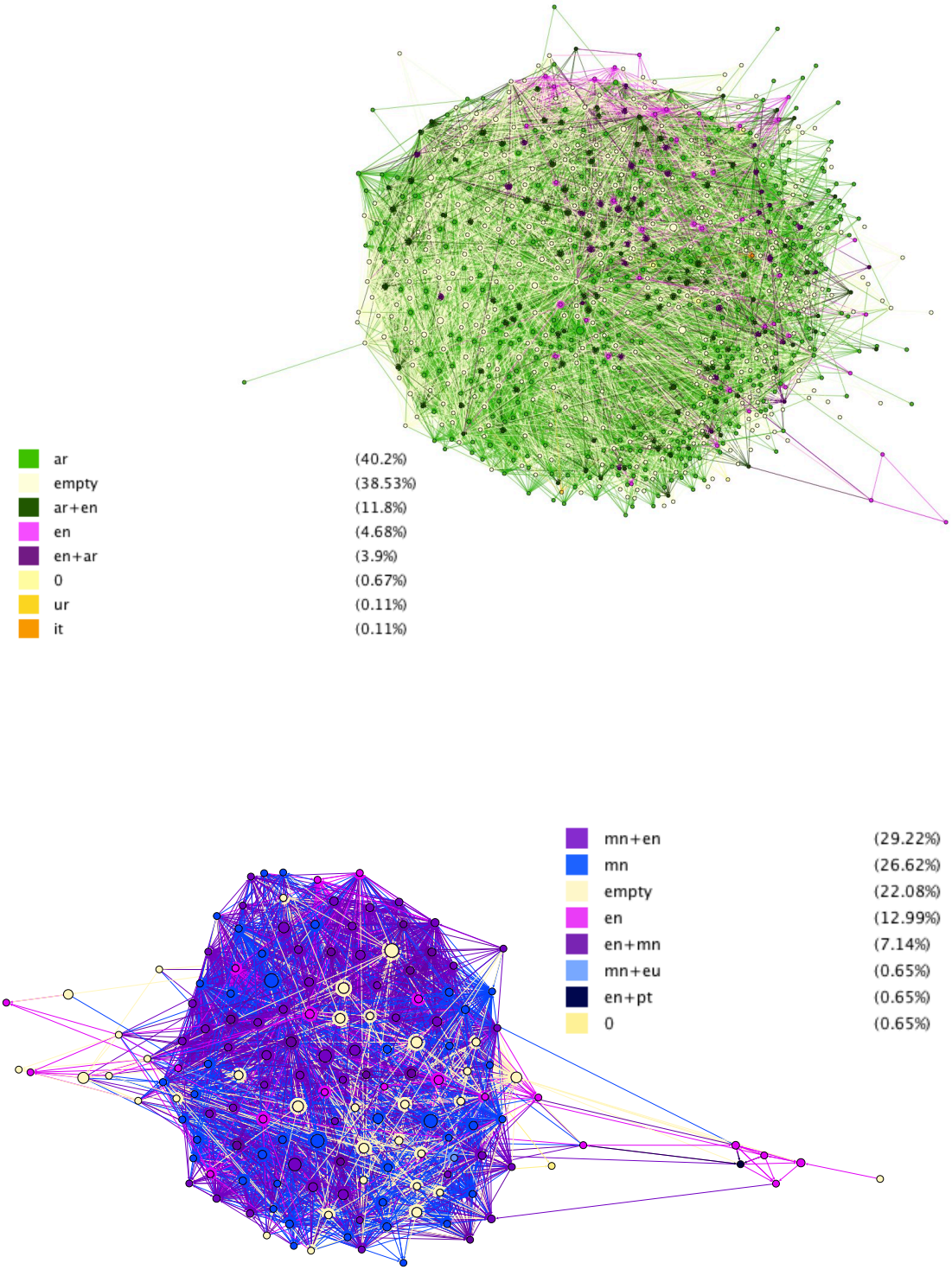


Figure A.27: Bilingual networks: integration type (8).

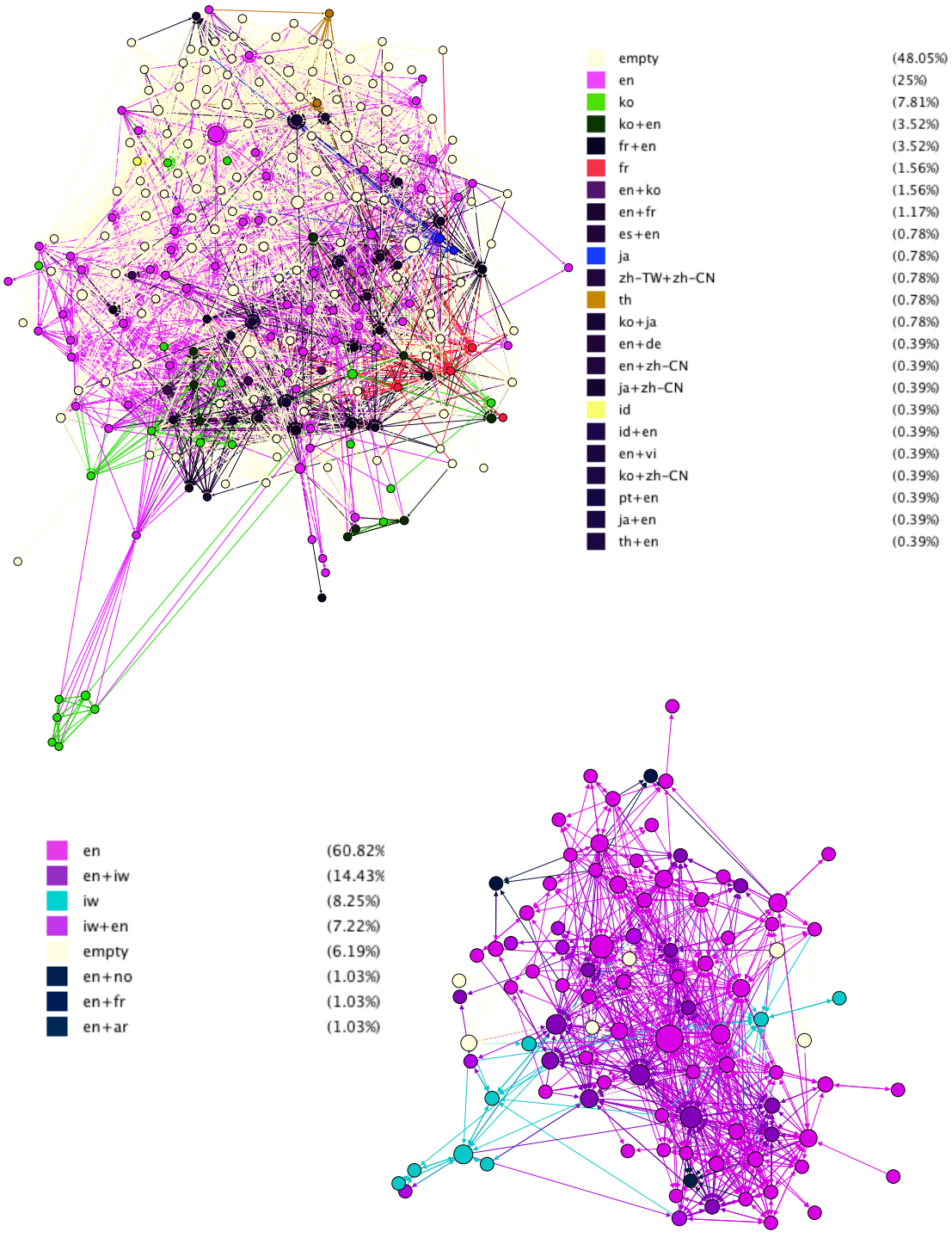


Figure A.28: Bilingual networks: peripheral language type (1).

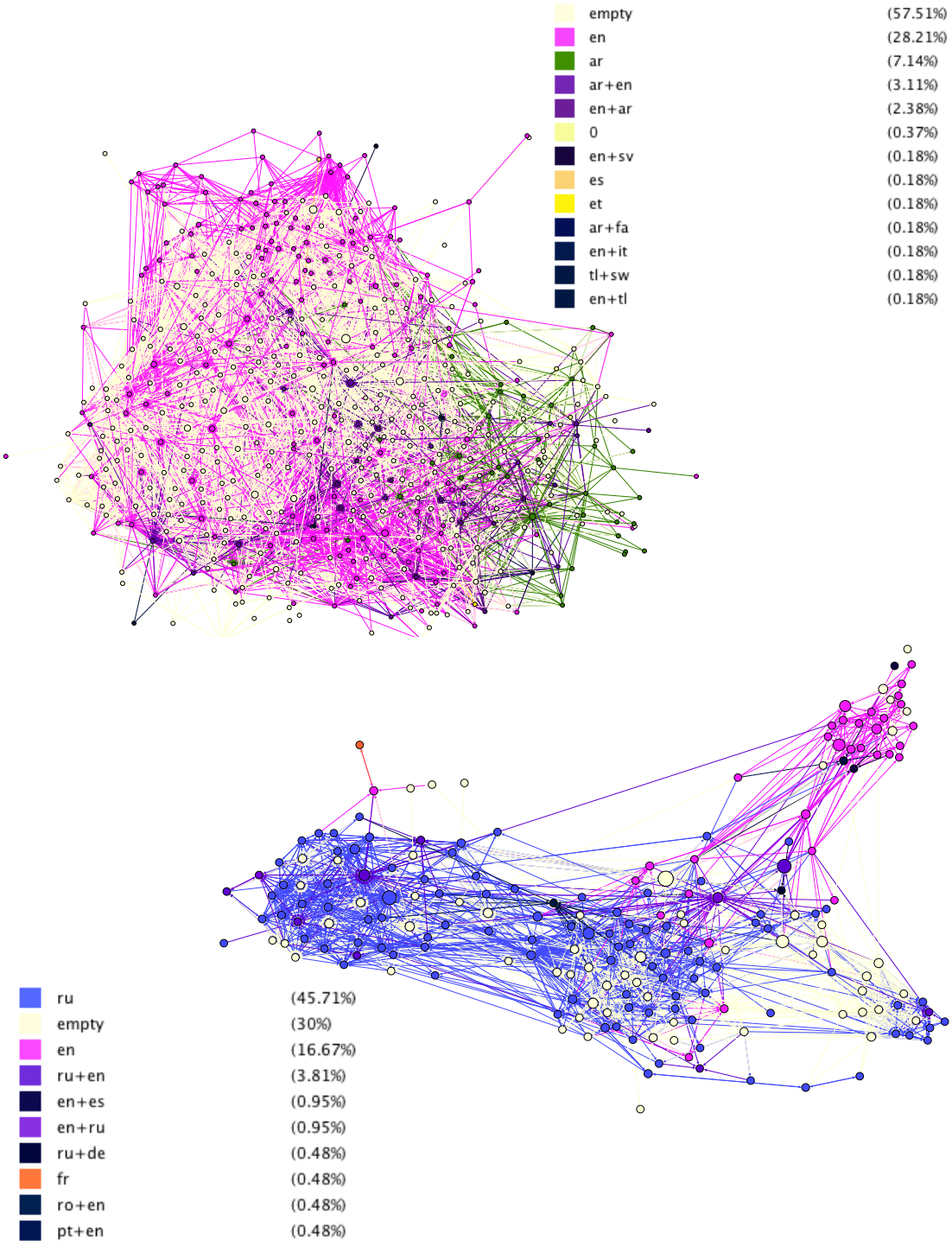


Figure A.29: Bilingual networks: peripheral language type (2).

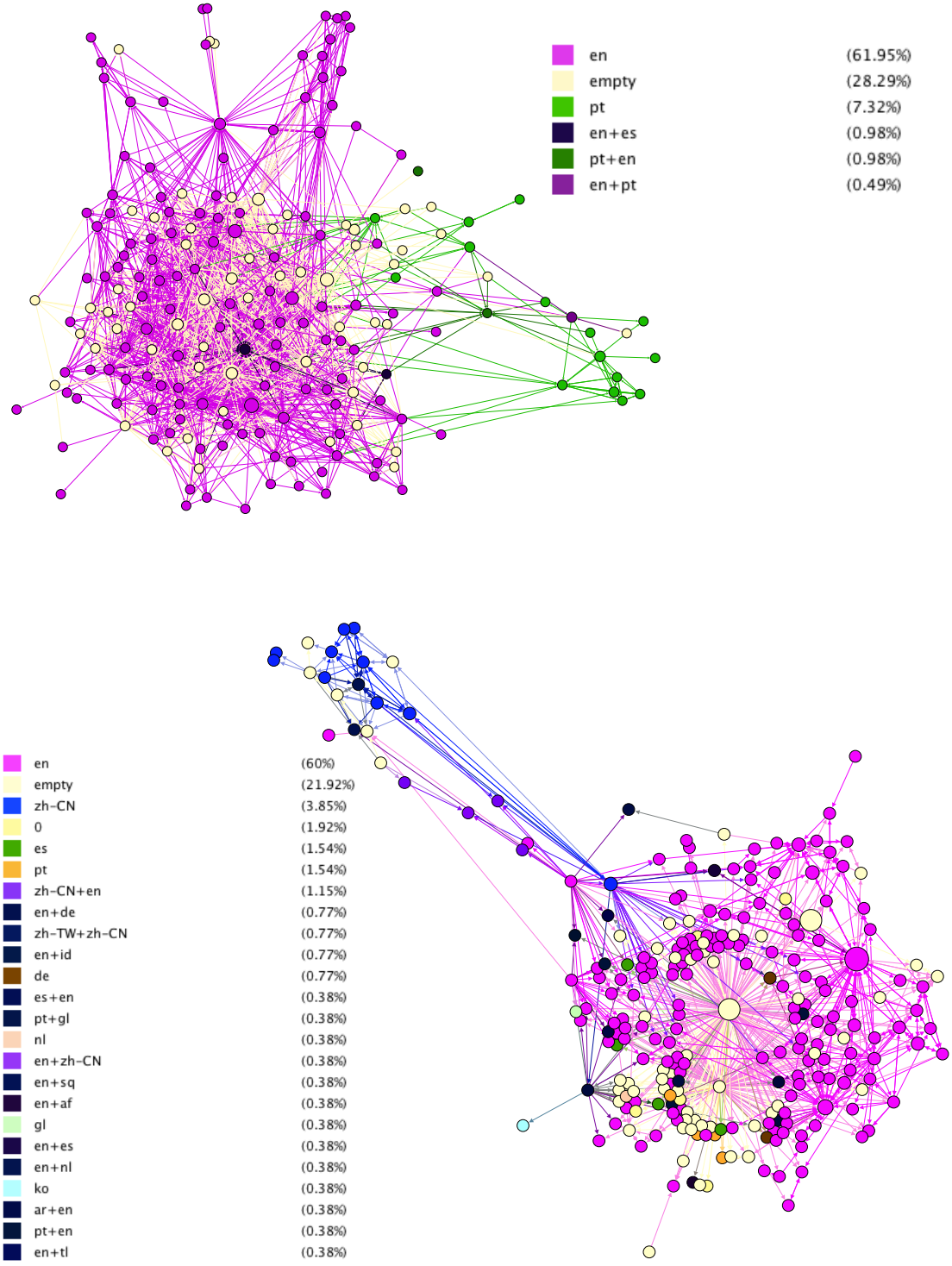


Figure A.30: Bilingual networks: peripheral language type (3).

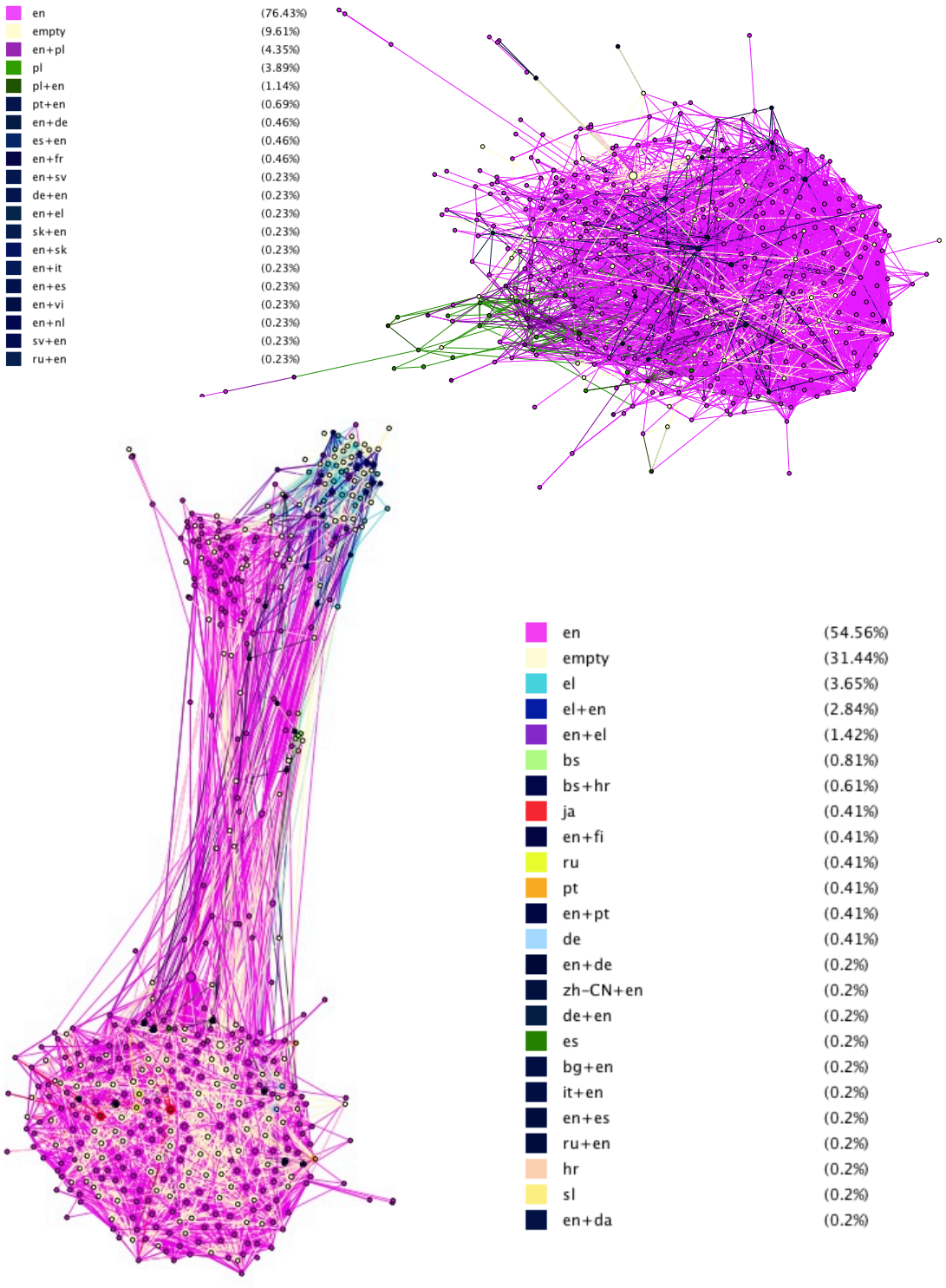


Figure A.31: Bilingual networks: peripheral language type (4).

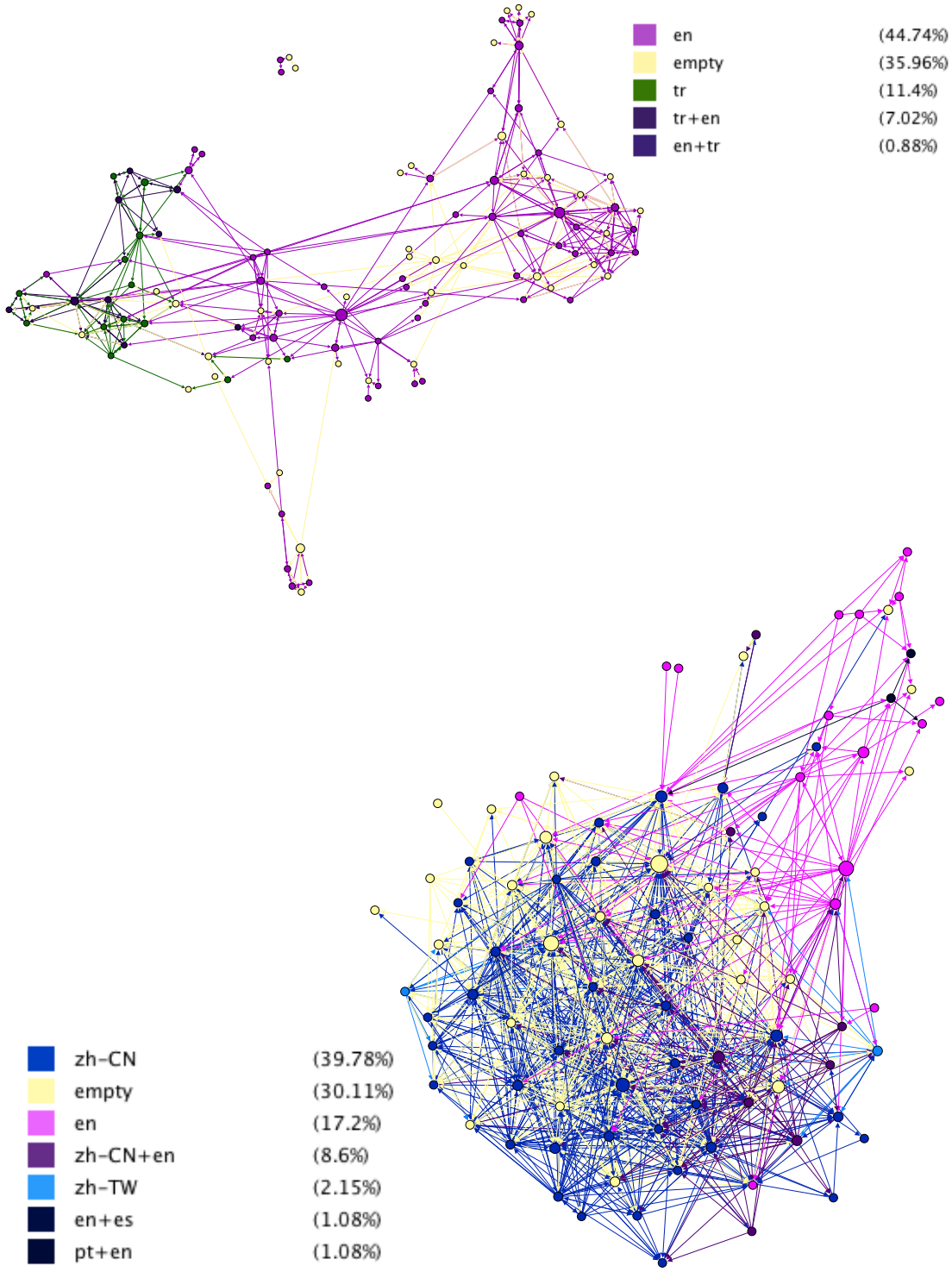


Figure A.32: Bilingual networks: peripheral language type (5).

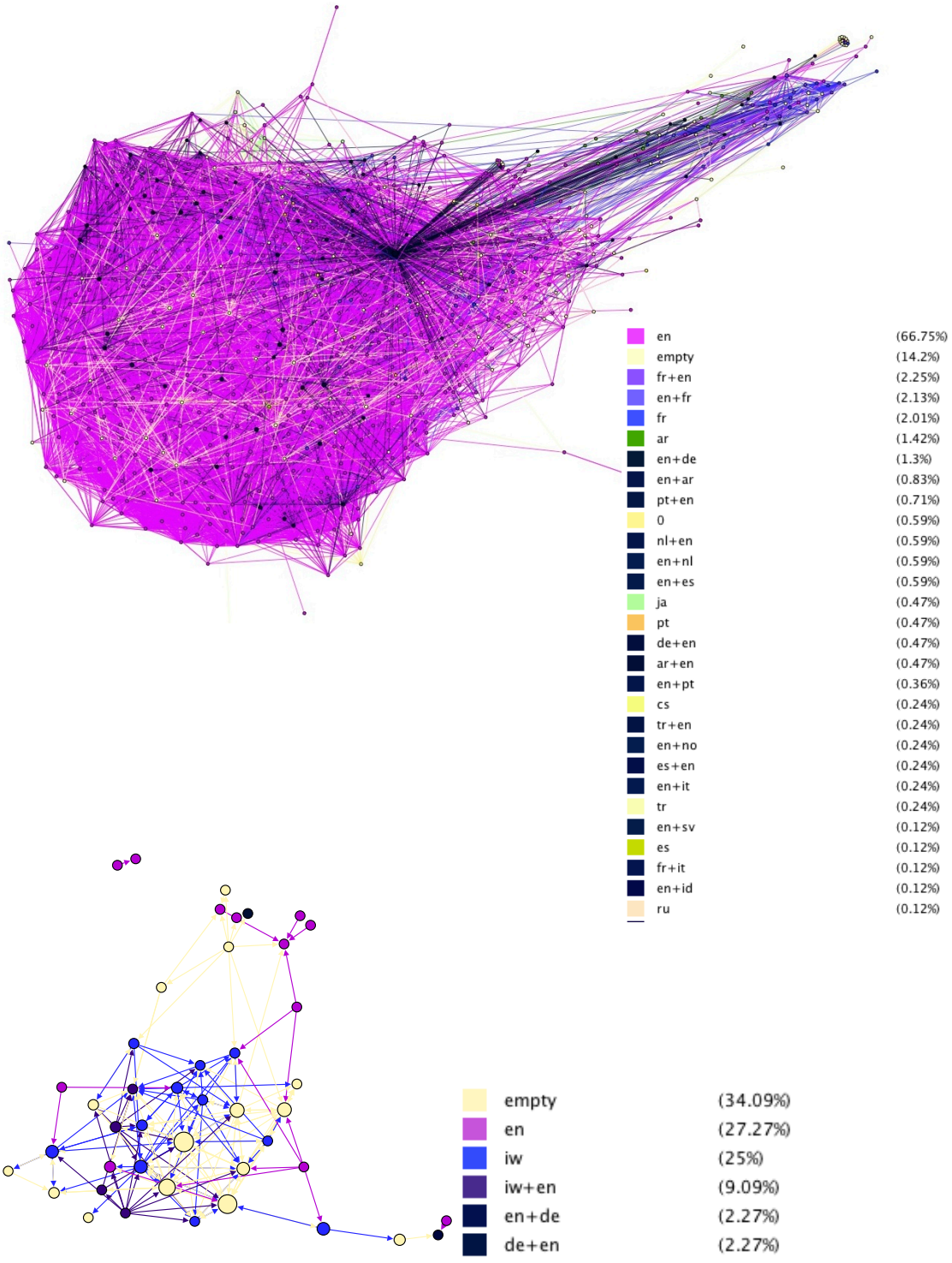


Figure A.33: Bilingual networks: peripheral language type (6).

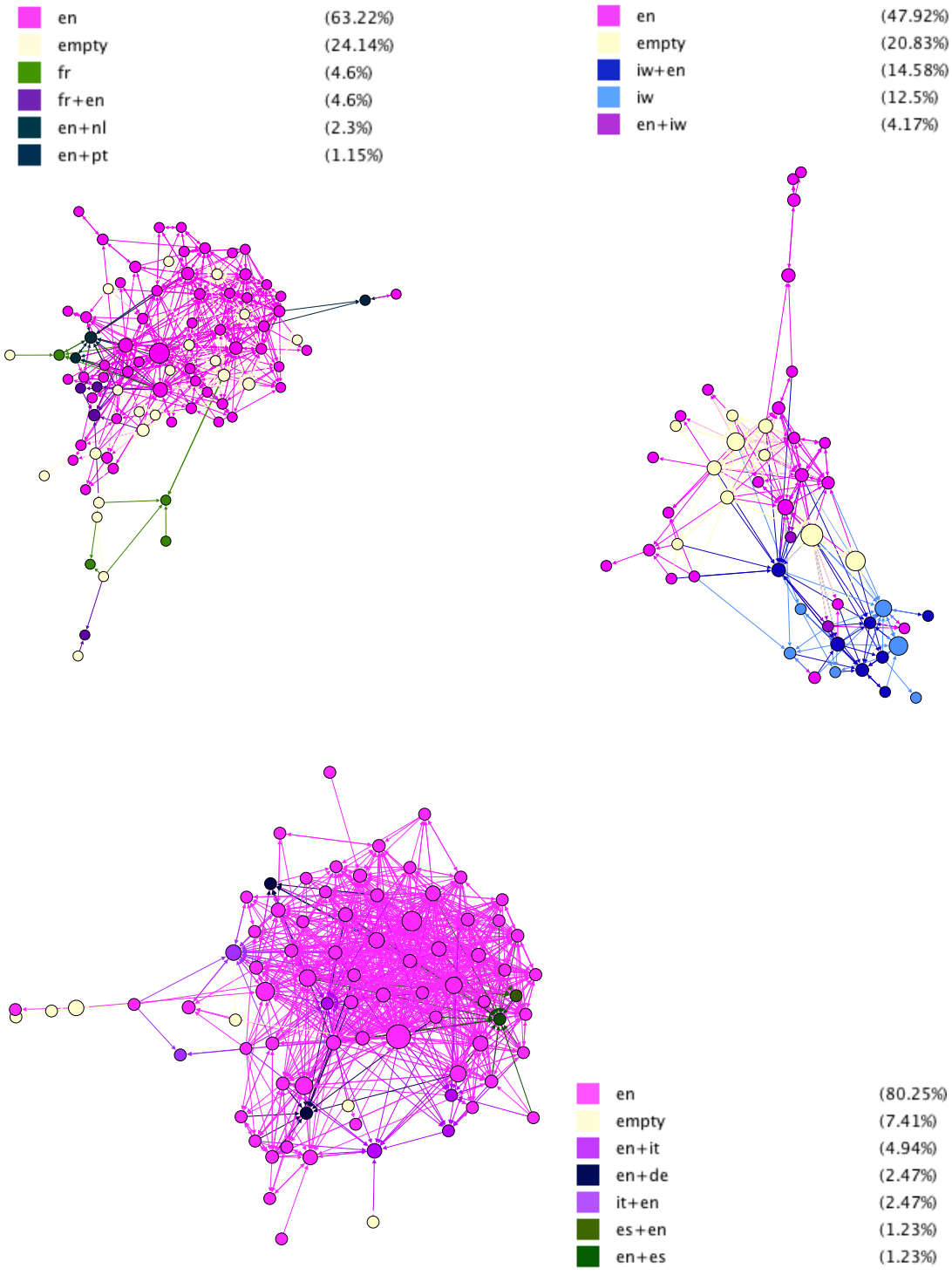


Figure A.34: Small and monolingual networks (1).

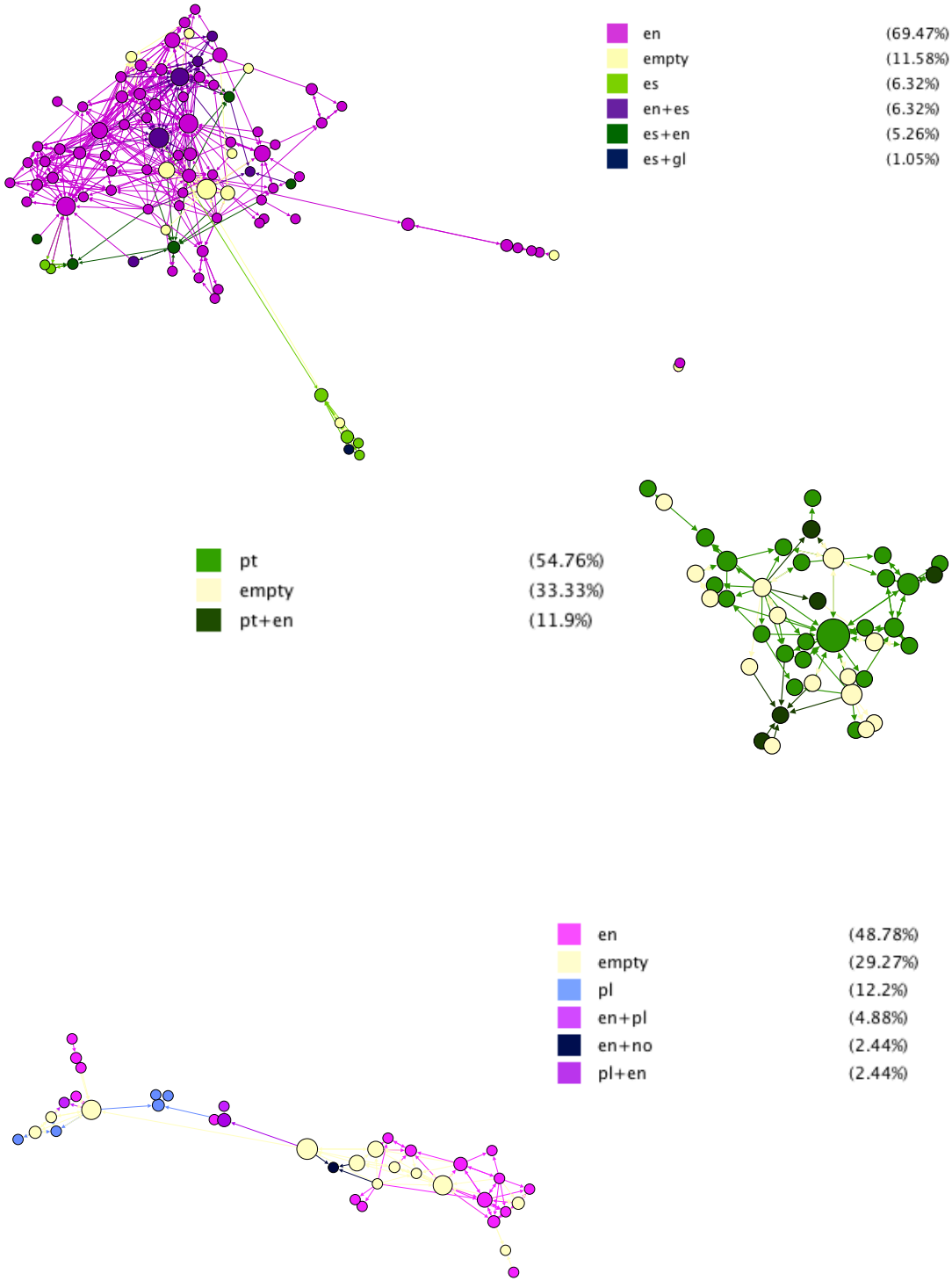


Figure A.35: Small and monolingual networks (2).

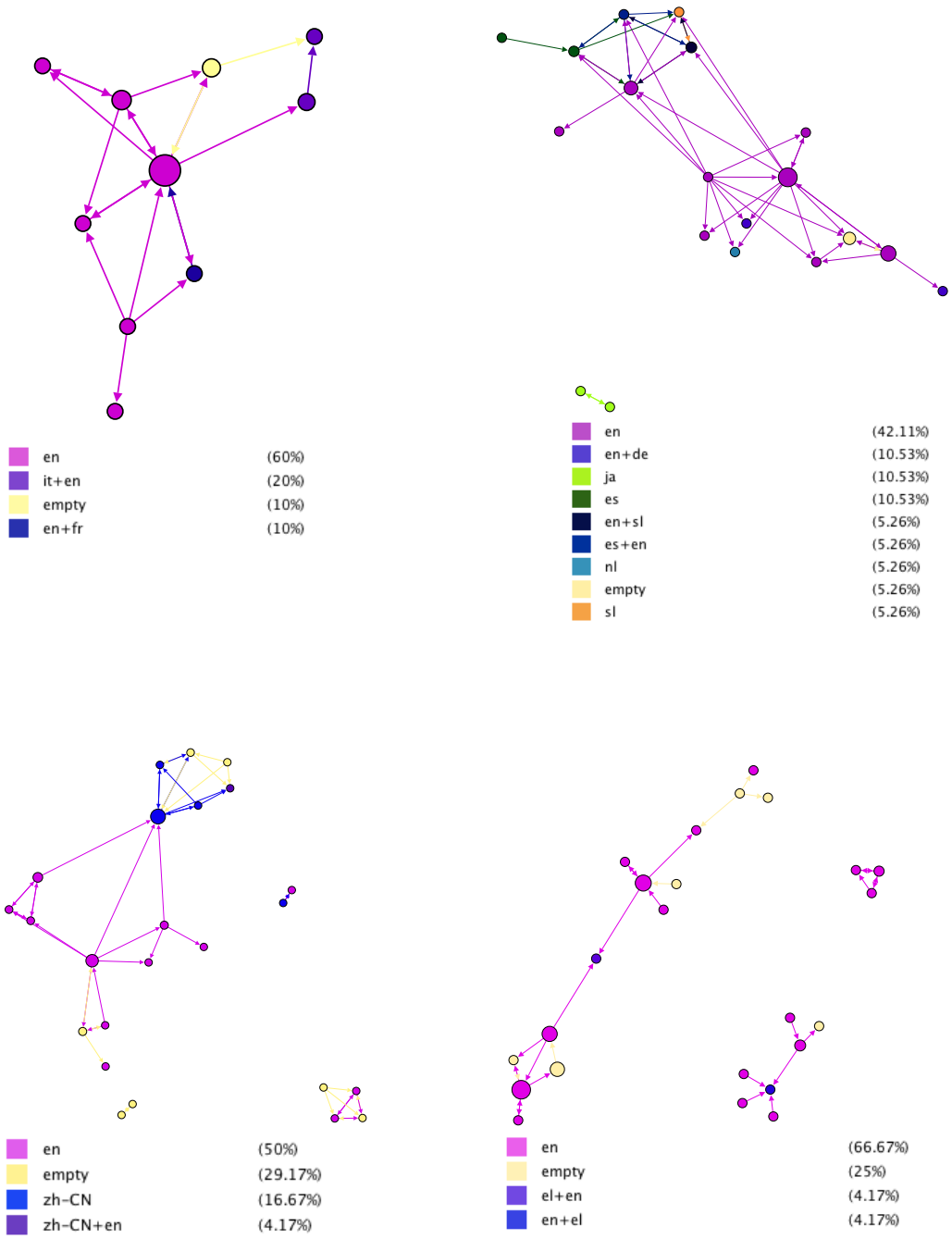


Figure A.36: Small and monolingual networks (3).

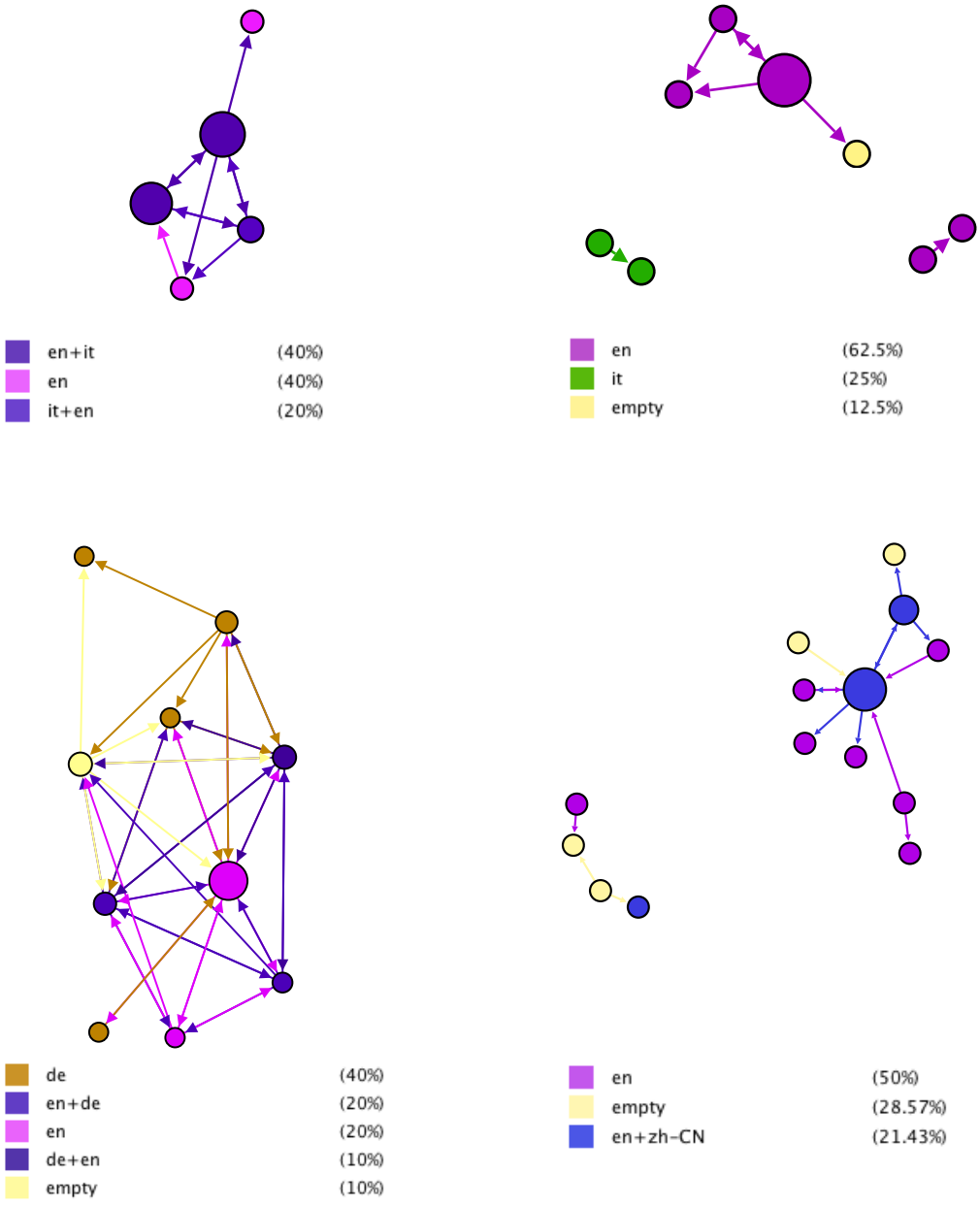


Figure A.37: Small and monolingual networks (4).

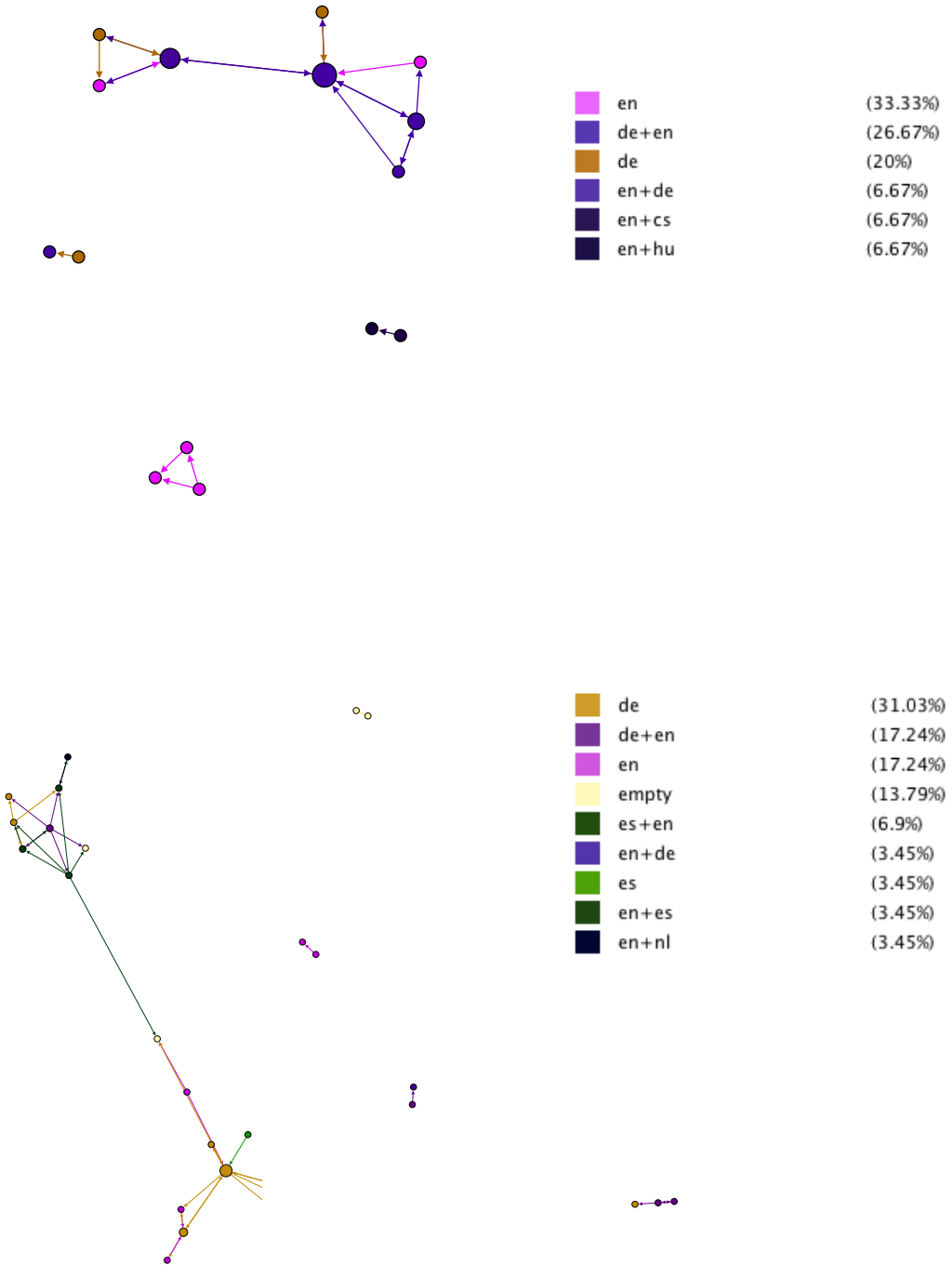


Figure A.38: Small and monolingual networks (5).

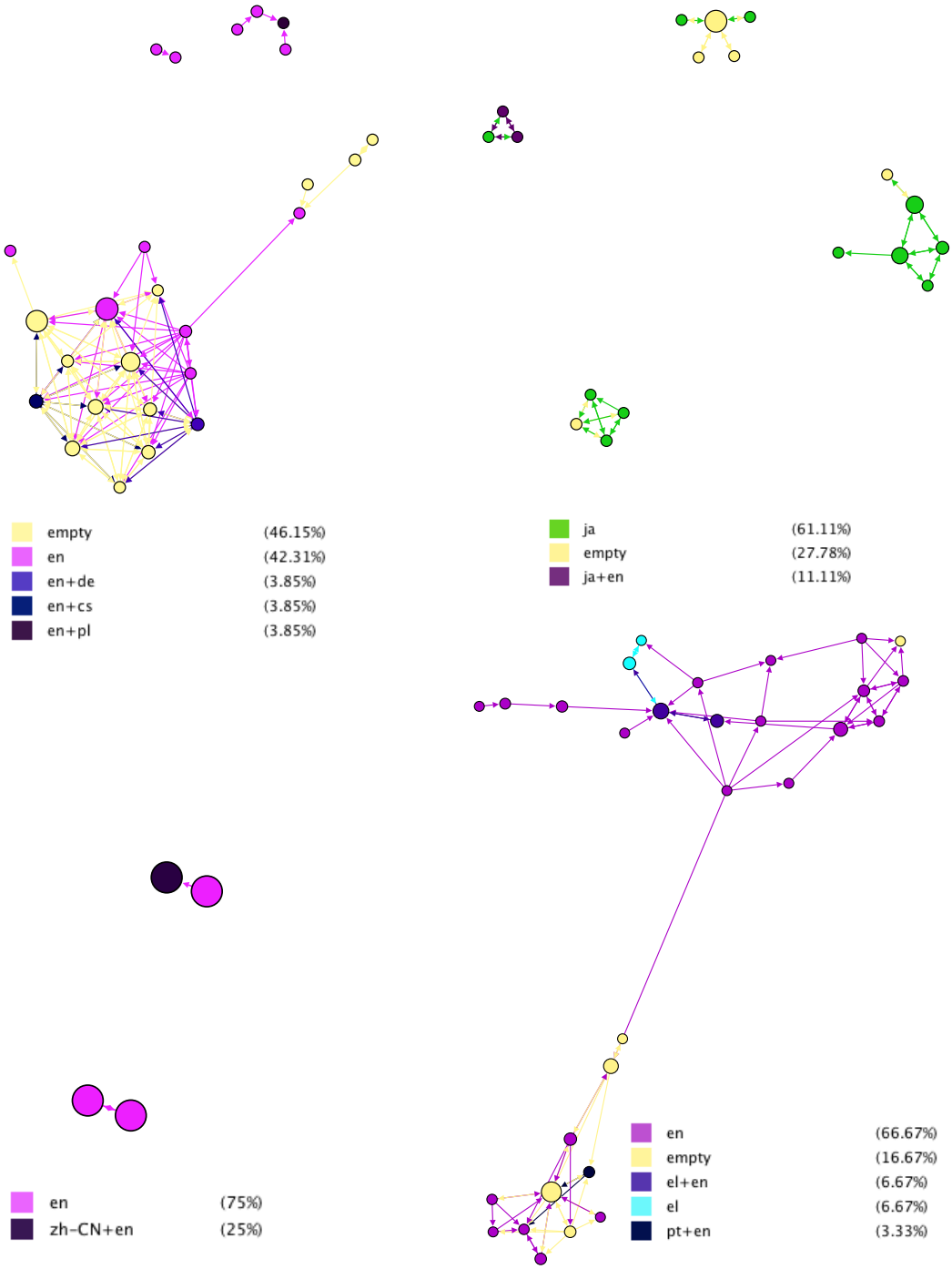


Figure A.39: Small and monolingual networks (6).

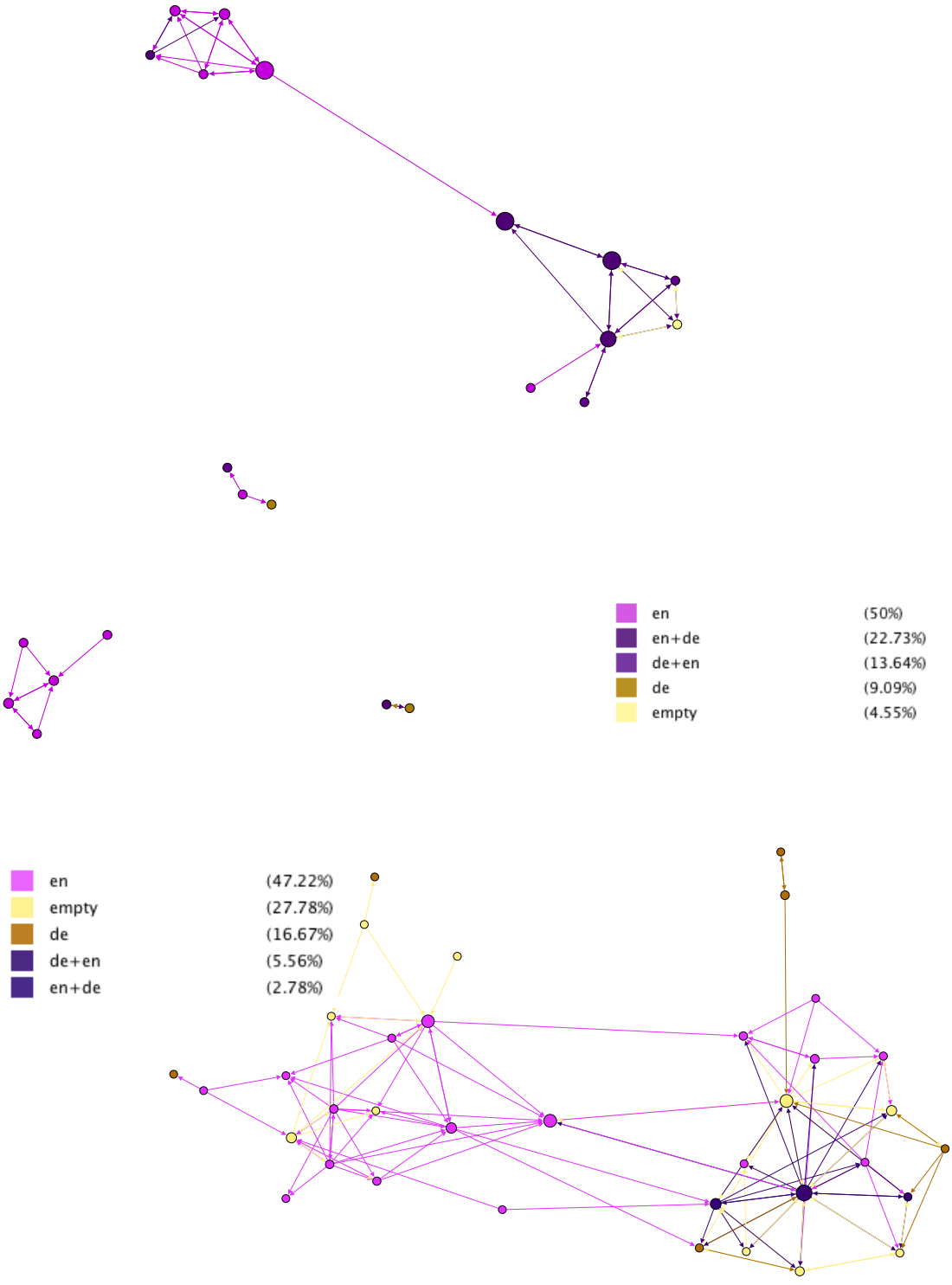


Figure A.40: Small and monolingual networks (7).

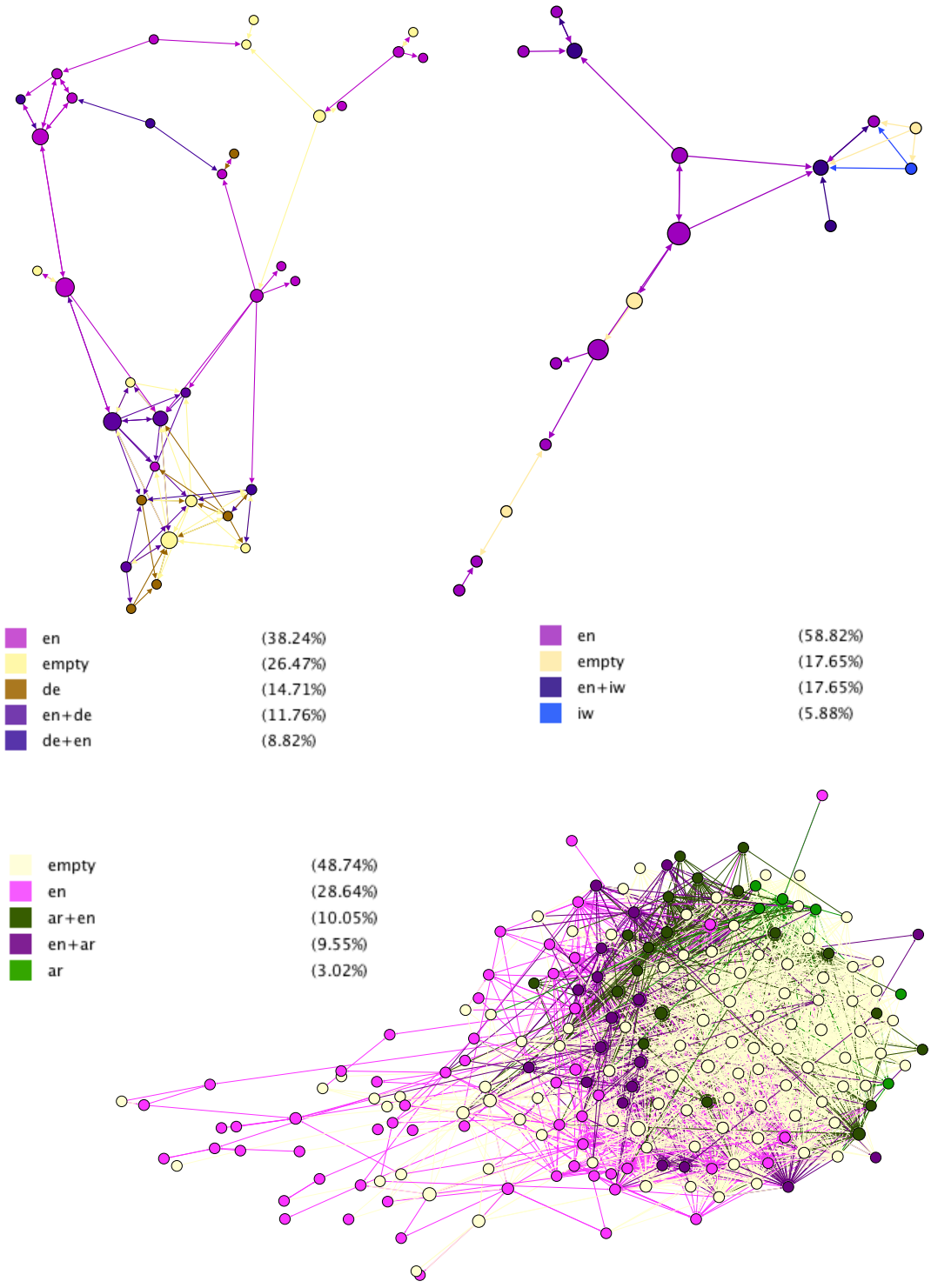


Figure A.41: Small and monolingual networks (8).

Appendix B

International Themes in English Posts

In the English language set of posts written by the 92 egos, totaling 2151 posts, there are 227 posts in which a non-English speaking place or language is mentioned. This appendix presents in landscape layout the complete list of places and languages in the central column, with an extract of the textual context on the left side, and associated theme on the right side.

TEXTUAL CONTEXT	PLACE OR LANGUAGE	THEME
Paris calling for a meeting at... [...]	Paris	travel plans
can't wait to fly to #Barcelona	Barcelona	travel plans
Inspiration for a little Lyon break […]	Lyon	travel recommendation
If a Lille visit is on the agenda […] I'd highly recommend the LAM museum […]	Lille	travel recommendation
Now considering Martinique & Guadeloupe for my next holiday. […]	Martinique & Guadeloupe	travel plans
Oh dear! Explosion @Moscow airport. […]	Moscow	International news
Is going to the Chinese for dinner […]	Chinese	gastronomy and restaurants
visited the 4th marketing-day in austria […]	austria	accomplished travel
REQUESTING THE #NKOTBCRUISE2012 ON THE MEDITERRANEAN SEA! […]	Mediterranean Sea	travel plans
[…] Hi back from germany ;o)	germany	accomplished travel
[…] Germany in 3 Days […]	Germany	travel plans
[…] Zimbabwe and Mugabe's Rule […]	Zimbabwe	International news
Afrikaans Video: giving fracking a drilling […]	Afrikaans	Language of resource
[...] It's 3AM here in France […]	France	location
[...] Venice, tomorrow 13:20? [...]	Venice	travel plans
Virtual #grocery shopping experience in Korea […]	Korea	tech internationalization
[…] Bye Germany	Germany	location
[...] you are at the Schokoladenmuseum. […]	Schokoladenmuseum	location
[…] we need you in Oberhausen! (: How long will you stay in Hamburg?	Oberhausen and Hamburg	location
[…] I have heard you speak german […]	german	Language
[…] stormy belgium	belgium	location
Belgium fries candidates to UNESCO patrimony!	Belgium	International news
Arab Revolution power […]	Arab Revolution	International news
Had to call my Pops so he could send me a couple of Mrs. Dash seasonings to Puerto Rico!	Puerto Rico	location
[…] Victor Manuelle is a Puerto Rican artist! […]	Puerto Rican	culture
[…] Have a good one and salute from Puerto Rico! […]	Puerto Rico	location
looking forward to eating china food […]	China	gastronomy and restaurants
We had so much fun today :D. Amazing audience in Küttigen!! [...]	Küttigen	location
Off to Geneva for a show at a fancy birthday party :-)	Geneva	travel plans
Tonight we're performing at Stadtcasino Frauenfeld. That's the theatre where […]	Stadtcasino Frauenfeld	location
Buskers Festival in Bern was fantastic!	Bern	event location
We have the honor of performing at one of Switzerland's most exclusive Hotels tonight […]	Switzerland	location
[…] at the annual Swiss Congress of Radiology […]	Swiss	event location
A special celebration video for the handball club Oberwil/BL (german) […]	german	Language of resource
This weekend we'll be traveling to Barcelona! Can't wait for another amazing European Yo-Yo Meeting. […]	Barcelona and European	travel plans
Tomorrow Saturday: Special midnight blacklight-performance at […] (Zürich Airport).	Zürich	event location
[…] Sven Epiney, Swiss TV presenter […]	Swiss	culture
We're heading to Augsburg/Germany today. […]	Augsburg/Germany	travel plans
We are grateful about contacts during Live! Global Performing Arts Exchange Singapore […]	Singapore	event location
Tomorrow is our concert in Dresden […]	Dresden	event location
Today CD-Release in Germany […]	Germany	event location
[...] will be playing at Kuala Lumpur, Malysia, mid-septembre this year […]	Kuala Lumpur, Malysia	event location
Got home tonight from Berlin […]	Berlin	accomplished travel
The Festival Jazztage Dresden is over. We had great musicians here! […]	Dresden	event location
[…] this Scharansky is very dangerous gangster and the leader of terrible russian triads, be afraid broth.	Russian	opinion
[…] be afraid, Lev Scharansky is native of Brighton-beach Russian mafia , very, very dangerous gangster.	Russian	opinion

I am for communism. Swedish communism	Swedish	opinion
[…] Watch servants of the people in Russia […]	Russia	culture
[…] in Russia and Paris, for the people of those countries are so willing to be amused […]	Russia and Paris	opinion
Did you know that about 50 thousand people are killed from snakebites for a year? It is only in India	India	International news
The Ukrainian experience for Arab world by Lionel Beehner […]	Ukrainian and Arab World	International news
Me-e-e-t U-u-u-kraine - the champion of all antiratings... Now the 18 Countries Most Likely To Default […]	Ukraine	International news
If you want to become a part of Euro-2012 […]	Euro 2012	sports
[…] Brazil... Here I come :D	Brazil	travel plans
[…] almost worth a trip from munich to berlin :-)	munich and berlin	travel plans
Today english garden, chinese tower, german beer! […]	chinese and german	gastronomy and restaurants
Forza Milan ! We are the champions !	Milan	sports
Who's got the bigger melons? Size em up... Turkish Style […]	Turkish	gastronomy and restaurants
[…] An Italian vibe. A moto and small tables. Thin crust pizzas […]	Italian	gastronomy and restaurants
Thinking of Sultans, Belly Dancing, Harem, Gypsy Music, Raki, Meze and Wine... the good old Ottoman Times...	Ottoman	culture
[…] The best stop for Midye Dolma in Town. Part of my IST - Asia tour […]	Asia	travel recommendation
[…] see my latest thoughts on "Where To Stay" in Istanbul. […]	Istanbul	travel recommendation
[…] great Istanbul food. See more at […]	Istanbul	gastronomy and restaurants
Loving Menemen -- Turkish scrambled eggs... […]	Turkish	gastronomy and restaurants
[…] turkish eating party […]	Turkish	gastronomy and restaurants
Just finished church in Nigeria […]	Nigeria	culture
wave two fingers in the air... Nigerian Clubbing etiquette […]	Nigerian	culture
[…] currently in Lagos, Nigeria for weekend between volunteer teaching in Ghana.	Lagos, Nigeria and Ghana	travel plans
[…] organizing a month of teaching English in Ghana for June	Ghana	travel plans
Is happy that the german and physics tests are over;))	german	Language
Haifa was a lot of fun... […]	Haifa	location
So I have a ticket to Amsterdam... now I need to find some one who will come with me to Berlin […]	Amsterdam and Berlin	travel plans
And I think that's the only thing I really really hate about munich	Munich	opinion
[…] oktoberfest, omg so many drunk people...	oktoberfest	opinion
Aaaaahhh so many tourists in my beloved Munich	Munich	opinion
In the aftermath of #Norway attacks, piece by NYTimes on the rise of right-wing movements in Europe […]	Norway and Europe	International news
[…] I hope to see you dancing in Italy (Milan) soon! ;-*	Italy (Milan)	travel plans
[…] Best Wishes and a lot of love from Italy!	Italy	location
[…] But I live in Italy!! :-(	Italy	location
Gestix ERP software penetrating Angola even without local resellers... […]	Angola	tech internationalization
[…] We bet in low cost high Q business software - Portugal / Euro-Asia / USA	Portugal / Euro-Asia	tech internationalization
[…] Gestix Certified from EUR 150 lifetime license […] - Europe and USA-ready	Europe	tech internationalization
Doctors have been sentenced to 15 years in prison in #bahrain for treating protesters […]	Bahrain	International news
bahrain tonight should burn.. […]	Bahrain	opinion
First district training in K-town tomorrow, afterwards first individual training in Stuttgart!	Stuttgart	travel plans
I will start working on […] in Saarbrücken	Saarbrücken	travel plans
I had a great meeting with the board members from US Youth Soccer Europe...[…]	Europe	sports
The Horn of Africa: Chronicle of a famine foretold […]	The Horn of Africa	International news
BBC News - Japan pensioners volunteer to tackle nuclear crisis […]	Japan	International news
The Norway attacks: Manifesto of a murderer […]	Norway	International news
[…] there's no Such a thing as tourism in Jordan !! How's anything done in this country ?	Jordan	opinion
Given the birth rate & z population figures in Egypt, I can't understand how sex is a taboo #justsaying	Egypt	opinion
next step Japan: Tokyo!! […]	Japan, Tokio	travel plans

My internet connection sucks in Chapala.	Chapala	location
[…] Intl. Conf. on Information, Process, and Knowledge Management in Valencia, Spain […]	Valencia, Spain	event location
[…] Austria adopts CC-BY as nation-wide default!	Austria	International news
In Haifa, Israel, supporting free, collaborative, and open knowledge at #wikimania 2011[…]	Haifa, Israel	event location
#icwsm2011 will take place next week in Barcelona […]	Barcelona	event location
Video of yesterday's presentation of #powerofopen at @eoi Madrid	Madrid	event location
At the presentation of @creativecommons book #thepowerofopen at @eoi Madrid	Madrid	event location
[…] can't wait to listen them at Japan:) […]	Japan	travel plans
[…] needs to tour with […] in Japan! pleaseeeeee!!	Japan	travel recommendation
[…] JAPAN IS ON #THEVERGE. HMV Tokyo w/ TFT on display […]	Japan, Tokio	opinion
Leopard Trek to lead tribute to Weylandt in Giro d'Italia [...]	Italia	sports
Mario Cipollini's Milan-San Remo form guide [...]	Milan-San Remo	sports
A visit to Orbea premises and the Basque Country is always fun […]	Basque Country	travel recommendation
Tour of Qatar already history, Tour of Oman starting.	Qatar and Oman	travel plans
Shanghai, China (PVG) Atlanta (ATL) Jun 5, 2011; Tues/Sun westbound; Wed/Mon eastbound.	Shanghai, China	travel plans
Bahnunglück in China - Train accident in China [...]	Bahnunglück in China	International news
[…] I use it in China to redirect my website […]	China	tech internationalization
at Sapienza...[...]	Sapienza	location
all you can eat at thai-jap rest...[...]	Thai-jap	gastronomy and restaurants
Going to Roma by train...	Roma	travel plans
[…] for chileans, a touristic place in London is "The Clinic".	chileans	opinion
#4deagosto y #cacerolazo banging on a pot for better education in #Chile	Chile	location
#chile #students #4deagosto several pictures	Chile	event location
#4deagosto […] trending since the students in Chile are protesting to reform education inequality and cost […]	Chile	International news
[…] German Rugby Championship of the Universities !!	german	sports
[…] useless trivia: Weihenstephan is the eldest brewery in the world ! Big Cheers from Munich !	Weihenstephan, Munich	culture
Italy wins vs France 22:21 so amazing	Italy, France	sports
[…] European Tech Tour Gala dinner on Wednesday night in berlin	European and berlin	event location
CROSS INNOVATION ACADEMY this Thursday in Bonn […]	Bonn	event location
Madvertise @ grow in munich	Munich	location
madvertise will celebrate its series A closing party on 29.04. in Berlin -- hope to see you there! […]	Berlin	event location
BBC News - Greece says debt talks to avert default 'productive' […]	Greece	International news
[…] to escape from miserable greek reality	greek	opinion
[…] are you sure of this piece of news coz Egypt can't take the consequences […] #Egypt#Jan25#Mubarak	Egypt	reaction to International news
Egypt we are your protectors and your builders and we will start from scratch […]	Egypt	opinion
[…] all i can think of is that after every rainfall must come a rainbow waiting for Egypt's rainbow […]	Egypt	opinion
Feeling 3 m high just for being an Egyptian,Dear country I love you […]	Egyptian	opinion
Security Theater Lessons From Utøya […]	Utøya	International news
[…] is off to a great start building a reliable and professional taxi network in Athens!	Athens	location
[…] I just read it: "Can Greeks Become Germans?" […]	Greeks, Germans	opinion
Fake Apple store in China […]	China	International news
[…] Student from Sweden sent me […]	Sweden	location
Greece definitelly needs its stateleaks, too […]	Greece	opinion
Getting ready for another sunset in Seychelles... […]	Seychelles	location
I know Israel is an internationally known start-ups maker, I just love being reminded […]	Israel	culture
Microsoft Country Manager In Libya Detained By Authorities […]	Libya	International news
[…] You are always invited back to Israel. The summer here is amazing :)	Israel	travel recommendation

[…] You should come to Israel during the summer, you'll have a blast!	Israel	travel recommendation
Israeli band Orphaned Land rocks Turkey, despite discord […]	Israeli and Turkey	culture
My daughter is coming back home from Israel and I'm waiting by the gate (@ Amsterdam Airport Schiphol)	Israel and Amsterdam	location
New week (@ UPC Nederland w/ 2 others) […]	UPC Nederland	location
[…] Liberation Day by the Papal State :: #welldone #italy #papal #carnival […]	Italy	culture
[…] Is this for real? (Hebrew) […]	Hebrew	Language of resource
Outbrain's weekend at Jerusalem :) […]	Jerusalem	location
[…] I didn't know there was one in Haifa […] #gdd11	Haifa	event location
We will play a guest show on the upcoming Black Trolls Over Europe Tour. […] next week in Traun, Austria!... […]	Europe and Traun, Austria	travel plans
Tomorrow we will rock Vienna!	Vienna	travel plans
Tickets for our 5th Anniversary concert in Graz are now available![…]	Graz	event location
[…] For this special occasion we will play a show in Graz […]	Graz	event location
Franz Loechinger will be drumming for ILLUMINATA on Fr. 4.3. in Graz (Explosiv) […]	Graz	event location
Franz Loechinger will hit the drums on the Black Trolls Over Europe Tour […]	Europe	event location
New Review 9/10 Points (German) […]	german	Language of resource
Apple removed an app of the Palestine Third Intifada just like facebook, Israel is controlling Media?	Palestine and Israel	opinion
@adobe should organise an event for nord africa just like @google (gmaghreb)	Nord Africa	event location
Hope that everyone in Japan is fine... #prayforjapan	Japan	reaction to International news
Google china is a joke	China	tech internationalization
Cooking in french :p	french	Language
[…] You might be interested at this: Are Chinese moms better than Western moms? […]	Chinese	opinion
Are Chinese moms better than Western moms? […]	Chinese	opinion
Staying in Montreal, learning French simultaneously […]	Montreal and French	location, language
Madrid is much better choice :) RT […] Paris is well located to […] between Seattle and Beijing […]	Madrid, Paris, Beijing	opinion
Please consider coming to Spain, too. Thanks for an unforgettable time!	Spain	travel recommendation
[…] Demonstrations all over #spain since last sunday #15m for #real #democracy	Spain	International news
[…] National Researchers System (SNI in Spanish) […]	Spanish	Language
Bomb attack at Moscow airport […]	Moscow	International news
Archaic Denisovans (hominin group) contributed to modern Melanesians! […]	Melanesians	International news
short english translation (sorry for the bad english) for out international Fans in UK, Russia, Brazil […]	Russia and Brazil	location
The Dissociates coming to Germany for the "HighFünf tour" […]	Germany	event location
Come to see us and our british friends from The Dissociates in Aachen […]	Aachen	travel plans
[…] All the video material is from their European Tour with us last year! […]	European	event location
Aida conference in Pavia on social networks	Pavia	event location
Editing an article about Italian case-law on liability of ISPs	Italian	a country's policy
Look forward to experiencing new Italian opposition procedure	Italian	a country's policy
Miss World Tourism 2011, In Kefalonia […]	Kefalonia	event location
Summer Night in Argostoli (Kefalonia)... […]	Argostoli (Kefalonia)	location
Swimming in the winter sea of Lixouri (Kefalonia)... […]	Lixouri (Kefalonia)	location
#TribalDDB Lisbon manages 2 of the most engaging Facebook Pages in Portugal […]	Lisbon, Portugal	tech internationalization
[…] I´m anxious to buy the Tower of Belem, here in Lisbon :-)	Lisbon	location
Got a great time at hyperisland in Barcelona […]	Barcelona	location
Under Siege™, the portuguese RTS videogame for PS3 won today the first prize […]	Portuguese	tech internationalization
Portuguese RTS game for PS3 gets an incredible cinematic trailer […]	Portuguese	tech internationalization
Portugal Gives Itself a Clean-Energy Makeover	Portugal	International news
@phillord @dullhunk I bet that's his name in greek	greek	Language
@timoreilly […] basques also :-) […]	basques	tech internationalization

We want to translate Twitter to Basque,support us! […]	Basque	tech internationalization
Tu BeAv - The jewish holyday of Love in Israel	Israel	culture
Travel in Eilat is over. The Red Sea is amazing and the water is so clear. but Tel Aviv weather […] much better.	Eilat, Red Sea, Tel Aviv	travel plans, opinion
Happy Israel Independence Day! Fireworks in the sky tonight!	Israel	culture
Good dim sums in Brussels? Does it even exist?	Brussels	opinion
in brussels ... no diving sites :(	Brussels	location
Arrived in Melaka in Malaysia, but everything is closed early tonight. Will check Chinese shopping tomorrow :)	Melaka, Malaysia, Chinese	location
Rain! And a spicy Chinese nuddles […]	Chinese	gastronomy and restaurants
[…]...but just vocabulary). I have one but in Polish and I want something like this in English […]	Polish	Language
[…] I'm looking for some computer program to learn German vocabulary […]	german	Language
preparing to dance my ass off for haiti!!!!	haiti	reaction to International news
tomorrow at mosaiic bar frankfurt...	Frankfurt	travel plans
[…]..finished munich...off to langenselbold with […]	munich, langenselbold	travel plans
tonight at mosaiic bar frankfurt...[…]	Frankfurt	location
back from berlin...tired now	Berlin	accomplished travel
getting my hair cut, then flying to berlin...	Berlin	travel plans
berlin is calling and i am following.everybody from berlin meet me at […]	Berlin	travel plans
[…] i love busted!!![…]....Spain are with you!!!;)))	Spain	opinion
Brazil: Death of Forest Defender Couple is a Shame to the Country […]	Brazil	International news
[…] Greek and Spanish young people occupying Trafalgar square #europeanrevolution #ukrevolution #London	Greek, Spanish, european	International news
Tim Hetherington is killed in #Misrata! […] #Libya	Misrata, Libya	International news
[…] Stay in Japan to work for this f**king company? NO WAY!!	Japan	location
Won won won!!! Japan has become the Queen!!!	Japan	sports
Sorting algorithms demonstrated with Hungarian folk dance […]	Hungarian	culture/humor
Which countries match the GDP of America's states? […] California is Italy! But... Italy has 20M more people...	Italy	International news
beautiful night view of Italy taken from International Space Station #ISS […]	Italy	International news
Why Young Italians Are Leaving […]	Italians	International news
[…] China's fake Apple stores […]	China	International news
World Cup joy for Japan […]	Japan	sports
Watching on ITV1 #England vs #Switzerland […]	Switzerland	sports
Hakuho Mongolia's best paid sports star - Yahoo! Eurosport […]	Mongolia	sports
[…] Chinese official media […] quoted ur opinion about Obama from twitter (a website blocked in China) !!! […]	Chinese, China	tech internationalization
[…] U ARE FAMOUS IN CHINESE TWITTER (WEIBO) […]	Chinese	tech internationalization
#molk BUT this thought is based on the report I read in the CHINA. […]	China	reaction to International news
Tomorrow will be my first spanish test […]	Spanish	Language
Bombing in Oslo and shooting at Utøya ! […]	Oslo and Utøya	International news
Japan won !! […] #worldcupfinal	Japan	sports
Netflix Brasil blog […]	Brasil	tech internationalization
Watching a 1965 BW Polish film set in Spain... […]	Polish, Spain	movies
Listening to Spanish football games […]	Spanish	sports

Bibliography

[1] Workshop on novelty and diversity in recommender systems - DiveRS 2011. In Pablo Castells, Jun Wang, Rubén Lara, and Dell Zhang, editors, *Proceedings of the fifth ACM conference on Recommender systems - RecSys '11*, pages 393–394, New York, New York, USA, October 2011. ACM Press. 1.2.1

[2] A Toolkit for Transnational Communication in Europe. In J. Normann Jø rgensen, editor, *The Copenhagen Studies in Bilingualism Vol. 64*, 2011. 2.1

[3] Proceedings of the 3rd Workshop on the Multilingual Semantic Web (MSW3). In Paul Buitelaar, Philipp Cimiano, David Lewis, James Pustejovsky, and Felix Sasaki, editors, *International Semantic Web Conference*, volume 936, Boston, 2012. CEUR. URL `http://ceur-ws.org/Vol-936/`. Last accessed Oct 30, 2013. 1.2.1

[4] Meshary AlMeshary and Abdolreza Abhari. A recommendation system for Twitter users in the same neighborhood. In *Proceedings of the 16th Communications & Networking Symposium*, pages 1–5, San Diego, California, April 2013. Society for Computer Simulation International. 8.2.1

[5] Jannis Androutsopoulos. Language Choice and Code Switching in German-Based Diasporic Web Forums. In Brenda Danet and Susan Herring, editors, *The Multilingual Internet: Language, Culture, and Communication Online*, chapter 15. Oxford University Press, New York, 2007. 2.5, 2.5, 3.1, 8.1

[6] Jannis Androutsopoulos. Localizing the Global on the Participatory Web. In Nikolas Coupland, editor, *The Handbook of Language and Globalization*, chapter 9, pages 203–231. Wiley-Blackwell, Malden, MA, 2010. 1.2.2, 2.4, 2.4.1, 2.5, 4.6, 8.2.2

[7] Albert-Laszlo Barabasi. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume, 2003. 2.3

[8] Bettina Berendt and Anett Kralisch. A user-centric approach to identifying best deployment strategies for language tools: the impact of content and access language on Web user behaviour and attitudes. *Information Retrieval*, 12(3): 380–399, January 2009. 1, 2.4, 2.5, 3.1

[9] Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. Language identification for creating language-specific Twitter collections. In *LSM '12 Proceedings of the Second Workshop on Language in Social Media*, pages 65–74. Association for Computational Linguistics, June 2012. 3.3, 4.7

[10] Carter T. Butts. Social network analysis: A methodological introduction. *Asian Journal Of Social Psychology*, 11(1):13–41, March 2008. 5

[11] Louis-Jean Calvet. *Towards an Ecology of World Languages*. Polity Press, Cambridge, 2006. 2.2

[12] Mónica Stella Cárdenas-Claros and Neny Isharyanti. Code switching and code mixing in Internet chatting: between yes, ya, and si a case study. *The Journal of the JALT CALL SIG*, 5(3):67–78, 2009. 3.1

[13] Manuel Castells. Communication, Power and Counter-power in the Network Society. *International Journal of Communication*, 1:238–266, 2007. 1, 2.2

[14] Vint Cerf. The Internet is for Everyone, 1999. URL `http://www.internetsociety.org/internet-everyone`. Last accessed Oct 30, 2013. 1.2

[15] Hsia-Ching Chang. A new perspective on Twitter hashtag use: Diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4, November 2010. 7.4.2

[16] Alok Choudhary, William Hendrix, Kathy Lee, Diana Palsetia, and Wei-Keng Liao. Social media evolution of the Egyptian revolution. *Communications of the ACM*, 55(5):74–80, May 2012. 1.2.2, 7.4.2

[17] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc, 3rd edition, 2007. 5

[18] Nikolas Coupland. Introduction: Sociolinguistics in the Global Era. In Nikolas Coupland, editor, *The Handbook of Language and Globalization*, chapter 0, pages 1–27. Wiley-Blackwell, Malden, MA, 2010. 1.2.2, 2.5

[19] Angela Creese and Peter Martin. Introduction to Volume 9: Ecology of Language. In Angela Creese, Peter Martin, and Nancy H. Hornberger, editors, *Ecology of Language - Encyclopedia of Language and Education Volume 9*, pages i–vi. Springer, 2nd edition, 2008. 2.2

[20] David Crystal. *English as a Global Language*. Cambridge University Press, 2nd edition, 2003. 2.4

[21] Daniel Cunliffe, Delyth Morris, and Cynog Prys. Investigating the Differential Use of Welsh in Young Speakers' Social Networks: A Comparison of Communication in Face-to-Face Settings, in Electronic Texts and on Social Networking Sites. In Elin Haf Gruffydd Jones and Enrique Uribe-Jongbloed, editors, *Social Media and Minority Languages: Convergence and the Creative Industries*, pages 75–86. Multilingual Matters, Bristol, Buffalo, Toronto, 2013. 3.1

[22] danah Boyd and Kate Crawford. Six Provocations for Big Data. In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. SSRN Electronic Journal, September 2011. URL `http://papers.ssrn.com/abstract=1926431`. Last accessed Oct 30, 2013. 4.9

[23] Brenda Danet and Susan Herring. Introduction: Welcome to the Multilingual Internet. In Brenda Danet and Susan Herring, editors, *The Multilingual Internet: Language, Culture, and Communication Online*, chapter 1. Oxford University Press, New York, 2007. 2.4, 4.7

[24] Stephen Dann. Twitter content classification. *First Monday*, 15(12), November 2010. URL `http://firstmonday.org/ojs/index.php/fm/article/view/2745/2681`. Last accessed Oct 30, 2013. 7.4

[25] Abram De Swaan. The Evolving European Language System: A Theory of Communication Potential and Language Competition. *International Political Science Review*, 14(3):241–255, January 1993. 2.1, 2.5, 5.4

[26] Abram De Swaan. The Emergent World Language System: An Introduction. *International Political Science Review*, 14(3):219–226, January 1993. 2.1

[27] Abram De Swaan. Language Systems. In Nikolas Coupland, editor, *The Handbook of Language and Globalization*, chapter 2, pages 56–76. Wiley-Blackwell, Malden, MA, 2010. 2.1

[28] Murat Demirbas, Murat Ali Bayir, Cuneyt Gurcan Akcora, Yavuz Selim Yilmaz, and Hakan Ferhatosmanoglu. Crowd-sourced sensing and collaboration using twitter. In *2010 IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 1–9. IEEE, June 2010. 1.1

[29] Jay L. Devore. *Probability and Statistics for Engineering and the Sciences*. Thomson Brooks/Cole, Belmont, CA, 7th edition, 2008. 6.2, 6.2

[30] Danny Dor. From Englishization to Imposed Multilingualism: Globalization, the Internet, and the Political Economy of the Linguistic Code. *Public Culture*, 16(1):97–118, 2004. 1, 2.2, 2.4

[31] Mercedes Durham. Language Choice on a Swiss Mailing List. In Brenda Danet and Susan Herring, editors, *The Multilingual Internet: Language, Culture, and Communication Online*, chapter 14. Oxford University Press, New York, 2007. 3.1, 7.2, 8.1

[32] Bruce Etling, John Kelly, Robert Faris, and John Palfrey. Mapping the Arabic blogosphere: politics and dissent online. *New Media & Society*, 12(8):1225–1243, December 2010. 1.2.2, 3.2, 5.4, 8.1

[33] Madelyn Flammia and Carol Saunders. Language as power on the Internet. *Journal of the American Society for Information Science and Technology*, 58 (12):1899–1903, October 2007. 2.4

[34] C. Fuchs. The Role of Income Inequality in a Multivariate Cross-National Analysis of the Digital Divide. *Social Science Computer Review*, 27(1):41–58, April 2008. 1.2, 2.4.2

[35] Gephi.org. Gephi Tutorial Layouts — Gephi.org, 2011. URL `http://gephi.org/tutorials/gephi-tutorial-layouts.pdf`. Last accessed Oct 30, 2013. 5

[36] Jean D. Gibbons. *Nonparametric Statistics: An Introduction (Quantitative Applications in the Social Sciences)*. SAGE Publications, Inc, 1993. 7.2

[37] Global Voices. About Global Voices, 2007. URL `http://globalvoicesonline.org/about/`. Last accessed Oct 28, 2013. 1.2.2

[38] Jennifer Golbeck. *Analyzing the Social Web*. Morgan Kaufmann, 2013. 2.3.1, 2.3.1, 2.3.1, 5, 5.4

[39] David Graddol. English Next. Technical report, British Council, 2006. URL `http://www.britishcouncil.org/learning-research-englishnext.htm`. Last accessed Oct 30, 2013. 2.4

[40] Mark Graham, Scott A. Hale, and Devin Gaffney. Where in the World are You? Geolocation and Language Identification in Twitter. *The Professional Geographer*, 2013. 3.3, 4.7

[41] Mark Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973. 2.3.1, 2.3.1

[42] Mark Granovetter. The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, 1(1983):201–233, 1983. 2.3.1

[43] Jeffrey Graves. Python Language Detector, 2012. URL `https://github.com/decultured/Python-Language-Detector/blob/master/README.md`. Last accessed Oct 4, 2013. 4.3.1

[44] Alexander Halavais. National Borders on the World Wide Web. *New Media & Society*, 2(1):7–28, March 2000. 1, 1.2.1

[45] Scott Hale. Translating Twitter, 2011. URL `http://www.scotthale.net/blog/?p=152`. Last accessed Oct 30, 2013. 1.2.2

[46] Scott Hale. Online language bubbles: the last frontier?, 2012. URL `http://freespeechdebate.com/en/discuss/online-language-bubbles-the-last-frontier/`. Last accessed Oct 23, 2013. 1.2.1, 5.4

[47] Scott A. Hale. Net Increase? Cross-Lingual Linking in the Blogosphere. *Journal of Computer-Mediated Communication*, 17(2):135–151, January 2012. 1, 1.2.1, 3.2, 7.4.1, 8.1

[48] Einar Haugen. The Ecology of Language. In Anwar S Dil, editor, *Essays by Einar Haugen*. Stanford University Press, Stanford, CA, 1972. 2.2

[49] Brent Hecht and Darren Gergle. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, pages 291–300, New York, New York, USA, April 2010. ACM Press. 1.2.1, 2.4.1

[50] Amir Helzer. Localizing for software, websites and global apps. *Multilingual*, 22(3):34–37, 2011. 1.2.1, 1.2.3

[51] Alfred Hermida. From TV to Twitter: How Ambient News Became Ambient Journalism. *Media/Culture Journal*, 13(2), 2010. URL `http://ssrn.com/paper=1732603`. Last accessed Oct 30, 2013. 1.1

[52] Susan Herring. Web Content Analysis: Expanding the Paradigm. In Jeremy Hunsinger, Lisbeth Klastrup, and Matthew Allen, editors, *International Handbook of Internet Research*, chapter 11, pages 233–249. Springer Verlag, Berlin, 2010. 1.4, 4

[53] Susan Herring, John Paolillo, Irene Ramos-Vielba, Inna Kouper, Elijah Wright, Sharon Stoerger, Lois Scheidt, and Benjamin Clark. Language Networks on LiveJournal. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 79–90. IEEE Computer Society, January 2007. 1, 1.2.1, 2.2, 3.2, 4, 5.4, 7.4.1, 8.1

[54] Courtenay Honeycutt and Susan C. Herring. Beyond Microblogging: Conversation and Collaboration via Twitter. In *Proceedings of the 42nd Annual Hawaii International Conference on System Sciences (HICSS'09)*, pages 1–10. IEEE Computer Society, December 2009. 7, 7.2

[55] Lichan Hong, Gregorio Convertino, and Ed Chi. Language Matters in Twitter: A Large Scale Study. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, volume 91, pages 518–521. AAAI Publications, 2011. 1, 2.5, 3.3, 3.3, 5.4, 7.3, 8.1

[56] Nancy H. Hornberger. Multilingual language policies and the continua of biliteracy: An ecological approach. *Language Policy*, 1(1):27–51, March 2002. 2.2, 2.5

[57] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational Tagging in Twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia - HT '10*, pages 173–178, New York, New York, USA, June 2010. ACM Press. 7.4.2

[58] International Telecommunication Union. ITU Measuring the Information Society. Technical report, Geneva, 2011. URL http://www.itu.int/ITU-D/ict/publications/idi/. Last accessed Oct 30, 2013. 1

[59] Internet Society. Who We Are. URL http://www.internetsociety.org/who-we-are. Last accessed Oct 28, 2013. 1.2

[60] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07*, pages 56–65, New York, New York, USA, August 2007. ACM Press. 2.5, 4.2

[61] Ian Johnson. Audience Design and Communication Accommodation Theory: Use of Twitter by Welsh-English Biliterates. In Elin Haf Gruffydd Jones and Enrique Uribe-Jongbloed, editors, *Social Media and Minority Languages: Convergence and the Creative Industries*, chapter 6, pages 99–118. Multilingual Matters, Bristol, Buffalo, Toronto, 2013. 2.5, 3.1, 7.3

[62] Aravind K. Joshi. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th conference on Computational linguistics* -, volume 1, pages 145–150, Morristown, NJ, USA, July 1982. Association for Computational Linguistics. 2.5, 7.4.2

[63] M Kaiser, M Görner, and C C Hilgetag. Criticality of spreading dynamics in hierarchical cluster networks without inhibition. *New Journal of Physics*, 9 (5):110–110, May 2007. 2.2

[64] Krishna Yeshwanth Kamath and James Caverlee. Transient crowd discovery on the real-time social web. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, pages 585–594, New York, New York, USA, February 2011. ACM Press. 4.6

[65] Helen Kelly Holmes. An Analysis of the Language Repertoires of Students in Higher Education and their Language Choices on the Internet. *International Journal of Multicultural Societies*, 6(1):52–75, 2004. 2.5, 3.1, 7.2, 8.1

[66] Farshad Kooti, Haeryun Yang, Meeyoung Cha, Krishna Gummadi, and Winter Mason. The Emergence of Conventions in Online Social Networks. In *International AAAI Conference on Weblogs and Social Media*, 2012. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4661. Last accessed Oct 30, 2013. 2.5, 7.1, 7.4.2

[67] A. Kralisch and B. Berendt. Language-sensitive search behaviour and the role of domain knowledge. *New Review of Hypermedia and Multimedia*, 11(2): 221–246, December 2005. 1, 3.1

[68] A. Kralisch and T. Mandl. Barriers to Information Access across Languages on the Internet: Network and Language Effects. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 3, page 54b. IEEE Computer Society, January 2006. 1

[69] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web - WWW '10*, pages 591–600, New York, New York, USA, April 2010. ACM Press. 1.1, 1.1

[70] David Laniado and Peter Mika. Making Sense of Twitter. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*, pages 470–485. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. 7.4.2

[71] Nadège Lechevrel. L'écolinguistique : une discipline émergente? *Revue des étudiants en linguistique du Québec - Quebec Student Journal of Linguistics*, 3(1):18–38, 2008. 2.2

[72] Julie Letierce, Alexandre Passant, John Breslin, and Stefan Decker. Understanding how Twitter is used to spread scientific messages. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, Raleigh, NC: US, 2010. URL `http://journal.webscience.org/314/`. Last accessed Oct 30, 2013. 7.4.2

[73] David Lewis, Stephen Curran, Gavin Doherty, Kevin Feeney, Nikiforos Karamanis, Saturnino Luz, and John McAuley. Supporting Flexibility and Awareness in Localisation Workflows. *The International Journal of Localisation*, 8 (1):29–38, 2009. 2.4

[74] Literature Across Frontiers. Publishing Translations in Europe. Trends 1990-2005. Technical report, Mercator Institute for Media, Languages and Culture, 2010. 2.3.2

[75] Gilad Lotan. #OccupyWallStreet: origin and spread visualized — SocialFlow blog, 2011. URL `http://blog.socialflow.com/post/7120244404/occupywallstreet-origin-and-spread-visualized`. Last accessed Oct 30, 2013. 1.2.2

[76] Gilad Lotan. Data Reveals That Occupying Twitter Trending Topics is Harder Than it Looks!, 2011. URL `http://blog.socialflow.com/post/7120244374/data-reveals-that-occupying-twitter-trending-topics-is-harder-than-it-looks`. Last accessed Oct 30, 2013. 1.2.2

[77] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and danah Boyd. The Revolutions Were Tweeted: Information Flows during the

2011 Tunisian and Egyptian Revolutions. *International Journal of Communication*, 5:1375–1405, 2011. 1, 1.1, 1.2.2

[78] Safari Mafu. From the Oral Tradition to the Information Era: The Case of Tanzania. *International Journal of Multicultural Societies*, 6(1):99–124, 2004. 1

[79] Christopher Manning. Logistic Regression (with R), 2007. URL `http://nlp.stanford.edu/~manning/courses/ling289/logistic.pdf`. Last accessed Oct 8, 2013. 6.2

[80] Cameron A. Marlow. *The Structural Determinants of Media Contagion*. Ph.d., Massachusetts Institute of Technology, 2005. 1.2.1, 2.3.1, 2.3.1, 5.1

[81] A. E. Marwick and d. Boyd. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13 (1):114–133, July 2010. 2.5, 3.1, 6.4, 7

[82] Cheryl Metoyer-Duran. *Gatekeepers in Ethnolinguistic Communities*. Information Management, Policy and Services. Ablex Publishing Corporation, Norwood, New Jersey, 1993. 2.3.1, 2.3.2, 2.5, 5.4

[83] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. The Twitter of Babel: mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981, January 2013. URL `http://dx.plos.org/10.1371/journal.pone.0061981`. Last accessed Oct 30, 2013. 2.3.2, 2.3, 2.5, 3.2, 3.3, 3.1, 3.3, 4.2, 4.7, 5.4, 8.1

[84] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, January 2007. 4.4.2

[85] Ory Okolloh. Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action*, 59(1):65–70, 2009. 1

[86] Eli Pariser. *The Filter Bubble: What the Internet is Hiding from You*. The Penguin Press, New York, 2011. 1.2.1

[87] Carol Peters, Martin Braschler, and Paul Clough. *Multilingual Information Retrieval: From Research To Practice*. Springer, 2012. 1.2.1

[88] Isabella Peters. *Folksonomies. Indexing and Retrieval in Web 2.0*. Knowledge and Information. De Gruyter, Berlin, 2009. 2.4.1

[89] Daniel Pimienta, Daniel Prado, and Álvaro Blanco. Twelve years of measuring linguistic diversity in the Internet: balance and perspectives — UNESCO publications for the World Summit on the Information Society. Technical report, United Nations Educational, Scientific and Cultural Organization, Paris, 2009. 1, 2.4

[90] Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. Do all birds tweet the same?: characterizing Twitter around the world. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, pages 1025–1030, New York, New York, USA, October 2011. ACM Press. 2.5, 3.2, 4.2

[91] James E. Prieger. The broadband digital divide and the economic benefits of mobile broadband for rural areas. *Telecommunications Policy*, 37(6):483–502, 2013. 1.2, 2.4.2

[92] Pei-Luen Patrick Rau, Tom Plocher, and Yee-Yin Choong. *Cross-Cultural Design for IT Products and Services*. CRC Press, 2012. 1.2.1

[93] Dana Rotman, Jennifer Preece, Yurong He, and Allison Druin. Extreme ethnography. In *Proceedings of the 2012 iConference*, pages 207–214, New York, New York, USA, February 2012. ACM Press. 4.6

[94] C.E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. 6.1, 6.1

[95] Katie Shilton, Jes A. Koepfler, and Kenneth R. Fleischmann. How to See Values in Social Computing: Methods for Studying Values Dimensions. In *(To appear in) Proceedings of the ACM 2014 conference on Computer Supported Cooperative Work - CSCW '14*. ACM Press, 2014. 1.2.3

[96] David Sims. Understanding place and space in a digital Babel. The nuances of location language, 2012. URL `http://radar.oreilly.com/2012/03/location-unstructured-non-english-health-outbreak.html`. Last accessed Oct 30, 2013. 8.2.1

[97] Richard L. Sites. Language Technology Ecosystem, 2011. URL `http://www.hltd.org/alex.pdf`. Last accessed Oct 30, 2013. 4.3.1

[98] Kate Starbird and Leysia Palen. "voluntweeters": Self-organizing by digital volunteers in times of crisis. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, pages 1071–1080, New York, New York, USA, May . ACM Press. 1, 1.2.2

[99] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of Twitter networks. *Social Networks*, 34(1):73–81, 2012. 3.2, 4.2, 4.3

[100] Steven L. Thorne, Rebecca W. Black, and Julie M. Sykes. Second Language Use, Socialization, and Learning in Internet Interest Communities and Online Gaming. *The Modern Language Journal*, 93:802–821, December 2009. 1, 2.4

[101] Twitter Help Center. Age screening on Twitter. URL `https://support.twitter.com/articles/20169945-age-screening-on-twitter`. Last accessed Oct 7, 2013. 4.9

[102] Claire Ulrich. Technological Developments for African Languages. *Multilingual*, 21(5):51–53, 2010. 1, 2.4

[103] UNESCO. Recommendation Concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace, 2003. URL `http://www.unesco.org/new/en/communication-and-information/about-us/how-we-work/strategy-and-programme/promotion-and-use-of-multilingualism-and-universal-access-to-cyberspace/`. Last accessed Oct 30, 2013. 1.2, 2.4

[104] Federico Vazquez, Xavier Castelló, and Maxi San Miguel. Agent based models of language competition: macroscopic descriptions and order-disorder transitions. *Journal of Statistical Mechanics: Theory and Experiment*, 2010 (04):P04007, 2010. URL `http://iopscience.iop.org/1742-5468/2010/04/P04007/`. Last accessed Oct 30, 2013. 2.3.2

[105] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, pages 1079–1088, New York, New York, USA, April 2010. ACM Press. 1.1

[106] Jessica Vitak, Cliff Lampe, Rebecca Gray, and Nicole B. Ellison. "Why won't you be my Facebook friend?". In *Proceedings of the 2012 iConference*, pages 555–557, New York, New York, USA, February 2012. ACM Press. 3.1

[107] Barney Warf. Geographies of global Internet censorship. *GeoJournal*, 76(1): 1–23, November 2010. 1.2, 2.4.2

[108] Mark Warschauer, Ghada El Said, and Ayman Zohry. Language Choice Online: Globalization and Identity in Egypt. In Brenda Danet and Susan Herring, editors, *The Multilingual Internet: Language, Culture, and Communication Online*, chapter 13. Oxford University Press, New York, 2007. 2.4, 2.5, 3.1, 5.4

[109] Duncan J. Watts. The "New" Science of Networks. *Annual Review of Sociology*, 30:243–270, 2004. 2.3

[110] Wouter Weerkamp, Simon Carter, and Manos Tsagkias. How People use Twitter in Different Languages. In *WebSci Conference 2011*, Koblenz, Germany, June 2011. URL `http://journal.webscience.org/539/2/Table1.png`. Last accessed Oct 30, 2013. 3.3, 7.2, 7.3, 7.4.2

[111] Li Wei. *The Bilingualism Reader*, volume 24. Routledge, London, July 2000. 2.5

[112] Howard T. Welser, Eric Gleave, Danyel Fisher, and Marc Smith. Visualizing the Signatures of Social Roles in Online Discussion Groups. *JoSS: The Journal of Social Structure*, 8(2):1–31, 2007. 5.2

[113] George Weyman. Translating Tweets from the Arab Spring: Towards a Translation Workbench for Twitter, 2012. URL http://meedan.org/2012/03/translation-twitter-middle-east-arabic/. Last accessed Oct 30, 2013. 8.2.1

[114] Leo Widrich. How Twitter evolved from 2006 to 2011, 2011. URL http://blog.bufferapp.com/how-twitter-evolved-from-2006-to-2011. Last accessed Oct 28, 2013. 1.1

[115] World Summit of the Information Society. Building the Information Society: a global challenge in the new Millennium. Declaration of Principles, 2003. URL http://www.itu.int/wsis/basic/about.html. Last accessed Oct 30, 2013. 1.2

[116] Sue Wright. Multilingualism on the Internet - Thematic introduction. *International Journal of Multicultural Societies*, 6(1):5–13, 2004. 1

[117] John Yunker. *Beyond Borders: Web Globalization Strategies*. New Riders, 2002. 1.2.1, 1.2.3

[118] John Yunker. Inside Google's language detection tool - Global by Design, 2010. URL http://www.globalbydesign.com/blog/2010/12/06/inside-googles-language-detection-tool/. Last accessed Oct 30, 2013. 4.3.1

[119] Ethan Zuckerman. CHI keynote: Desperately Seeking Serendipity, 2011. URL http://www.ethanzuckerman.com/blog/2011/05/12/chi-keynote-desperately-seeking-serendipity/. Last accessed Oct 28, 2013. 1.2, 2.4.1, 5.4, 8.1