Towards the evaluation of written proficiency on a collaborative social network for learning languages: Yask An English language study case

Fabio Nelson Silva Penagos

Instituto Caro y Cuervo

Seminario Andrés Bello

Maestría en Lingüística

Bogotá, Colombia

2020

Towards the evaluation of written proficiency on a collaborative social network for learning languages: Yask An English Language case study

Fabio Nelson Silva Penagos

Trabajo de grado presentado como requisito parcial para optar al título de:

Magíster en Lingüística

Director: Sergio Gonzalo Jiménez Vargas

Doctor en Ingeniería de sistemas y computación

Co-director: George Enrique Dueñas Luna

Magíster en Educación

Instituto Caro y Cuervo

Seminario Andrés Bello

Maestría en Lingüística

Bogotá, Colombia

2020

"—No es la primera vez que alude al empobrecimiento del lenguaje —dijo Etienne—. Podría citar varios momentos en que los personajes desconfían de sí mismos en la medida en que se sienten como dibujados por su pensamiento y su discurso, y temen que el dibujo sea engañoso. Honneur des hommes, Saint Langage... Estamos lejos de eso."

Rayuela

Julio Cortázar

Dedicated to...

Myself, the one whose life has molded just like wind or water does with the rock, that seen as hard, would be as soft as cotton in the deep ocean of time.

I shall return to the soil as humble as I arrived!

Hamish, how I miss you! Luigi and Tito who have been there silently for me, gazing at me with purity and innocence.
F.N.S

CARTA DE AUTORIZACIÓN DE LOS AUTORES PARA LA CONSULTA Y PUBLICACIÓN ELECTRÓNICA DEL TEXTO COMPLETO

Bogotá, D.C., 9 de noviembre de 2020

Señores
BIBLIOTECA JOSÉ MANUEL RIVAS SACCONI
Cuidad

Estimados Señores:

Yo **FABIO NELSON SILVA PENAGOS**, identificado con C.C. No. **80140956**, autor del trabajo de grado titulado **TOWARDS THE EVALUATION OF WRITTEN PROFICIENCY ON A COLLABORATIVE SOCIAL NETWORK FOR LEARNING LANGUAGES: YASK – AN ENGLISH LANGUAGE CASE STUDY** presentado en el año de 2020 como requisito para optar el título de MAGISTER EN LINGÜÍSTICA; autorizo a la Biblioteca José Manuel Rivas Sacconi del Instituto Caro y Cuervo para que con fines académicos:

- Ponga el contenido de este trabajo a disposición de los usuarios en la biblioteca digital Palabra, así como en redes de información del país y del exterior, con las cuales tenga convenio el Seminario Andrés Bello y el Instituto Caro Y Cuervo.

- Permita la consulta a los usuarios interesados en el contenido de este trabajo, para todos los usos que tengan finalidad académica, ya sea formato impreso, CD-ROM o digital desde Internet.

- Muestre al mundo la producción intelectual de los egresados de las Maestrías del Instituto Caro y Cuervo.

- Todos los usos, que tengan finalidad académica; de manera especial la divulgación a través de redes de información académica.

De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, ***"Los derechos morales sobre el trabajo son propiedad de los autores"***, los cuales son irrenunciables, imprescriptibles, inembargables e inalienables. Atendiendo lo anterior, siempre que se consulte la obra, mediante cita bibliográfica se debe dar crédito al trabajo y a su (s) autor (es).

Firma y documento de identidad

DESCRIPCIÓN TRABAJO DE GRADO

AUTOR

Apellidos	Nombres
Silva Penagos	Fabio Nelson

DIRECTOR (ES)

Apellidos	Nombres
Jiménez Vargas Dueñas Luna	Sergio Gonzalo George Enrique

TRABAJO PARA OPTAR POR EL TÍTULO DE: Magister en Lingüística

TÍTULO DEL TRABAJO: Towards the evaluation of written proficiency on a collaborative social network for learning languages: YASK

SUBTÍTULO DEL TRABAJO: An English Language case study

NOMBRE DEL PROGRAMA ACADÉMICO: Maestría en Lingüística

CIUDAD: ____BOGOTA_____ AÑO DE PRESENTACIÓN DEL TRABAJO: 2020

NÚMERO DE PÁGINAS: 160

TIPO DE ILUSTRACIONES: Ilustraciones _X_ Mapas ___ Retratos ___ Tablas, gráficos y diagramas _x__ Planos ___ Láminas ___ Fotografías __

MATERIAL ANEXO (Vídeo, audio, multimedia):

Duración del audiovisual: _____ Minutos.

Número de casetes de vídeo: _____ Formato: ___ Mini DV ___ DV Cam ___ DVC Pro ___ Vídeo 8 ____

Hi 8 ____ Otro. Cual? _____

Sistema: Americano NTSC _____ Europeo PAL _____ SECAM _____

Número de casetes de audio: _____

Número de archivos dentro del CD (En caso de incluirse un CD-ROM diferente al trabajo de grado: _____

PREMIO O DISTINCIÓN (En caso de ser Laureadas o tener una mención especial): Tesis Meritoria.

DESCRIPTORES O PALABRAS CLAVES: Son los términos que definen los temas que identifican el contenido. *(En caso de duda para designar estos descriptores, se recomienda consultar a la dirección de biblioteca en el correo electrónico biblioteca@caroycuervo.gov.co):*

ESPAÑOL	**INGLES**
Suficiencia lingüística	Languaqe proficiency
Redes sociales	Social networks
Comunidades de habla	Speech communities
Trabajo cooperativo/colaborativo	Cooperative/collaborative work
Ambientes de aprendizaje	Learning environments
YASK	YASK

RESUMEN DEL CONTENIDO Español (máximo 250 palabras):

YASK es una red social colaborativa en línea que permite practicar idiomas dentro de una dinámica que incluye solicitudes, respuestas y votos. Partiendo del hecho que la medición de la competencia lingüística mediante el uso de diversas metodologías es muy difícil, costosa y en muchas ocasiones imprecisa. Presentamos una nueva alternativa metodológica denominada ProficiencyRank, usando como contexto la aplicación YASK. Nuestro método, amplia el reconocido algoritmo PageRank al incluir información correspondiente a los votos positivos y negativos. Se identificaron cuatro tipos de señales dentro del grafo de interacción social. Entre ellos están los acuerdos y desacuerdos entre usuarios, los cuales permiten calcular clasificaciones para la mayoría de los usuarios en la red, superando así las limitaciones intrínsecas de PageRank. Nuestros experimentos muestran que la reputación de los usuarios en la aplicación YASK, medida con ProficiencyRank, está significativamente correlacionada con la suficiencia en el idioma aprendido, mientras que la producción escrita está correlacionada débilmente con los perfiles de vocabulario del Marco Común Europeo de Referencia. Además, descubrimos que los votos negativos son considerablemente más informativos que los positivos, los votos explícitos son más informativos que los implícitos, y una combinación ponderada de todas las señales produce los mejores resultados. Concluimos que el uso de esta tecnología es una herramienta prometedora para medir la suficiencia en L2, incluso para grupos relativamente pequeños de estudiantes, y potencialmente aplicable a otras redes sociales colaborativas.

RESUMEN DEL CONTENIDO Inglés (máximo 250 palabras):

YASK is an online collaborative social network for practicing languages in a framework that includes requests, answers, and votes. Since measuring linguistic competence using current approaches is difficult, expensive, and in many cases imprecise, we present a new alternative

called ProficiencyRank using YASK as context. Our method extends the well-known PageRank algorithm by allowing positive/negative votes, and explicit/implicit information. We identified four types of implicit signals in the social graph from agreements and disagreements between users allowing the computation of rankings for the majority of users in the network overcoming the intrinsic limitations of PageRank. Our experiments showed that the reputation of the users in YASK measured by ProficiencyRank is significantly correlated with their language proficiency, while their written production was poorly correlated with the vocabulary profiles of the Common European Framework of Reference. In addition, we found that negative votes are considerably more informative than positive ones, explicit votes are more informative than implicit ones, and a weighted combination of all signals produce the best results. We concluded that the use of this technology is a promising tool for measuring L2 proficiency, even for relatively small groups of learners and potentially applicable to other collaborative social networks.

Agradecimientos

La vida debería ser es en sí misma un acto de agradecimiento, ¡qué tan fácil dejamos de ver el milagro de los dones y las posibilidades recibidas!

Volviendo la mirada, descubro que está maestría en Lingüística fue un regalo, una oportunidad, llena de bendiciones, que como signos en lenguas perdidas fueron apareciendo día tras día.

¡Tan difícil fue entrar al instituto, como mantenerme, pero más difícil fue entregar este documento, este minúsculo legado que representa una visión del mundo en sí mismo.

Gracias Sergio, un profesor, un amigo, un ser humano incondicional. Gracias George por su desinteresado apoyo.

Gracias a mis maestros, a aquellos que vieron bondad en mi corazón, a aquellos que les resulte molesto y vano.

Gracias a los funcionarios del instituto, solo me puedo llevar el calor de su amable trato...

Gracias a mis compañeros de cohorte, a todos y a todas, aprendí de ustedes, les deseo una vida próspera.

Gracias a mi familia, por su apoyo moral y su fe en mis decisiones.

Abstract

YASK is an online collaborative social network for practicing languages in a framework that includes requests, answers, and votes. Since measuring linguistic competence using current approaches is difficult, expensive, and in many cases imprecise, we present a new alternative called ProficiencyRank using YASK as context. Our method extends the well-known PageRank algorithm by allowing positive/negative votes, and explicit/implicit information. We identified four types of implicit signals in the social graph from agreements and disagreements between users allowing the computation of rankings for the majority of users in the network overcoming the intrinsic limitations of PageRank. Our experiments showed that the reputation of the users in YASK measured by ProficiencyRank is significantly correlated with their language proficiency, while their written production was poorly correlated with the vocabulary profiles of the Common European Framework of Reference. In addition, we found that negative votes are considerably more informative than positive ones, explicit votes are more informative than implicit ones, and a weighted combination of all signals produce the best results. We concluded that the use of this technology is a promising tool for measuring L2 proficiency, even for relatively small groups of learners and potentially applicable to other collaborative social networks.

Keywords: Language proficiency measurement, Second language learning, Reputation in social networks, Cooperative/collaborative learning, Data science applications in education, Distributed learning environments, Evaluation methodologies, Learning communities.

Resumen

YASK es una red social colaborativa en línea que permite practicar idiomas dentro de una dinámica que incluye solicitudes, respuestas y votos. Partiendo del hecho que la medición de la competencia lingüísitica mediante el uso de diversas metodologías es muy difícil, costosa y en muchas ocasiones imprecisa. Presentamos una nueva alternativa metodológica denominada ProficiencyRank, usando como contexto la aplicación YASK. Nuestro método amplia el reconocido algorítmo PageRank al incluir información correspondiente a los votos positivos y negativos. Se identificaron cuatro tipos de señales dentro del grafo de interaccion social. Entres ellos están los acuerdos y desacuerdos entre usuarios, los cuales permiten calcular clasificaciones para la mayoría de los usuarios en la red, superando así las limitaciones intrínsecas de PageRank. Nuestros experimentos muestran que la reputación de los usuarios en la aplicación Yask, medida con ProficiencyRank, está significativamente correlacionada con la proficiencia en el idioma aprendido, mientras que la producción escrita está correlacionada débilmente con los perfiles de vocabulario del Marco Común Europeo de Referencia. Además, descubrimos que los votos negativos son considerablemente más informativos que los positivos, los votos explícitos son más informativos que los implícitos, y una combinación ponderada de todas las señales produce los mejores resultados. Concluimos que el uso de esta tecnología es una herramienta prometedora para medir la suficiencia en L2, incluso para grupos relativamente pequeños de estudiantes, y potencialmente aplicable a otras redes sociales colaborativas.

Palabras claves: Medición de la suficiencia lingüística, Aprendizaje de una segunda lengua, Reputación en redes sociales,Aprendizaje cooperativo/colaborativo, Ciencia de datos en educación, Ambientes de aprendizaje distribuidos, Metodologías de evaluación, Comunidades de aprendizaje.

Contents

List of Figures

List of Tables

1 Introduction

Quantitative evaluation of a phenomenon consists of measuring one or several variables associated to it while controlling other intervening variables that alter the measure, but that are not linked to the phenomenon, e.g. noise (Black, 1999). That unavoidable situation produces differences between what is wanted to be measured and what is actually being measured. The accuracy of a particular measurement method depends significantly on whether the variables used are effectively aimed at the target and on the robustness of the method against unwanted or unavoidable factors.

The evaluation for educational purposes also obeys that principle. That is, the tools used to measure a particular skill or knowledge (for example, an exam or a written test) sometimes point to a "moving target" and are usually affected by external factors. For instance, the tests for second language proficiency assessment can deviate from its intrinsic objective if they only take into account what is taught in the teaching curriculum and discard diverse cultural and linguistic backgrounds (Sandberg and Reschly, 2011). Also, they are affected by factors such as the artificial preparation of the individuals being evaluated (González Moncada, 2009; Menken, 2006), the test takers' ability to discriminate between plausible options, the handling of a particular set of keywords (Matthiesen, 2017), the test time conditions (Knoch and Elder, 2010), the misalignment with the language used by the population (Gu and So, 2015), among others. In addition, most tests are usually based on prescriptive curricula that become inconvenient to evaluate other learning approaches such as informal learning (Jurkovič, 2019) and deductive data-driven learning (Godwin-Jones, 2017; Lee and Lin, 2019). Consequently, many tests of linguistic competence actually measure many factors that may

or may not be related to their actual linguistic competence.

Meanwhile, the current information era and the rise of social networks provide new approaches for quantitative evaluation based on the principle of the "wisdom of the crowd" (Golub and Jackson, 2010). Consider the case of StackOverflow[2], a social network where computer programmers ask questions that are collaboratively answered by the online community. Traditionally, a programmer's degree of technical competence is determined by written, oral or automated tests, which suffer from many of the aforementioned problems in the language-proficiency domain (Douce et al., 2005). A recent study (Movshovitz-Attias et al., 2013) showed that the reputation gained from the social interactions on StackOverflow is an accurate predictor of the programming skills of the users of that social network.

Although most of the research has been dedicated to the identification of experts (Balog et al., 2012; Lin et al., 2017), the identification of high-intermediate degrees of expertise has also been addressed (Pal et al., 2011), and also the entire spectrum (Zhang et al., 2007). Measuring the reputation of the users fulfills the double objective of motivating self-learning of the users and that of providing opportunities for personal promotion in the real world. Although it is difficult to compare this type of evaluation with the traditional concepts of formative and summative evaluation, reputation acquisition could be associated with the objectives of facilitating learning and providing means to access jobs, business opportunities and academic positions (Hall and Graham, 2004). Although the evaluation based on social reputation has been receiving acceptance in different domains, as far as our knowledge goes this has not been attempted in the academic domain for student or learner evaluation.

Recently, YASK[3], a new collaborative social network for learning and practicing languages (Spanish, English, German, French, Italian, Portuguese, Russian, and Haitian Creole), has been gaining popularity, recognition, and an increasing number of active users (Chile, 2018). YASK has a similar structure to StackOverflow, opening the research perspective of measu-

[2] `https://stackoverflow.com/`

[3] `https://www.yask.ai`

ring the written proficiency in L2 of the users based on their interactions in the social network.

The methods for measuring the user importance/reputation in a social network are based on the analysis of the structure of the social graph. A well-known method for that is the PageRank algorithm (Page et al., 1999). Our method, called ProficiencyRank, extends PageRank by integrating positive, negative, implicit, and explicit signals to the social graph. In this work, we are focused on determining if ProficiencyRank is an appropriate approach for measuring the language competence of a group of users in a social network like Yask.

Similarly to other online knowledge-sharing communities, Yask provides its users with answers that would be related to a formative assessment, assessment for learning, and peer feedback. This would be because a language doubt posted by a user would receive answers, comments, and votes, which could fulfill the function of feedback. Most likely, some of these comments answer the doubt posted by the user. Thus, the user can read them and think critically about the knowledge that has just been acquired. In this way, the level of appropriation of the topic by the user is shown and, at the same time, it guarantees that the user manages the topic discussed before addressing another one, which is broadly known as formative assessment (Black and Wiliam, 2009). Our approach recombines the ensemble of votes of the entire community to provide a quantitative measure of the degree of language expertise for each user, which could be assimilated as a summative evaluation in the sense that it provides a global measure relative to the entire community of each user's progress.

1.1. Summary of contributions

Given the emerging importance of collaborative, distributed, mass, and distance learning, automated educational assessment methods become a necessity. Furthermore, some skills such as proficiency in a second language are difficult to assess due to the lack of consensus on the objective to be measured. That is, should learners be assessed against teaching curricula or against native speakers? Are these curricula representative of real communicative environments?

This study addresses these two issues (the necessity and the difficulty) by adapting social networking analysis technologies to a learning environment. However, current methods are mainly aimed at identifying minorities of authorities and experts among users, ignoring other degrees of expertise. In this scenario the contributions of this study are the following:

1. A method for ranking users on a collaborative social network from novices to experts, even for users who only vote and do not contribute by posting requests or responses. Such method is independent of the domain, language, or modality of the users' posts, as it only recombines voting pools and does not make any use of information from the users' content.

2. The use of an alternative scope for curricula and assessment, The online interaction-based approach.

3. An application and evaluation of the method in the natural L2 learning domain that demonstrates its effectiveness and convenience compared to curriculum-based methods.

1.2. Dissertation outline

The rest of this thesis is structured as follows. In Chapter 2 the topics required for understanding the following chapters are briefly presented. Given the interdisciplinary nature of this dissertation, that chapter provides the necessary concepts from the fields of education, linguistics, and computing to understand its content. In Chapter 3 the current approaches for automatic assessment of written proficiency are reviewed (section 3.1), along with the current approaches aimed to measure expertise in social networks (section 3.2). In addition, in section 3.3, the explicit links between the reviewed state of the art and this dissertation are briefly explained. In Chapter 4, the research problem outlined in this introduction is formally justified, stated, and developed. In Chapter 5 the *ProficiencyRank* method (based on social media) along with the *CERF Baseline* method (based on textual content) are explained in detail. In Chapter 6, our experimental validation is presented including: experimental setup,

results, and discussion. Finally, in Chapter 7, the final remarks and conclusions that can be drawn from our investigation are presented.

2 Background

The evolution of language is dynamic and constant, sometimes is evident, and sometimes hides underneath historical, cultural, and political facts. Nowadays technology has made perceptible these processes and it has dared individuals to update constantly to fulfill requirements for connecting to others faster and easier. In terms of interaction, people communicate using internet-based tools, using languages according to geographical location mostly. Nevertheless, the English language has imposed its dominance over the trade markets and has served as an acculturation vehicle to spread the Anglo-western civilization (Xue and Zuo, 2013). The prestigious worldwide information tends to be published in English, forcing people to be in contact with it. Entertainment industries, commerce, and academic fields evolve using the English language as a lingua franca.

From a variationist perspective, English enters the speech communities around the world affecting inland languages at different aspects such as phonetic, lexical, and semantic. For instance, in Colombia where the official language is Spanish and where there are over 65 indigenous languages spoken by small communities along the country, students must study a foreign language (English) following a continental trend in education policies. Beyond that in academic fields and university contexts, premium information and publishing requirements involve English language proficiency. English is a contact language in Latin American countries as Spanish is a contact language in countries like the USA and Canada due to immigration phenomenon.

In this thesis, we addressed the problem of determining the degree of English proficiency in

learners from a perspective based on social media interaction. Since this is a rather novel approach[4], it is necessary to support our hypothesis using selected concepts and theories from the fields of linguistics, education, and computing. The subjects from these fields are presented with an intermediate depth, therefore, readers with different background education should skip some sections. The relevant topics from the linguistics perspective are covered from sections 2.1 to 2.3. The educational topics are focused on the second language (L2) teaching, acquisition, and evaluation (sections 2.4 to 2.9. Finally, in section 2.10 the concepts needed to support the computational methods proposed in this thesis are briefly explained.

2.1. Linguistic social networks

Studies on social networks analyze the properties and structures of relationships happening in the individuals' interactions (Milroy and Llamas, 2013). In language studies, researchers have focused on the maintenance of nonstandard and minority languages by some specific social groups (Lippi et al., 1997).

From sociolinguistics perspectives, social networks are perceived as environments to capture the speaker's underlying dynamics or "speaker variables" (Eckert, 2000). In anthropological and sociological studies, the aim deals with policies and supporting population segments with the economical, sanitary, educational, and political deficit (Cochran et al., 1993; Johnson, 1994). The general assumption in the genesis of communities indicates that individuals create their communities (gathering selected members) to provide a meaningful framework for solving daily life problems (Mitchell, 1986).

The term *ego* in Milroy and Llamas (2013) is conceived as the central participant or "anchor" in any network. In highly dense and many-stranded networks multiple ego's ties are linked to each other, which means that these kinds of networks have the capacity to support their members in practical and symbolic ways. On the contrary, the influence on networks could

[4]Language related practices could be considered as the "original interdisciplinary knowledge, due to its inherent quality of explaining everything, just like maths does

be also negative when imposing unwanted and stressful constraints on their members.

The social network model can be represented as webs of ties reaching out through the whole society. Ties can be direct (first-order network ties) or indirect (second-order network ties) among participants. Milroy and Llamas (2013) also clarified ties subdivision as strong ties and weak ties in the consequent interaction with family members and friends and the distant connections to acquaintances. Milardo (1988) separated exchange from interactive networks. Exchange networks, where members and whom ego interact frequently and exchange direct support, advice, aid, and criticism. In interactive networks despite the frequent interaction among members, the ego does not rely on the material or symbolic resources.

Wei (1994) included Milardo's proposal an extra category "passive tie" when there is no regular contact and absence of material support. It is found in the immigration context where the ego receives moral and social influence from relatives and friends from the distance.

2.1.1. The community of practice

Eckert (2000) employed the community of practice to locate interactional social sites where meanings are conventionalized and constructed after linguistic factors of change and variation take place. A simple definition of community of practice is the collectivity of people sharing and mobilizing towards a common enterprise.

Linguistic norm (as a whole) and linguistic styles are placed in the interactional site within the close-knit networks. The members of a social network are immersed in the particular type of relationships, conventionalized and accepted, inside their community of practice. To some extend, without external influence, they could not be aware of different norms apart from their one. Linguistic influence takes place without an explicit sign or comment. The adoption of a way of speaking as many other cultural expressions requires in the community both access and entitlement, as part of the group identity (Eckert, 2000).

Immersive methodologies

The traditional methodology to access to language communities inside social networks is the ethnographically-oriented data-collection procedure. The researcher enters the social network with member permission, observing and interacting directly and randomly. The usual findings of this approach allow the researcher to collect spontaneous speech samples and relevant social, cultural, and demographic information. Data collected serves to compare new information to the previously collected one.

In the literature, most of the authors have reviewed some common findings like i) long-established communities are minimally impacted by social or geographical mobility; ii) close-knit network communities use vernacular variants as an identity sign of integration; iii) gender mark positively use of variant forms among members of a community (Chambers, 1995; Cohen, 1982; Docherty et al., 1997; Milroy and Milroy, 1993; Milroy, 1999; Young and Wilmott, 2013).

2.1.2. Measuring social network structures

Each research process supposes a definition of relevant features on which information would be analyzed, thus in the case reviewed by Milroy (1987) from a study held in Belfast, Ireland; the chosen indicators were membership in a neighbor group, kinship ties with at least two households, a same working place with at least two neighborhood inhabitants, a same working place with at least two same-gender inhabitants from the neighborhood, and voluntary-leisure association with workmates. The study findings show correlations between personal network structures and phonological variation, affected by age and gender (Milroy and Llamas, 2013).

Applicability of studies in social networks

Milroy and Llamas (2013) described the preference of network approach in linguistic va-

riationist studies, due to the multiple advantages like the possibility of selecting whether smaller or bigger groups of informants for data recollection, the intrinsically features of some social networks regarding ethnic groups and minorities, rural population, immigrants and non-industrialized societies. Perhaps, the best advantage deals with closer participation of informants, allowing linguistics to elucidate variants driving language variation and change.

Bilingual communities

It has been proved after many studies, that the social mechanism of language maintenance and shift in bilingual communities are influenced by general principles in network ties constitutions. Thus strong ties maintain linguistic variety standstill, despite external influences, and only if these networks ties weaken is probably the linguistic shift to happen.

In the case of traditional working-class immigrants like Italian American urban villager of New York have transformed their initial rural close-knit networks in their Italian soil to urban close-knit networks in the ghettos in the USA. Additionally, newer immigrants have kept this network transition for generations (Bourhis and Marshall, 1999; Dabène and Moore, 1995; Gan, 1962; Giddens and Sutton, 1989).

Situational contexts also influence speakers to use different codes and languages (Zentella, 1997) describes the shift in linguistic code among Puerto Rican people in the USA. First, age influence preferences, the elder (all genres) use Puerto Rican Spanish while "young dudes" favor African American Vernacular English (Labov, 1972), although they also access to other varieties of Spanish and English. Children who speak very little Spanish tend to mix it into English and produce of the Nuyorican Code.

2.1.3. Language variation on weak ties

Social networks have been subsequently studied to identify and measure circumstances and phenomena producing language changes. The prototypical social network observed in most

of the previously published researches included strong ties communities, where speakers tend to preserve specific features through time. In fact, according to Chambers (1995), most of sociolinguistic is generally oriented to non-mobile speakers living in isolated or marginal areas. On the contrary, social networks regarding loose-knit ties are more susceptible to allow language changes, but ironically this type has been undervalued for researching. From a traditional perspective in humanities or the minorities protection, loose-knit or weak ties are not relevant at all; nonetheless, in linguistic revision or when analyzing glottopolitical conditions associated with linguistic imperialism, weak ties are quite relevant, indicating speakers' motivations or constraints.

2.2. Speech community

Patrick (2001) drew the influence of speech community (*SpCom*) as a widespread theoretical concept, used by many sholars dealing with language from various perspectives. The aim of studies would be diverse (grammar, syntax, phonetics, or discourse analysis), but linguistic research always includes the inherent fact of language development socially in communities. Commonly language change, whether geographical or social, refers to *SpCom* as a boundary of urban and rural, large, and small areas. *SpCom* as the label of different kinds of minorities or linguistic subgroups inside communities. Even though *SpCom* also applies to studies about children and gender language.

Theoretical proposals

Traditionally, *SpCom* is the concept relating people, culture, and language into a singular entity. In the study of aboriginal peoples in isolated communities, anthropology based, at very early stages, in structuralism sources (Herder and Scheibe, 1949; Humboldt, 1988).

In 20th century authors like Sapir (1921) defined speech as historical heritages of human groups, then language, culture, and race are not necessarily correlated unless in a historical form.

Bloomfield (1922) and Bloomfield (1933) identified speech to utterance, and those utterances run within certain communities. Thus, *SpCom* is a relative-value concept, varying in size and overlapping with other communities and normative heterogeneity. *SpCom* represents a human group interacting through speech.

Lyons (1970) shared Bloomfield's definition and included the term dialect. Gumperz (1962) stayed apart from the anthropologist's perspective and concentrated his interest in code-switching highlighting the relation between *SpCom* and multilingual settings. The notion of "bilingual speech community" introduced by Weinreich (1953) joins structural and functional approaches. A community may be either monolingual or multilingual according to the specific characteristics (Gumperz and Hernandez-Chavez, 1972).

Patrick (2001) featured the importance of code (and code-switching) in the theoretical proposal of *SpCom* due to the inherent communicative matrix weaved by speakers. If a multilingual community suddenly skips one language code many situations, objects, and parts of the life contexts would have restrictions to be expressed meaningfully. Also, Gumperz and Hernandez-Chavez (1972) indicated that *SpCom* could cover small zones (face-to-face among speakers) or larger regions, depending level of abstraction driving the analysis, thus social cohesion is optional.

Gumperz (1996) came out with the idea that *SpCom* is a complex body composed of the similarities among used codes and their shared meaning across the social group. The understanding of the conventionalized features is possible only for *SpCom* members. Patrick (2001) separated grammatical competence from the organization and interpretation in sociolinguistic norms. There are common extends among Gumperz, Labov, and Hymes, dealing with the functionalist character of *SpCom*. Members of such a community differ from each other in certain beliefs and behaviors, if such conditions are analyzed in individuals separately, irregularities appear. Nevertheless, in a holistic revision, systematic regularities reveal

themselves.

Hymes et al. (1974) declared thus *SpCom* defines itself by the concurrence of rules of grammar and rules of use. This affirmation is possible in an environment where the ethnographer can identify verbal repertoires, classify speech events, and rules in communicative situations. These organized tasks allow ethnographers to describe the communicative competence of speakers exploiting language resources inside the *SpCom*. The linguistic knowledge (speaking norms) are unequally distributed within communities. How much knowledge is required to distinguish participants (random users) from members (usual users)? (Dorian, 1982).

Hymes et al. (1974) pointed out difficulties about proper notions of community and membership, besides his approach emphasized shared norms over interaction, that is why ways of speaking imply various types of knowledge of form, constructions, coherence among them, but also their social distribution and function. Patrick (2001) indicated a shift in the focus for *SpCom* from linguistic production to community-based.

A disruptive practice-based proposal came from on-campus research addressed by Labov (1966) in New York City. His proposal focused on language structure and change by developing specific sets of questions and answers to test and describe. Labov's work characterized by the wide range of methods used and the demonstrability of uniformity and normative sociolinguistic structures. Labov (1989) highlighted linguistic uniformity as the main evidence for *SpCom* membership. Despite criticism, this conception is not an outcome but an interpretative practice resembling previous researching conditions from New York city *SpCom*. Ideally, the rest of the on-campus observations would show similar features like the ones described for Labov's New York City study. Under this conception and methodology, the research process reveals the *SpCom* which is not evident per se. A *SpCom* is not an assumption, does not deal with theory. On the contrary, it is a matter of observation.

In the New York City (NYC) Language case, Labov primarily revised social conditions of re-

sidence and stratification relations among neighborhoods, including validity items like social and ethnic representative inclusion, immigrant groups influence, social mobility and loyalty, and typical structure of the residence. Secondly, after obtaining fine samples, Labov analyzed and compared peculiar phonological aspects. Labov twisted the aim of his investigation by not pursuing the dialect divisions, instead, he focused on uniformity across the whole collectivity. He concluded that "NYC is a single speech community united by a set of evaluative norms, though divergent in the application of these norms" (Labov, 1966). Finally, Labov (1972) concluded that a *SpCom* is basically "defined by participation in a set of shared norms, which uniformity of abstract patterns of variation is invariant respect to particular levels of usage".

The speech community according to the model of society

In philosophy and sociology, the study of the inner evolution of societies has produced several different models and theories regarding social, economic, or political features (Marx, 2015; Weber, 2002).

The sociolinguistics uses these models due to class struggle influence the whole human dimensions including language. Language varieties are classified, according to relation to the standard, legitimized, and literate form of language (Kerswill et al., 1994).

The use of historical approaches allows scholars to understand events, causes, and consequences of human dynamics affecting the use of codes and languages. Despite the given historical conditions, linguist research must select carefully the analytical choices of a research question and methodology, criteria of legitimacy depend directly from the *SpCom* anatomy, legitimacy criteria depend directly from the *SpCom* anatomy.

One of the prime goals in sociolinguistic deals with the question about the choice in the use of certain code or language over others. Is it a free act? or is it a constraint and mandatory

reaction? Answers depend on the research aim.

During the 1990s, opinions about *SpCom* were divided. Hudson (1996) and Wardhaugh (1998) proposed language as a subjective and individual concept, dependant from the community only to allow the speaker to identify oneself with others. A cultural identity sort of speak. Duranti (1997) went beyond recommending to move *SpCom* from being an object of inquiry to be "the product of the communicative activities engaged in by a given group of people". Patrick (2001) considers these previous suggestions inadequate. Instead, he presented the historical development of the *SpCom* concept from the 1950s, 1960s, and 1970s used in prestigious and important research papers (Ferguson, 1959; Hymes et al., 1974; Saville-Troike, 1982; Stewart, 1962; Weinreich, 1953). came with an innovative notion of simultaneous membership in multiple overlapping *SpCom* (Patrick, 2001).

The linguistic analysis can be divided into five levels of abstraction in speakers: individual, network, social group, speech community, and general language (Romaine, 1982). Considering different communicational contexts like individual speakers, dyads, multi-party face to face interaction, communities of practice, and large communities, Hank (1996) concludes that no metalanguage suits all levels of communicative interactions. Participants' intersections occur only in face-to-face contexts and some large-scale discursive formations.

2.3. Language acquisition

Language acquisition has been one of the most researched items in linguistics' history. Different models dealing with diverse disciplines and approaches have been proposed over the years. Some of the traditional models linked language acquisition to behaviors resulting in environmental stimuli over children (Skinner, 1957). Language acquisition as an innate mental ability from human nurture developing simultaneously to physiological development (Chomsky, 2006) and Piaget's proposed cognitive-staged development in children that included a linguistic phase (Piaget, 1977).

In recent years, scholars have presented many theories holding innovative paradigms. For practical purposes in this specific work, only two methods will be reviewed, the competition model and the usage-based Theory.

2.3.1. The competition model

The competition model (CM) is a model proposed by Bates and MacWhinney (1982). The CM focuses on language variation studied from psycholinguistics. The CM has been applied to different study aims of language such as acquisition mechanisms, comprehension, production, and impaired language processes. It also has been useful in the analysis made on different languages around the world. The most important theoretical construct is the cue or an information source that allows the user to successfully link form with the meaning (Li and MacWhinney, 2012). *Cues* are divided into types according to levels of linguistic study (morphological, syntactic, prosodic, semantic, and pragmatic). *Cues* also have an aspect of availability dealing with how often they are present in a language, and reliability or the percentage in which they lead to a correct interpretation. *Cues* could obtain a third feature validity if they have availability and reliability in a language.

The unified competition model

Learning a second language requires the individual interaction with norms and rules for the second language (L2) but also requires that learners contrast them to the Mother language (L1). There are some models regarding similarities and differences between using different paradigms in the entrenchment of acquisition processes (Firth and Wagner, 1998; Friederici, 2009; Zhao et al., 2008), whereas the unified competition model (UCM) focuses on the dependency of L2 in L1, which brings some risk factors like negative transfer, social isolation, parasitism and incorrect connections between processing ideas (Li and MacWhinney, 2012).

2.3.2. The usague-based theory

Tomasello (2009) points out the two main principles of The usage-based theory (UBT) "meaning is use" relating to the semantic function of linguistic communication and "structure emerges from use" representing the structural dimension of linguistic communication. The UBT extends to explain first language acquisition adding a couple of cognitive skills: Intention reading and pattern finding to the principles above. Intention reading is a functional dimension that refers to what children must do to distinguish intentions from mature speakers when they use linguistic conventions. Linguistic conventions are transmitted from older to younger culturally. Pattern finding is a grammatical dimension is what children must do creatively, passing beyond the mere reproduction of utterances, it means new situational constructions (Tomasello, 2009).

First language acquisition is divided into stages, for each stage, there are some specific features dealing with children's age, communicative tool, effectiveness in the message. The stages are:

Prelinguistic communication

In this stage, infants have not acquired the linguistic conventions yet. It has been proved by scholars (Goldin-Meadow, 2009) that infants have alternative and sophisticated manners to communicate, for instance, children point directions to objects they want and also use some generalized gestures representing needs. Tomasello (2009) affirms that these features of prelinguistic communication embody forms of social cognition and communicative motivation that are unique to the species, stressing it as the initial feature of linguistic convention both phylogenetically and ontogenetically.

Utterances and words

The manner children acquire language is, as said before, by using communication tools or linguistic convention elements. Tomasello (2009) indicates that children use the utterance –the smallest unit capable to express a complete communicative intention- because it is the

most accessible item of linguistic communication. At a certain point in the physiological development, children comprehend utterances, and more importantly, they understand words' senses, their functional aspect, then words can be used creatively.

Schemas and constructions

Children's language development under the UBT is not an isolated phenomenon, but a social construction that takes place at communicative linguistic situations. Children produce different types of utterances, some are quite more sophisticated than others. There are utterances with intonational contours expressing communicative motives. The usual result from this kind of interaction is highly concrete linguistic schemas or constructions, based mostly on particular words and phrases. On the other hand, there are abstract linguistic constructions linked to idioms. Tomasello (2009) claims that utterance-level constructions underlie multi-word utterances.

There are three types of utterance level constructions: word combination, pivot schemas, and item-based constructions. Utterances used by children in these early stages of communication, depend directly from usage, restricting the option to analyze those utterances without taking into consideration their intention. In the example "toy sofa" a child refers to its location after throwing it away. Grammar revision can drive to a quite different conclusion instead. Maybe a syntactic mistake or a missing word. There are cases when children multiword production shows more systematic patterns. Usually there a unit that acts as a pivot schema, due to the possibility to be with many other words. A typical example "no ____; no soup, no shot, no cold."

Tomasello (2009) indicates item-based construction as the most advanced structure for this children's development stage. An item-based construction is a kind marked linguistic figure including aspects like the canonical order of a language, transitivity mode, etc. Children who perform this type of constructions are capable of use transitive verbs –previously taught- using a diverse set of words and in distinct communicative environments. As partial con-

clusions of these stages can be listed as follows: i) the acquisition of a language follows a chronological order and well-structured patterns, ii) the utterance is the initial unit of cognitive process, and iii) patterns and functions can be observed and extracted only form usage situations.

Abstract constructions

The abstract constructions required high training and ability, this stage starts at the age of three approximately. Tomasello (2009) enlists the abstract constructions into groups: i) identificational, attributives, and possessives, ii) simple transitives and intransitives, iii) datives, ditransitives, and benefactives, iv) collocatives, resultatives, and causatives, v) passives and reflexives, vi) imperatives and questions.

Gentner and Markman (1997) indicate as a goal for observation of the acquisition processes in children, focusing on constructional patterns conventionally associated with semantic content (Tomasello, 2009). Children must see when a specific utterance is produced to fits a particular linguistic pattern. The linguistic discrimination for the utterance required in children's abilities of schematization and analogy, used along for other cognitive activities.

Theoretical critics for UBT

Among objections critical detractors mention three main incongruent items: i) UBT can not deal with more complex constructions, such as those regarding multiple verbs and syntactic embedding, ii) UBT fails explaining how generalization and abstractions processes are to be constrained (Tomasello, 2009) and iii) the manifested poverty of the stimulus.

The claims for UBT defenders point that in fact "simple constructions" in the early stages of language acquisition may not be so different in comparison to complex constructions (Diessel, 2004). As an illustration of this declaration, the authors mention the prototype of

utterance with a sentential function e.g. “I know you tell her” and “I think I can do it” (Diessel and Tomasello, 2001).

In children between three to five years old, it is quite common to find utterances with sentential complements. Besides, there is an evident use of multiple-categorized types of verbs from epistemic (think and know) used in affirmative, negatives, and exclamatory variety of sentences; to attention-getting verbs (look and see) applied to imperatives.

The second objection relating to the constraining in generalization and abstraction processes is explained by Tomasello (2009), who clarifies that constraining influence takes place mostly in early stages, due to the lack of generalization and abstraction processes in children. As individuals advance in the acquisition of elements and the interaction grows stronger, they are suited to use a wide range of functions and appropriate application situations for words such as verbs. Construction constraining is less used and in fact, restricted to particular circumstances.

About the poverty of stimulus, Tomasello (2009) highlights two separate theoretical perspectives, on one hand, Chomsky’s innate universal grammar that holds the idea of children behavioristic learning regarding blind associations and inductive inferences with no conceptual understanding of linguistic function at all. On the other hand, the UBT can not *“conceived as set of formal, algebraic rules but as a structured inventory of meaningful grammatical constructions, with the child possessing sophisticated learning skills involving categorization, analogy and distributional learning”* (Tomasello, 2009). The acquisition of constructions is determined by some aspects like the frequency, consistency, and complexity (Lieven and Tomasello, 2008).

2.4. Second language acquisition

The usage-based (UB) is a term that labels a whole set of different approaches to second language acquisition (L2A), Wulff and Ellis (2018) mark the common assumptions to each

approach as i) the primary source for learners of a second language (L2) is the linguistic input they receive, and ii) L2 learners employ all kind of learning mechanism available in their language learning acquisition.

2.4.1. L2 learning constructions

Learning language comprises various tasks but mainly to learn conventionalized constructions (form-function mappings) to express meanings and intentions inside speech communities (Wulff and Ellis, 2018).

According to Goldberg (2006) language constructions range from morphemes to words and phrases. Morphemes are suffixes such as *less*, indicating lack of an object or characteristic e.g. homeless, endless, useless, or tasteless. Words such as 'bill' with multiple meanings like *twenty-dollar bill, pay the bill or bill of rights*.

Constructions have several levels of complexity and abstraction, therefore they are stored in multiple forms, thus the word bill and the morpheme plural *s*. Both are simple constructions, probably stored independently one from another. Also, they are constituents of a complex construction: *bills*.

In this respect, Wulff and Ellis (2018) point out the levels of constructional abstraction or schematization present in lexical formulas. There are two main types of lexical formulas known as fully lexicalized formula (thank you) and partial schematized slot-and frame patterns like [Good + (time of the day)] rendering lexicalized greetings schemas like *Good morning* and *Good night*. Many other constructions can be obtained from this formula sample.

This widely encompassing definition in Wulff and Ellis (2018) is a manner in which division between lexicon and grammar, words, and rules fade away. Consequently, a sentence is a product of combining several constructions rather than a product of applying rules for

word order. Learning a second language requires learning associations within and among constructions.

2.4.2. L2 learning processing

The children's perceptual system gradually includes and attunes these entries in the input. Ellis (2008b) enlist several psycholinguistic construction-related and learner-related factors in the L2 learning process. Factors for construction, significance of meaning, redundancy vs. surprise value, prototypicality, frequency of experience, contingency of form, and function seem to matter. The factors for the learner are, attention, transfer, automaticity, overshadowing, blocking play crucial roles. According to Ellis (2008a), these psycholinguistic factors conspire in the acquisition and use any linguistic construction. The L2's learner early stages constructions usually are those with higher frequency, e.g head is acquired earlier than kidney. The learner's perceptual system gradually includes and attunes to these entries in the input.

Exemplar-based rational contingency analysis

The first time a construction comes to an L2 learner's mind is a unitary representation that binds all its properties at once (i.e phonological, spelling, etc.). The construction activates as its properties would be present in the language environment serving as a pattern-recognition unit (Wulff and Ellis, 2018). After this initial approach to construction by the usage environments, learners build a memory representation, and gradually alters and adapts it to the accumulative experiences, properties, and contexts in which construction develops. Constructions serve each other as a comparative item in encounters and also produce prototype construction (more typical ones). Prototypes are considered as main items for categorization due to the similarity or difference to other constructions' prototypes.

Learners do not have statistics about the frequency of constructions in cognitive or linguistic contexts, perhaps they are aware of more common ones in usage settings. Statistical

knowledge happens unconsciously (Ellis, 1994; Rebuschat, 2015).

2.4.3. Constructicon

The form-function mappings are connected in a network –called constructicon- collecting forms and meanings. The constructicon is a system that depends on each individual on the particular form-interpretation associations built by speakers during the lifespan. This system performs at any time of language development, is custom-tailored, highly adaptatively and quite precise in showing learners own linguistic experience, these special features can support the idea that language learning is rational and is located in the rational cognitive studies of human psychology due to the goal of adapt behavior to the environmental conditions (Anderson, 1989).

Considering language as a complex-adaptative system because it involves many agents or people, many different configurations (social and cultural roles), operating at different levels (brain processes, body expressions, phonemes, constructions, social interaction contexts, and discourses) and on multiple scales (biological features, social-interactional and historical) (Ellis et al., 2016; MacWhinney and O'Grady, 2015).

Lexical and grammatical constructions L1 and L2

Construction learning is forced by the frequency in the use of specific linguistic items and functionalities. Not all constructions are learnable at the same rhythm and speed by all learners. Most of the learner's attention focuses on word classes than on grammatical cues. In some cases, learners hardly evolve from a lexical focused knowledge to a grammar-system knowledge. This stage is known as "Basic Variety" applicable for L2 learners whose command of L2 is quite less sophisticated than L1 ability (Klein, 1998).

Learnability of a construction is defined by three main factors, salience, contingency of form-function association, and learned attention.

Wulff and Ellis (2018) emphasized the idea that despite blocking limitations on L2, there are no qualitative differences in learning mechanisms between L1 and L2. L1 requires learners to focus attention on particular language environments in a natural way, on the contrary for L2 must reconfigure learning attention after the environments have changed.

Implication for Language Teaching

L2 learners must learn how to adjust their attention biases from L1 to be able of acquiring constructions formation items in L2. The basic assumption of the implicit acquisition process implies the learning of complex data without a selective attention on the matters learned. On the contrary, L2A implies for instance an explicit learning process and then an obvious attention on information items (Wulff and Ellis, 2018).

Schmidt's (2001) noticing hypothesis argues that attention to linguistic form is a precondition to learn. In traditional learning processes, teachers or tutors demand from their students' attention and synthesis from the information treated during classes and sessions. A quite common limitation to learning appears when students have to use information or knowledge in practical activities, but they do not know how to proceed, because they did not attend instructions. In language learning attention means input. A teachers' must-do is to use strategies to address students' attention on linguistic forms, make the acquisition process more efficient.

2.5. Educational Framework

There are two basic forms of learning English, on one hand, an immersion that is traveling to an English-speaking country and learns the language by interacting with or without tutoring. On the other hand, learning English in a non-English-speaking country, by taking a course, learn structures and content and practice in the classroom. Access to one or another depends on funds. This dichotomy draws two labels applied to the features in the acquisition

of a non-native language.

Approach dychotomy

English as a Second language (ESL) occurs in English-speaking countries, where the real practice is available everywhere, producing meaningful knowledge and the use of structures according to specific communicative situations.

English as a Foreign Language (EFL) results from learning that takes place in non-English-speaking countries, where language practice is restricted to the classroom or class-related activities. Communicative situations usually are simulated, thus the accuracy of appropriate expression according to the communicative situation is hard to obtain by learners.

In this thesis paper, ESL and EFL labels are replaced by a unique EL2 (indicating the English language as the one use after the native language, no matter form, and spatial environment) avoiding any further misunderstanding and focusing on the English language specifically.

2.5.1. English Language Education Development

The English Language is an educational factor that holds most of the institutions of society through countries nowadays, its development has been a result of multiple combining historical facts, the environment of application and politic and economic decisions.

Historical Overview

The English language is the lingua franca, used internationally for trading, politics, international affairs, industry, technology, and education following a wave of globalization and informatization addressing communication worldwide (Xue and Zuo, 2013). As language is the core of culture, no culture possibly exists without its language like historical struggles in America‘s conquest process has demonstrated (Jacobs, 2006). Xue and Zuo (2013) high-

lighted emphatically that "culture is the reflection of politics and economy and ideology. [...] important means for maintaining and developing countries".

The English language dominance in the modern era started as a result of political and economic factors of the British Empire colonization. Later with the independence of the United States of America in 1783, the English language served to White House's imperialist enterprises. The sacred duty of enlightening those unknown and savage regions in the middle of jungles, deserts, and seas. Many scholars and authors have supported the idea of the European supremacy, a clear example is found in Rudyard Kipling's The White Man's Burden, a poem exhorting Americans to assume political control of the Philippines Islands. Currently, English-speaking countries are having unequal relations with poorer non-speaking countries, due to the cultural hegemony hold by economic, military, and political factors. Besides, the English-speaking countries promote vigorously expansion of the English language as an apparent vehicle of progress, hiding the negative influence of it as a weapon violating third-world countries' national identities.

Influence of English in Media

Certain aspects have to be reviewed to truly measure the influence of the English language, i) number of users, ii) fields of application, iii) Learning policies.

English number of users

The approximate number of native speakers of the English language is around 380 million people. As a second or foreign language, the number is about 280 million people worldwide. According to some official institution like The British Council, about 1 billion people is currently learning English, and at least 2 billion are in contact with it (Xue and Zuo, 2013).

English fields of application

After World War Two (WW2) Britain and the United States declared themselves as the victorious party after defeating Nazi Germany. They took advantage by monopolizing political influence in institutions like United Nations; industry such as automobile, machinery, mining and agriculture tools; science, especially in applied chemistry and physics; trading developing branding and product suppliers; education, proposing models and researching and; communication implementing mass media, entertainment and written publication indexing in English. The communication worldwide usually uses the English language, since broadcasters such as BBC and NBC started the mass media content production in the 1950s. Technology has improved velocity and connections means by using networks such as the internet, initially created and operated by the US Army, uses mainly English –in coding for instance-.

English learning policies

The British government agency for English education and cultural relation known as The British Council was founded in 1934, in a moment of profound social, commercial, economical and political instability. The British council's initial mission was 'to create in a country overseas a basis of friendly knowledge and understanding of the people of this country, of their philosophy and way of life, which will lead to a sympathetic appreciation of British foreign policy, whatever for the moment that policy may be and from whatever political conviction it may spring.' The mission of the British Council was to "friendly promote" the use of English as a foreign language in non-English-speaking countries. In Colombia for instance, the British Council was established in 1940. Behind the British Council, there is some renowned academic institution supporting the use of the English language, especially the University of Cambridge. UCLES (University of Cambridge Local Examinations Syndicate) was the first Cambridge official exam (1913) initially to measure standards of school education in the transition from secondary to the tertiary level of education (Secretariat, 1998).

The United States of America has had some initiatives of cultural exchange programs such as the Fulbright scholar program, established under legislation by Senator J. William Fulbright in 1946. The offering includes several types of scholarships for American and foreign graduate and undergraduate students, and institutions. The program includes certified studies, teaching and research residences, and lectures funds. In terms of English for educative purposes, the most representative is ETS, a non-profit private organization devoted to education and research through testing. ETS develops different kinds of exams and tests regarding academics, business, and education such as TOEFL. A 16-members board of trustees, representing levels and areas of interest in education and business governs ETS.

Along with the USA and Britain Australia is the third global power in the English language. Australia considers education as an economic income industry contributing $37.6 billion (AUD) to the Australian economy in 2018/19. English language courses contribute an estimated 2.4 billion annually. Students who choose Australia to study due to the outstanding quality programs, innovative environments to study and practice, and high-standard protective policies for international students. The Australian government encourages universities and educational institutions to ensure admission requirements and support students to meet the Higher Education Standards Framework (Threshold Standards) 2015 and the National Code of Practice for Providers of Education and Training to Overseas Students 2018. Institutions must verify previous academic enroll admission requirements and levels of English language proficiency. Australian government proposes English Language Teaching International Engagement Strategy 2025 to outline official support to the English language teaching sector.

2.5.2. Traditional and Technological EL2 Models

Traditional EL2 model

The traditional EL2 teaching approach is face-to-face interaction, among participants in learning contexts. By this extend, traditional classrooms, labs, and auditoriums, imply student

and teacher co-presence and co-assistance. The didactics for the lessons can vary depending on participants' needs. The main aim of traditional methodologies is knowledge (content or skills) efficient synthesis and acquisition. Initially, for this paper, the traditional EL2 methods are only those developed without computer technology assistance. Most of these method proposals were introduced between the 1960s and the 1980s and have survived until today due to their relevance. Larsen-Freeman reviewed some of the most widespread ESL methods (Asher, 1969).

Table **2-1** enlists the methods with their corresponding principles. Most of the approaches gravitate around the role of students, emphasizing their human dimension, including aspects like self-esteem, personal opinions, expressive skills, and well-being affecting the construction of knowledge through direct interaction. Teachers facilitate conditions for students to learn, such as comfortable environments, availability of appropriate activities, students' performance monitoring, and personal support in public as well in private.

Traditional EL2 teaching follows some protocols and stages in the planning. The selection of materials, resources, and strategies depend on the structure of the lesson. Usually, lessons are divided into sections or moments:

- Greetings: The teacher welcomes students to the lesson. Short dialogue.
- Warm-up: The teacher sets a quite short activity reviewing previous content related to the current lesson.
- Class settings: The teacher indicates the topic, instructions of the class, and lesson methodology.
- Initial question: The teacher sets some questions or goals for the lesson to answer or fulfill.
- New information: The teacher introduces new information about the lesson's topic.

METHOD	PRINCIPLES
Audio-Lingual Method	1. Learning is a habit formation.
	2. Early error correction by teacher to students.
Author:	3. Sentence pattern overlearning by students.
Nelson Brooks (1964)	4. Positive reinforcement to provoke correct habits.
Community Language Learning	1. Students are whole persons.
	2. Self-security improves learning.
Author:	3. The students must have choices to produe language they prefer.
Charles Curran (1972)	4. Teacher must try to understand Student's feelings.
Comprehension approach /	1. Meaning in the target language can be conveyed through actions.
Total Physical Response	2. Retention enhanced when learners respond physically.
Author:	3. Feelings of success and low anxiety facilitate learning.
James Asher (1969)	4. Listening comprehension comes first, speaking come as soon as speakers are ready.
Suggestopedia	1. Learning is facilitated by confortable environments.
	2. Confident students learn better and easier.
Author:	3. Students unity enhance learning.
Georgi Lozanov (1978)	4. Didactics and materials must be activated by playful activities.
Communicative Approach	1. Language teaching should enable students to use language to communicate.
	2. Language is used in social contexts. It should be appropiated to the communicative settings.
Authors:	3. Teaching should provide students the chance for themselves to understand meaning.
Dell Hymes and	4. Students should be able to epress their ideas, questions and opinions.
Michael Halliday (1985) Silent Way	1. Teaching should be subordinate to learning.
	2. Practice should drive students to develop their own ïnner criteria"for correctness.
Author:	3. Errors are essential in learning process.
Caleb Gattengo (1963)	4. Practice should focus on students, not on teachers.

Table 2-1: L2 Teaching methods with their corresponding principles

- Practice: Students develop activities related to the topic and to reinforce new information.
- Socialization and correction: The teacher asks students to participate in socializing the findings of the practice. If there are errors, or doubts, students and the teacher answer.
- Conclusion: The teacher highlights the outstanding information from the lesson, and assign homework.

In terms of quality, traditional EL2 lessons should include four skills. Therefore, planning has to manage different types of activities to ensure proficient standards. The EL2 lessons

take place mostly in classrooms (Tsui, 2001) where teachers can address lesson pace, by interacting directly with students, performing control over the learning environment and conditions of language acquisition.

Assessment during the class deal with teachers monitoring and student's perceptions about activities, through dynamic of active questioning and error correction. In collaborative classroom contexts, assessment comes from different participants, the teacher, the partner student, and oneself student (Peacock, 2017).

Classroom healthy practices ensure continuous feedback among teachers and students. Roles should switch to empower students to conceal language acquisition through creative pathways. The active participative learning approach.

Independent Students in traditional language learning settings

The traditional approach to English language learning also comprises independent individuals studying on their own. The use of books is common. Two types of resources have been widely used, printed course-books (Soars and Soars, 2001) and grammar books (Raymond, 2015).

Oral interaction obliges students to find speakers to practice, while listening is supplied by audio samples. The own pacing study is an opportunity to focus on written aspects of the language.

Technological Model EL2

The computer technology (computer-assistance language learning CALL) incepts a breakthrough in EL2 teaching and learning expanding environments, tools, resources, and time to interact using the English language in non-native contexts.

By the late 1980s, computers became popular at homes and in schools as mandatory gadgets. First, they replaced typing machines -text interfaces-, after the video and audio interfaces (Komlodi et al., 2006).

By the late 1990s, computer science constituted a common school subject worldwide. Basic computer skills became essential at different levels (Hahnel et al., 2016).

Initially, educational agents, found in computers, an auxiliary resource for students to work more efficiently (Grant et al., 2009). For instance, complex mathematical calculus using computers took a fraction of the time in comparison to the manual processes.

At that time, teachers continued the use of traditional methodologies, based on classroom interactions, and computers were just the subsidiary tool to complete tasks. Sooner with internet-based innovations, many linguistic skills were involved in practice. The internet allows individuals to access information from multiple sources, including books, exercises, audio samples, videos, recording software among others (Manovich, 2009).

By the 2000s, online education was widely established (Dede et al., 2018) using the internet and online platforms, emulating scholar structures of academic courses, subjects, syllabus, and class content formats. Teachers and students interacted, no physically but virtually, through emails, chat rooms, and blogs (Ros Martínez de Lahidalga, 2008). The internet-based education processes required the use of new pedagogical approaches as well as a whole new set of didactics strategies. Nevertheless, as Compton (2009) mentioned online education literature has scarcely promoted publications carrying out specific language teaching under this extend.

Essentially, work for online education should converge on successive aspects just as teachers' roles in the 21st century regarding technology-related teaching issues, important for students' engagement and motivation. Among these issues are enlisted the software proficiency,

internet basics enrollment, permanent search of new resources and tools (Chapelle and Hegelheimer, 2004) Apart from the technical knowledge, Bennett and Marsh (2002) identify two relational skills: i) identification of relevant differences and similarities between face-to-face and online teaching contexts; ii) identification of strategies and techniques to facilitate online learning and allow students to exploit advantages of independent and collaborative learning. To a certain extend, online teaching manages skills that go from general levels to very specific micro-levels, for instance, the skills pyramid proposed by Hampel and Stickler (2005).

The skills pyramid encompasses seven levels introducing in the first level the basic ICT competence arranging the minimum knowledge about computer technologies. The second level suggests specific technical competence for the software specifically designed to teach languages. Teachers should know the basic platform maintenance and a wide range of associated software programs and applications, whether free or under license permission.

The third level, dealing with constraints and possibilities of the medium, requires teachers to understand limitations and restrictions for every single software, due to the purpose of the creation or from external conditions like advertising policies for the free products.

The fourth level is about online socialization, where every kind of interaction is possible from the socially established community (Palloff and Pratt, 1999).

From the teachers' perspective, socialization should be flexible according to the level of students and their confidence towards class interaction. Besides, socialization addresses students to maturate inherent linguistics abilities.

The fifth level, facilitating competence, informs about teachers' guide role. Ideally, through online education, students interact and fulfill planned tasks to improve accumulatively expressing, rote, and linguistic skills Hampel and Stickler (2005).

highlighted social cohesion as a prime factor for assertive communication.

The sixth scale in the pyramid is creativity and choice. Shallowly, creativity expresses through the different types of resources and activities delivered by teachers, the proposed use, and learning objectives. Moreover, creativity and encouraging environments drive students to critical and propositive thinking.

The most up, the seventh level refers to own style teachers enhance to be. Honest and creative teachers investigate and encourage students positively to go further. Disparities among teachers are possible and needed to develop an identity that ensures and attract class dynamics towards effective learning.

Compton's framework for online language teaching skills

Compton (2009) criticized Hampel and Stickler (2005) skill pyramid, in the sense that technological and pedagogical skills can not be placed one after another in a sequence, those skills are concurrent and simultaneous during the whole process, they are present in different roles. Furthermore, Compton improves skill pyramid proposal, introducing a framework for online language skills (see Table **??**) considering three basic aspects of online language teaching: technology, pedagogy, and evaluation. Teachers are classified as well, conforming to expertise: novice, proficient, and expert.

Novice teachers have elementary technical abilities allowing them to use and identify software, discerning their possibilities and limitations. In terms of pedagogy, Compton features at least a content or theoretical knowledge of didactic strategies, interactives phenomenon, curriculum, and assessment. For the evaluative aspect, novice teachers at least must know basics on task and course evaluation.

The second group is compound by the proficient teachers, who have an upper level of expertise. In the technological aspect, these teachers can select appropriate software tools for

the learning tasks, from a set of possible options, considering advantages and disadvantages. Proficient teacher top ability deals with the build of web pages. In the pedagogic dimension, proficient teachers exceed the knowledge of novice teachers in terms of identify didactic strategies, interactives phenomenon, curriculum, and assessment. Proficient teachers create conditions for students learning processes, modifying abilities accordingly. Following tendencies, evaluation for proficient teachers should be done after the observation of previous pattern frameworks.

The third group is the advanced teachers, whose performance features creativity in the usage and building of resources pursuing pedagogic objectives. They integrate and combine technical abilities for an intuitive development of task and course evaluation (Compton, 2009).

Independent student in technological and online settings

Moore and Kearsley (1996) identified responsibilities for stakeholders in distant education systems. Students self direct their learning process autonomously. Online language learning is a type of distance learning which provides tools, resources, and platforms for learning content and practicing the language through communication.

Traditional language learning focuses on content, expressed in vocabulary, syntax structures, grammar rules, and reading-writing tips. Online language teaching aims to converge on interaction as a vehicle to acquire skills, allowing people to communicate effectively (Kukulska-Hulme and Shield, 2008).

Online resources available on the internet, usually are created and posted under strict organization parameters, which trace simultaneously to the school curriculums (Mccloskey et al., 2008).

Language learning mobile applications

The growing industry of communication through smartphones and tablets has produced in manufactures and software developers the building of multiple-use and themes kind of applications (APPs) suitable for those gadgets' operative systems. Most popular APPs are related to games and entertainment topics. Nowadays, educational APPs are acquiring importance gradually, considering aspects such as accessibility, affordability (free or very low cost), flexibility (use everywhere), and intuitive interface (Rosell-Aguilar, 2017).

From a pedagogic perspective, language learning APPs extend the influence of L2 beyond the classroom. Learning evolves from a duty to a challenge. Features from the APP industry, have innovated towards high-quality multimedia interfaces, that easily attract users (Basri et al., 2019). The gamification of language learning APPs has numerous advantages profitable in real-scenario practices (Rachels and Rockinson-Szapkiw, 2018).

As mentioned in previous studies (Castaneda and Cho, 2016; Kim et al., 2013; Lys, 2013; Yıldız and Ozger, 2012) among positives effects on language, can be listed vocabulary acquisition, phonological awareness, listening comprehension skills, and verb conjugation knowledge after using an app. Also, users acquire positive attitudes towards APPs for language learning. Language learning (LL) APPs are divided into two types: Full LL package APPs contain a high variety of resources, grammatical explanations, and interaction among native and non-native users. Duolingo is a clear example of this kind of APP.

Taxonomy models for Learning APPs

Rosell-Aguilar (2017) presented a taxonomy for APPs (especially for the educational and learning-associated ones) to provide, after Bloom's ideas (1956) "a common language of reference" to define educational and research scopes. In his model Rosell-Aguilar (2017), divided APPs for language learning into three groups: the first refers to the APPs designed specifically for language learning, the second presents dictionaries and thesaurus, and the third shows those APPs designed for purposes different from language learning.

Language learning (LL) APPs are divided into two types: Full LL package APPs contain a high variety of resources, grammatical explanations, and interaction among native and non-native users. Duolingo is a clear example of this kind of APP. Separate LL skill APPs offer specif skill exercises, thus users only use those APPs for improving that area isolated. an example of this type of APP is BookBox, a tool focused on listening.

As alluded above, some APPs designed for different purposes from language learning, actually help users to learn and practice linguistic skills. In this category, the diversity of APPs is quite high, as follows:

- Web browsers, include LL APPs search.
- Multilingual text input like grammar and autocorrect features.
- Speech-to-text tools, generally including pronunciation and spelling checker.
- Communication tools such as email, instant messaging services, and videoconferencing.
- Photo and video camera software.
- Any other application using language with a communicative purpose.

In his model, Rosell-Aguilar (2017), included dictionaries and translator APPs that alternatively can be independent APPs or be embodied within other APPs, e.g.: Google Translator.

2.6. CEFR and Curriculum

Origin and evolution of the CEFR

The Council of Europe is a multilateral organization created by the European Cultural Convention on December 19th, 1954. The Council aimed to pursue a policy of common action designed to safeguard and encourage the development of European culture. To fulfill this

mission, efforts would target language, history, and civilization of singular countries inclusive of the continent as a whole.

Historically in the 1960s, European countries started programs for language communication, with learner-centered approaches. The objective was to provide citizens with opportunities to learn other languages and to reinforce the mother language. During the decades of the 1970s and 1980s was developed the Threshold Level or specifications for language learning attaining personal communication. The cognitive perspective from the Threshold Level stated that linguistic competence depends not only on linguistic knowledge. The Council proposed five dimension of communicative competence: linguistic, sociolinguistic, discourse, socio-cultural, and social competence[5].

By the 1990s the Council of Europe considered the development of a framework for language learning. In 1991 the idea was announced during a continental symposium Transparency and coherence in language learning in Europe. Objectives, evaluation, certification (Rüschlikon, Switzerland, 10-16 November). The first draft was presented in 1995, followed by a consultation made by experts along with European universities in 1996. Between 1997 and 1998, the second draft, its corrections, and piloting were conducted. Last revisions and pre-publishing by 2000. The CEFR document was launched in 2001 and updated in 2018.

The Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) published in 2001 is available in 40 languages, and its one of the Council of Europe's best known and most used policy instruments (COUNCIL, 2018).

The aims of the CEFR

The CEFR is a six-points scale international standard for the description of language ability. The scale classifies from A1 for beginners, up to C2 for mastered-language users. The do-

[5] `https://www.coe.int/en/web/common-european-framework-reference-languages/history`

cument develops a descriptive scheme of language proficiency, including for the scale levels, illustrative descriptors, and curriculum designing options, under intercultural and plurilingual education promotion.

International institutions often use the CEFR to provide transparent reference points in education programs and enterprises. Furthermore, it has been used as a solid background for informing curriculum reform and pedagogy (COUNCIL, 2018). The stated aims of CEFR are to:

- promote and facilitate institutional cooperation among countries.
- provide a sound basis for mutual recognition of language qualifications.
- help language learning and teaching participants (learners, teachers, course designers, examining bodies, and educational administrators) to work under the same understanding (COUNCIL, 2018).

2.6.1. The CEFR's action-oriented approach

The CEFR explicitly announced the nature of its approach, which is action-oriented, meaning that language users are the active participants in the learning process driven by communication purposes following task complexions. The CEFR breakthrough from traditional syllabuses is based on not considering linear language progressions. Instead, the CEFR pursuits syllabuses based on user's need analysis, oriented toward real-life situations constructed encompass selected notions and functions.

Descriptors have the form "Can do" indicating the actual abilities and skills of the users, as well as pointing out those skills users have not acquired yet. The CEFR is a basic tool assisting planning curricula, courses, and examinations, working backward in defining those abilities, skills, and content users or learners should have and know.

The implementation of the action-oriented approach for the CEFR does not intent to be mandatory, on the contrary, it tends to be neutral. Pedagogic operators are free to choose the best strategies to teach and assess according to their particular learning conditions. Although, CEFR does propose to understand learners as social agents, whose needs and petitions should be taken into consideration for planning at all levels.

In methodology implementation of teaching and assessment, the CEFR considers language learning should develop within real-life situations, accomplishing diverse-natures tasks. The concordance between teaching and assessment environments, strategies, and actions is called criterion-reference.

Plurilingual and pluricultural competence

The CEFR distinguishes between multilingualism and plurilingualism, the first is the coexistence of different languages at the social or individual level, the second is the dynamic and developing linguistic repertoire of an individual user or learner (COUNCIL, 2018).

Plurilingualism is an irregular competence, that could restrict learner's learning process because the nature of resources and strategies in one language or variety could differ in other languages. According to the CEFR (COUNCIL, 2018), plurilingual competence is a flexible ability allowing users/learners to :

- switch easily from one language to another.
- express orally in one language, comprehend listen in another language.
- use skills or knowledge from various languages to make sense of a text.
- recognize common international words e.g. trademarks.
- mediate with individuals from other languages.
- employ paralinguistics like gestures, mime, facial expression, etc .

2.6.2. The CEFR descriptive scheme

The CEFR is an organized standardization document, that provides a common descriptive metalanguage to talk about language proficiency (COUNCIL, 2018). The CEFR is divided into competences, which can be general competences, such as knowledge of the world, socio-cultural competence, intercultural competence, and professional experience. Communicative language competences are linguistic, sociolinguistic, and pragmatic competences. Moreover, for the real-life classroom or other learning contexts, two additional competences are present: Communicative language activities and communicative language strategies (COUNCIL, 2018).

The aim of CEFR considers inadequate the old-fashion model of language learning skills regarding speaking, listening, reading, and writing since it can not capture the complexity of real communication (see Figure **??**). Conversely, the CEFR emphasizes real-life context and proposes a model established on interaction in which meaning is co-constructed. Thus, the four forms of communication descriptors for the CEFR are reception, production, interaction, and mediation.

Communication descriptors

Reception

Receiving and processing input activate in users building representations schemata of meaning being expressed or pragmatic hidden intentions. Reception is constituted by three auxiliary types:

- "In aural reception (one-way listening) activities, the language user receives and processes a spoken input produced by one or more speakers."
- "In visual reception (reading) activities the user receives and processes as input written texts produced by one or more writers."

- "In audio-visual reception, for which one scale (watching TV and film) is provided, the user watches TV, video, or a film and uses multimedia, with or without subtitles and voiceovers."(COUNCIL, 2018)

Interaction

Interaction involves discourse and meaning co-construction, usually by two parties. Spoken interaction is the most common, as it back-traced in communication genesis due to collaborative, interpersonal, and transactional functions (COUNCIL, 2018). In the learning dimension, CEFR reviews interaction reflects in strategies such as turn-taking, cooperating, and asking for clarification. Writing interaction explain phenomena including modern chat APPs massively used (e.g.whatsapp), these APPs show speaking styles in written form, this is what CEFR calls online interaction.

Online interaction

The CEFR exemplifies online interaction as Online conversation discussion and Goal-oriented online transactions collaboration COUNCIL (2018).

Mediation

Mediation in CEFR is an innovation for the communication models included in the 2018 paper update. Mediation is a communication factor occurring when a user, usually unconsciously acts as an intermediary between interlocutors unable to communicate one to another, it usually happens when speakers from different languages assemble. When users meditate activities, two basic kinds of mediation can take place oral or written (North and Piccardo, 2016).

Oral mediation

Oral mediation comprises three types of interpretation, simultaneous, consecutive, and informal. The differences are given by types of interactions, formality, speech format, speaking turns, etc (North and Piccardo 2016).

Written mediation

Written mediation is also divided into categories, relating text types, style, genre, text length among other features. The kinds of written mediation are exact translation, literary translation, summarising gist, paraphrasing (North and Piccardo, 2016).

Under her proposal Piccardo (2012), identified four types of mediation as follow:

Linguistic Mediation

Linguistic mediation is a non-restrictive dimension comprises at least two different processes as interlinguistic mediation listing knowledge about interpretation and translation formally and informally, or transforming texts into others. The other type is intralinguistic mediation constitutes arrangements develop inside a language (L1 or L2 separately), for instance summarizing texts, change lexical elements, etc. Complementary contributions locate linguistic mediation in multilingual collaborative settings such as classrooms and workplaces (Creese and Blackledge, 2010; King and Chetty, 2014; Lewis et al., 2012).

Cultural Mediation

As mentioned above, language and culture are complementary and subsidiary to each other. whether linguistic mediation implies translation or interpretation, results must preserve the cultural load of words in terms of inner language-cultural idiolects and sociolects (Brown, 2007; Gohard-Radenkovic et al., 2004; Lévy and Zarate, 2003; Neuner and Byram, 2003).

Social Mediation

Linguistic and cultural practices can be considered as results from social interactions (Piccardo, 2012). When users and learners are immersed in multilingual contexts, situations beyond the meaning interpretation or translation from external cultures could happen as well. Zarate (2003) classified social mediation into three possible groups mediation for introducing new partners into a specific context, mediation to solve conflicts, and mediation installing "third areas" to prevent cultural and linguistic confrontation. The "thridness" concept developed by Kramsch (1993) is a well-spread critical idea relevant for several disciplines like literature criticism, foreign languages, and semiotics. The "thridness" describes hypothetical abandonment from roles -dominant and dominated- and the construction a third place where there is equality between individuals.

Pedagogic mediation

Education systems around the world differ in their own pedagogic practices (Alexander, 2008). Nevertheless, most of them present a continuum feature in the teacher-centered approaches accompanied by collaborative methodologies and strategies (Mercer and Hodgkinson, 2008). Classroom interaction time establishes certain relationships and factors such as facilitating and encouraging people to develop their critical thinking knowledge, collaboratively construction of meaning, and building environments for creativity.

2.6.3. Written production

The CEFR illustrative descriptor scales (COUNCIL, 2018) consider production as a major unit involving two extensive types of production: spoken production (including sustained monologues that describe and give information, putting a case, announce in public, and address audiences) and written production (comprising creative witting and written reports and essays).

textbfCreative writing

This specific kind of written production covers a wide variety of text types, with a personal, imaginative, and expressive language style. Some key concepts (COUNCIL, 2018) may include the following aspects:

- Descriptions based on multiple aspects such as everyday situations, diverse fields of interest, engaging stories, and experience.
- Any type of text is permitted, for instance, diaries extracts, biographies, poetry pieces, well-structured descriptions, and other imaginative written pieces.
- Discourse complexity level, creative writing comprises simple words, expressions and phrases, clear-connected texts, established conventions texts, and smoothly flowing texts.
- Language use contains elementary vocabulary and simple sentences, to very personal style appropriate to both the genre adopted and the reader.

Written reports and essays

More formal types of transactional and evaluative writing include reports and essays. The CEFR (COUNCIL, 2018) descriptor scale includes some attached aspects :

- content: range from factual familiar and routine information of interest to complex topics (academic and professional contexts), distinguishing one's own viewpoints from those in the sources;
- type of texts: comprehend from short content reports and posters to convoluted texts presenting cases, or giving a critical appreciation of proposals or literary works;
- the complexity of discourse: from sentence connectors simple linking to effective logical structure accompanied by consisted expositions.

2.6.4. Written interaction

Interaction in written form has changed from an initial extend developed in 2001, from interpersonal nature to information transfer. Interactive writing uses a type of language that is similar to the spoken one. Some errors and confusion are permitted and contextually supported. Through interaction, strategies are possible to ask for clarification, for help, and to clarify misunderstanding (COUNCIL, 2018). Carefully structure texts are not a requirement in this type of interaction.

Correspondence

In CEFR paper from 2001, only personal correspondence was taken into account and scaled. In CEFR 2018's update, formal correspondence was included since it is a useful activity carried out by many users/learners. Two key concepts (COUNCIL, 2018) address scale analysis:

- type of message: range from simple and personal correspondence to in-depth, sophisticated, personal, and professional correspondence.
- type of language: from emotional linked expression to appropriate expression tone and style aiming text type.

Notes, messages and forms

The transactional set of interactive writing under this category envelops, filling forms with personal information, taking or leaving short messages or notes. The CEFR scale (COUNCIL, 2018) includes some key concepts:

- Pre-A1 to A2 : Characterized by filling forms with personal information
- Leaving and taking telephone messages (simple content messages, e.g., numbers)
- Writing notes (from short and simple to more complex ones)

2.6.5. Online interaction

This type of interaction is always mediated by the use of machines. Thus, its nature differs greatly from face-to-face interaction. Technological innovation pushes emerging new properties (real-time resources) affecting group interaction, which are unlikely to capture and define in traditional competences scales. Usually, misunderstandings and errors are not spotted nor corrected very often and faster as in face-to-face interaction (COUNCIL, 2018). Among prescripted requirements for successful communication, can be mentioned the need for more redundancy in messages, need to check that messages have been correctly understood, the ability for message reformulation to help comprehension and understanding, and ability to handle emotions (COUNCIL, 2018).

Online conversation and discussions

Online conversation and discussion imply a multi-modal phenomenon of consistent communicative interaction. The emphasis is the usage of online communication, regarding all types of social exchanges including random serious issues. The operationalized concepts for the scale consist of simultaneous and consecutive interaction contexts, multiple interlocutors sustained interaction, posting and commenting, and contributing to others' interactions. Also, it comprises the ability to include several additional tools such as images, symbols, codes, tones, stress, and prosody to exploit irony, affective and emotional sides (COUNCIL, 2018).

Goal-oriented online transactions and collaborations

Online interaction is undermined by specific goals, as regular features of contemporary life. After technological innovations, there is no clear line separating written and oral aspects. Instead, there is an increase in multimodal tools and resources linked to particular uses and backgrounds. Some key concepts have to be highlighted:

- Online good purchasing and services

- Client service engagement transactions
- Collaborative project works
- Dealing with communication problems (technical difficulties)

2.6.6. Testing and Common European Framework (CEFR)

For a long time, linguistics has proposed many different ideas to explain the mechanisms and processes underlying L1 and L2 acquisition. Most of the L1 learners converge on the same grammar and a "native" proficiency level. On the contrary, L2 learners have personal and differential grammar and proficiency among them (Andringa et al., 2019).
Differences between L1 and L2 are support under the biological factor's assumption dealing with maturity and strength of neuronal links, as a matter of fact, some research has provided evidence on this paradigm (Abrahamsson and Hyltenstam, 2009; Monner et al., 2013) On the opposite side, there are some other researchers explaining that causes of the variety in language proficiency among L2 learners are related to contexts factors as for example quantity and quality of L2 input, motivational states, and extralinguistic factors. (Birdsong, 2005,0; Birdsong and Vanhove, 2016). Linguistic competence in native speakers can vary according to their own experiences, educational level, learning abilities, intelligence quotient, meta-linguistic abilities and need for cognition (Andringa et al., 2019; Hulstijn, 2011) points out the need to measure differences among native speaker to have a model to measure differences between native and non-native communicative competences.

It is indisputable that learning a new language brings benefits in different domains for the learner and for some surrounding contexts (Nakamura, 2019); for example, protection against dementia (Bialystok et al., 2012). Other examples of benefits are related to better opportunities to study, work, travel, and so on. Hence, some languages are more learned than others. According to data from The Washington Post (Noack and Gamio, 2015) and Duolingo (Pajak, 2016), there are seven languages that people learn the most. The data of these two sources agree in five languages: English, French, Spanish, German, Italian and they differ in

the rest: Japanese and Chinese for The Washington Post, and Swedish, and Turkish for Duolingo. In the case of South America, the most popular language studied is English. Regardless of the language to be learned, there is agreement on the way in which language learning is accredited and recognized: a language proficiency test; for example, TOEFL,IELTS, TOEIC for English; DELF, DAL Language proficiency test for French; DELE for Spanish, Goethe-Zertifikat for German, among others.

2.7. Language testing and assessment

Language testing has been defined as a relatively new, rich academic discipline into the applied linguistic field regarding reliable measurements and evaluation of language proficiency in test and assessment (Ginther and McIntosh, 2018). Language testing applies to L1 and L2 acquisition. Nevertheless, most approaches have centralized only on L2 acquisition, after English as a Second Language (ESL) teaching and learning development (Ginther and McIntosh, 2018).

Historical development

As cited by Ginther and McIntosh (2018), Spolsky (1995), identified three historical stages in language testing, the prescientific, the psychometric-structuralist, and the psycholinguistic-sociolinguistic.

The prescientific period characterized by a continuous dependence on examinations by learners and teachers as well. Even so, the judgment relied on a singular examiner.

The scientific-structuralist period focused on reliability. Individual evaluators applied different standards. Nonetheless, such as narrow scope including a single teacher/ examiner sooner showed defects.

The psycholinguistic-sociolinguistic period, had a broader conception evolving from reliabi-

lity to validity, It also proposed scores' underlying constructs. During this period, the concept of proficiency was used with conceptual limitations.

According to Ginther and McIntosh (2018), language testing historical stages of development produced initiatives of assessment like the Test of English as a Foreign Language (TOEFL). Two theoretical considerations also influenced this kind of examination test. First, Lado (1961) argued the possibility of identifying and then developing representatives samples of structural and phonological items based on the similarities or differences between examinees' first and second languages. Sampling must contain meaningful features of the continuum from easy to difficult. Second language mastery can be displayed from these tests.

A second consideration, made by Carroll (1961) whose works were addressed to ensure real-world settings for productive and receptive modes. At a certain level, both trends were included in TOEFL's initial attempts. Lado's discrete-structuralist approach concentred on structural aspects of language (e.g. verb tenses). Carroll's integrative approach was included in exercises, exploring language beyond mere structures (e.g reading comprehension items).

Language testing not only comprises linguistic approaches methods since it belongs to general test analysis, which blends into the psychological and educational measurement (Davies, 1984). In 1966 the National Council on Measurement in Education (NCME) along with associate institutions like the American Psychological Association (APA) and the American Educational Research Association, published the Standards for Educational and Psychological Testing, providing book updates every ten years. This text introduced several core concepts highlighting especially validity.

Validity

By the early 1980s (Ginther and McIntosh, 2018), four central types of validity were identified and accepted: face validity, content validity, criterion-related validity, and construct validity.

- Face validity or the apparent relationship between a test and the subject matter to be represented. It arranges evident credibility to establish validity.
- Content validity also covers the relationship between the contents of the test and the area tested, but it only considers empirical evidence and expert judgment.
- Criterion-related validity relates relationships among tests that measure similar abilities.
- Construct validity engages required investigation about underlying qualities of matters measured in a test (underlying construct correspondence).

The Standards for Educational and Psychological Testing in the version of 1985, proposed that validity was a unitary concept, pointing out the "[...] degree to which that evidence supports the inferences that are made from the scores"(p. 9) This kind of cognitive revolution is understood as a rejection of behaviorism in psychology and structuralism/empiricism in linguistics (Ginther and McIntosh, 2018).

Validity conceptualization went further with Cronbach (1984) validation is a matter of constructing validation. Angoff (1988), stated that "construct validation is a process, not a production, that requires many lines of evidence".

Kane (2013), proposed an assumption of validation as the evaluation of factors, such as coherence and completeness of the argument relation, interpretation-use dealing with the plausibility of the analysis inferences. Keane's argument-based approach draws back from philosopher Stephen Toulmin (2003,0) with his famous proposal presenting interdependence among claims, data, and warrants to define argument's quality basis.

John Oller Jr (1983), a scholar influenced by Carroll on intelligence research, proposed the Unitary Trait Hypothesis, following the discrete-point/integrative approach. Oller's hypothesis introduced an intelligence-related single factor underlying language proficiency. This factor was disaggregated from common underlying factors: reading, writing, listening, and

speaking. From the 1990s, a new trend impulsed the idea of proficiency concept as both unitary and divisible at the same time (Alderson, 1991; Sawaki et al., 2009) organizing the model into factors, specific factors, components, and subcomponents.

Consequential validity

Two quite apart dimensions are implicit in the validity of tests. One is inner, depicting the user's ability as part of the human integrality. The other one deals with public policies regarding work or academic admissions requirements. It is expected that both dimensions collide, despite in practical terms they are distant from each other (Messick, 1998).

Communicative competence and language proficiency models

After Chomsky (2014) notion ideal speaker-listener, unaffected by environmental conditions (speech communities). Hymes (1972) introduced the concept of communicative competence, including strongly communities' influence in applied linguistics. Canale and Swain (1980) developed a communicative competence approach for language teaching and testing. Among the innovations, can be listed: grammatical, sociolinguistic, and strategic competences. Some scholars consider, there was an expansion of language from restrictive structural contexts to real-world communicative contexts, a jump towards the functional, communicative test.

Discrepancies in proficiency scales

As mentioned above, the Council of Europe created an extended document regarding language ability teaching, curriculum, and assessment. The counterpart, the USA education system, has not proposed such as descriptive paper, but it did propose a proficiency scale, taking into consideration its scope for the descriptors.

The American Council on the Teaching of Foreign Languages (ACTFL) prepared proficiency

guidelines (ACTFL and Portal, 2012) attempting to represent communicative competence and language proficiency testing, directly referenced to classroom teaching. This proficiency guidelines indicate what users/learners "can do" in terms of speaking, writing, listening, and reading in real-world situations in a spontaneous and non-rehearsed context. Under this extent, there are five levels: Distinguished, Superior, Advanced, Intermediate, and Novice. These levels can be subdivided as well into high, mid, and low sub-levels.

The ACTFL Guidelines and the CEFR have in common few but important elements, first the positive inclusive descriptor "can do", second the real-world grounded aims, and third the applicability of these headlines to teaching, curriculum design, and assessment.

Language proficiency test

Language testing is a widespread discipline applicable to most of the world languages - minority languages barely transcendent to the business spheres -. The scope of this document is the English language. Thus, is necessary to understand purposes underlying international examination tests of English language proficiency (ELP).

Brown (2019) differentiated among three usual types of English language applicable to proficiency: World English (WE), English as lingua franca (ELF), and English as an international language (EIL). WE is a general concept of English gathering different language communities of the English language around the world (initially, including those places where English has a strong impact on population). Brown (2014) opposes WE to Native Standard English, which is for him, an idealized concept. ELF describes the use of English in international contexts by speakers who are non-native speakers of English. EIL indicates the massive usage of English for cross-cultural processes.

Critics and concerns on international standardized english proficiency test

Davidson (2006) featured concerns about "the disconnection of powerful language English test from the insights of the analysis of the English language in the world context".

The two major English proficiency tests are the Test of English as a Foreign Language Internet Based Test (TOEFLiBT), and The International English Language Testing System (IELTS). The advertising and the names may imply their scope is on general and international English. TOEFL is designed indeed to evaluate English-language proficiency (Jamieson et al., 2000). IELTS frameworks also include the note of being an exam for measuring English-language proficiency (Humphreys et al., 2012). According to the analysis made by Brown (2019), the test takers could make mistakes by taking the wrong examination in agreement with their needs.

Now, as the International language test is based on proficiency measurement, it needs to be clarified that proficiency is the sum of knowledge and ability to do something (specific task). Proficiency fixes any skill by all means (e.g. Piloting -aircraft- proficiency). Hence proficiency criteria bases upon native speakers' standard English. Davies (2003) featured the native speaker's standard English with the following conditions:

- Language acquisition during childhood
- Linguistic intuition, grammar and idiolectal.
- Sensibilities about grammar and idiolectal variations.
- Production of fluent spontaneous English discourse.
- Wide-ranging communicative competence.
- Creative writing.
- Interpretation and translation into English.

After knowing these criteria, an obvious question appears, Why do International Proficiency Test measures skills exclusive to native speakers in non-native speakers? this is evidence of

the disconnection between the exams and the functional development of language.

Amidst scholars, alternatives to native speakers' standards have been presented (McKay and Brown, 2015) considers an amplification of the standard into a Global English Standard (GES) including language varieties and dialects worldwide.

Evolution of international language tests

Several changes have proved the evolution of International English language tests, according to Brown (2019) incoming modification on IELTS have included materials -reading and listening- come from various countries like Australia and New Zealand. Also (Taylor, 2006) references towards native speaker's standards slowly faded away.

Brown (2019) examines some cutting-edge approaches regarding proficiency measurement, they divide into two main groups:

Top-Down, language-focused Approaches:

- Truth-in-advertising approach (real information for users) (Brown, 2014)
- Multiple WE approach (Include more English speaking and non-speaking countries' varieties and dialects) (Jenkins, 2006)
- ELF approach (Focus on communication between speakers from different languages) (Jenkins, 2006)
- GES approach (Focus on functional grammar, written forms, and recognition of language varieties) (McKay and Brown, 2015)
- Functional approach (assessment is restricted to what is functional to the users)

Botton-Up, person-focused Approaches:

- Effective communicator approach (focused on levels of effective communication)
- Scope proficiency approach (show differences among internationally, nationally and locally effective) (Canagarajah, 2006)
- The scale of range approach (Melchers et al., 2019)
- EIL intelligibility approach (focus on the rate of intelligibility of utterances and writings (Munro and Derwing, 1995)
- Resourcefulness approach (interactional use of resources, two types high and resourcefulness) (Firth and Wagner, 1997)
- Symbolic competence approach (focus on the ability to approximate or appropriate a language and the ability to shape language learning context) (Kramsch, 2011)
- Intercultural communicative skills approach (intercultural communication among non-natives only) (Clyne and Sharifian, 2008)
- Performative ability approach (communicative negotiation) (Canagarajah, 2006)

Traditional assessment

In the international English language tests reviewed above, common issues glitter over the rest, descriptors, which are modalized declarative statements using the form “can do”. These descriptors are subsequently, placed into a rubric or scale (see section 2.6.3).

Any assessment under a particular examination test type follows the criteria impressed in descriptors and scales. Thus, examiners contrast test-taker answers or procedures to conclude if “can do” or “can not do”. Paper-based test and computer-test based version of the test, follow this procedure.

The particular scope of international English language tests defines the type of scale and its descriptors. As reported by Brown (2019) (Hudson et al., 1992) scales have at least two

forms: unidimensional (e.g. when having a single scale for written interaction), and multi-dimensional having multiple assess items, it would be represented by the sum up of various unidimensional scales.

The new approaches through the evolution of the international English language test, authors like (Canagarajah, 2006) broadened the use of scales including ratings of newly assessed aspects such as pragmatics. Nonetheless, the use of scales and rubrics is perceived as the preferred resource, for not saying the only one, in measurement proficiency.

2.8. Proficiency

To make a comprehensive concept of Language proficiency, some clarifications are needed. First Language Proficiency is divided into two constructs (Hulstijn, 2011).
Language Proficiency of native speakers (LP1) and Language Proficiency of second language speakers (LP2).

2.8.1. Native Language Proficiency (LP1)

Speakers of a native language L1 can communicate successfully with each other to a certain degree to the extent of linguistic knowledge they share, this is called Basic Language Cognition (BLC) (Hulstijn, 2011) BLC is restricted to speech reception and speech production and it does not comprise writing and reading[6].

BLC refers only to the frequent morphosyntactic and lexical structures that may occur in any communicative situations, that is, common to all adults L1 speakers. BLC occurs regardless social, age, educational level, literacy and geographical distribution factors. For (Hulstijn, 2011) BLC is restricted to speech reception and A second kind of cognition complements

[6](Hulstijn, 2011) indicates that her proposal makes this distinction in accordance with linguistic structuralism principle that indicates as a human attribute more the speech than literacy. (Saussure, 1916); (Bloomfield, 1933)

BLC, High Language Cognition or (HLC). Descriptively HLC is identical to BLC, but it differs at the lexical and morphosyntactic levels, due to the utterances produced and understood contain low-frequency lexical entries and uncommon morphosyntactic structures (Hulstijn, 2011). As a highlighted contrast from BLC, utterances produced in HLC pertain both written and spoken language as well. This feature indicates their more complex nature. The HLC utterances usually are produced in academic, work, and technical contexts.

2.8.2. Second Language Proficiency (LP2)

First attempts to conceptualize LP2 were proposed over 50 years ago, (Lado, 1961) and (Carroll, 1961; Carroll et al., 1971).Early models consisted basically in a two-dimensional grid, containing linguistics knowledge along with one knowledge lexis, morphology, syntax, phonology and orthography or spelling, crossing with, the basic language skills listening, speaking, reading and writing (Hulstijn, 2011) (Canale and Swain, 1980) proposed a model of communicative competence consisting in three interactive competences: grammatical, sociolinguistic and strategic. This LP was extended later by (Bachman et al., 1996), where they included levels of hierarchy layered using three language abilities (see Figure 1) organizational language knowledge (grammar and textual knowledge) pragmatic language knowledge(functional and sociolinguistic knowledge) and strategic competence (metacognitive components and strategies). Despite the popularity of this theory, scholar soon found that it was extremely difficult to obtain empirical support (Hulstijn, 2011) (Bachman et al., 1996) developed a study conducted to find evidence of the three traits from their theory (grammatical, pragmatic, and sociolinguistic competences). Participants were 116 English as a Second Language students, in the United States, with heterogeneous descriptive files referring location, age and length of residence. The date recollection used had multiple instruments like oral interviews, writing tasks, multiple-choice test and self-ratings. The evidential findings obtained from the study confirmed the existence of a general factor and two specific factors: Sociolinguistic competence and grammatical/pragmatic competence (Hulstijn, 2011). (Sasaki, 1993) made a study on homogenous English as a Second Language students from Japanese universities who had learned English in controlled academic environments for over seven years. 11 scores

BLC		
Unconscious (Largely) implicit Knowledge	**Conscious (Largely) Explicit Knowledge**	**Automaticity**
Phonetics	Lexical domain, form-meaning mappings	Automacity[7] with these types of knowledge (implicit and explicit) are processed
Prosody		
Phonology		
Morphology		
Syntax		

Table 2-2: Previous descriptions are proposed for individuals without mental or medical disorders interfering in the normal speech production and reception –oral and written-.[8]

for subtest were proposed after they were derived from three tests comprising a free composition paper, a short-text multiple-choice test (SMC) and a long-text multiple-choice test (LMC). LMC included listening and reading comprehension and fill-in-the-gap.

Recently studies conducted in the Netherlands took students and adults with L1 Dutch and L2 English to measure and analyze the componential structures of reading and writing skills (Schoonen et al., 2003; Van Gelderen et al., 2007) these studies recognized in the researching investigation.

2.9. Second Language Proficiency Assessment

Language proficiency in a second foreign language (L2) comprises the ability to do "something" with the language but also knowing about it (Harsch, 2016). This proficiency can be understood as pragmatic knowledge to face real-life communicative situations. After Hymes (1972), two parallel approaches were suggested. The first one takes into consideration the

[7]Automacity can be understood as the speed in the processing of speech.

[8]Table designed especially for this paper from the information taken explicitly from (Hulstijn, 2011, p. 230)

sociolinguistic and discourse abilities needed to communicate appropriately. The second one is based on the performance as a result of intertwinedness among linguistic, mental and social competence and their mutual dependence on context (Bachman et al., 1996). Nonetheless in the educational domain there are other assumptions for proficiency (Cummins, 1979) divided mainly into two groups according to the linguistic aim, namely: the Basic Interpersonal Communication Skills (BICS) or everyday interaction skills, and the Communicative Academic Language Proficiency (CALP) or academic and schooling knowledge communication. In terms of language testing proficiency discussion focuses on its innate nature as unitary (Oller, 1979) or divisible (Palmer and Bachman, 1981). The unitary vision supposes the existence of an indivisible underlying structure (Tesnière, 1959).

The divisible proficiency theory (Palmer and Bachman, 1981) holds that proficiency could be divided into subcategories e.g.: writing, speaking, listening, and reading. Over the last two decades a "multidimensional conceptualization of language proficiency" has pointed out the existence of a set of different communicative skills and strategies. The Common European Framework of Reference for Languages (CEFR) (of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, 2001), based on a divisible language proficiency model, depicts six different ascending levels of proficiency as follows A1, A2, B1, B2, C1, and C2. The A-labels correspond to Elementary level, B-labels to Intermediate and C-labels to Advanced. CEFR describes proficiency from overall skills and abilities to particular and less important aspects of human communication. CEFR has become a mandatory tool for teaching materials, curricula, and assessment since it was proposed in 2001. Language proficiency exams are presented into separate sections like reading, listening, speaking, writing. Each section assesses a particular skill needed for communication, expression and language understanding. The language aim for these exams is expected to cover everyday contexts.

2.10. Concepts from computational science

2.10.1. Wisdom of the crowd

The concept of the wisdom of the crowd refers to "the aggregated opinion of a crowd" (Hosio et al., 2016) which is a statistical phenomenon. There is no social nor psychological nature underlying the wisdom of the crowd. Historically, many authors have mentioned this phenomenon but was Galton (1907) who first included in an academic paper. Galton's research pointed out that the collective member knowledge in the audience of a weight-judging contest of a fat ox remarkably outperformed opinions from field experts (butchers). Since those days many scholars in different disciplines have verified similar findings (Page et al., 1999; Surowiecki, 2005). Recently, wisdom of the crowd have been applied to computationally challenging problems (Hosio et al., 2016; Surowiecki, 2005) proposes four qualities that "validate" crowd to be statistically relevant:

1. The crowd needs to be diverse, this ensures that individuals can offer different bits of information to the analysis.

2. The crowd needs to be decentralized to avoid dictation of collective outputs by hierarchy structures.

3. It is essential to have mechanisms and tools for summarizing different opinions.

4. The members of the group have to be independent, to avoid affectation from other's member's opinions.

Conversely, there is a factor that undermines the validity of wisdom of the crowd, social influence, which refers to how opinions of peers affect individual judgments. Ideally, crowd members should not be aware of each other's opinions, due to the natural human tendency to seek consensus (Yaniv and Milyavsky, 2007).

The question and answer sites (QA) in the internet along with some social network sites use the principles of wisdom of the crowd for analyzing interactions among users e.g.: votes,

surveys, rankings, etc. Usually, the data collected come from a large number of people (representing the crowd). Independence of the individual's opinions can vary according to the nature of the site itself. In this case of study, internet ensures a diverse and decentralized crowd as well as many different tools and information processing resources.

2.10.2. Stack overflow

Stack overflow[9] is a well-known Question-and-Answer (Q&A) site specialized in computer programming. Created in 2008 by Jeff Atwood and Joel Spolsky. The site serves as a platform for enthusiastic users[10] to ask and answer about programming-related topics. Users can vote questions and answers up and down, also QA edition is possible. Users earn reputation when receiving positive votes and answers to posted questions, besides after worthy contributions users receive badges. Site privileges get unlocked with the increase of reputation. Privileges bring to users the ability to vote, comment and edit other user's posts.

Stack overflow uses close questions in the search for improving quality of the posts, differentiating from other Q&A sites such as Yahoo! Answers. In January 2019 Stack Overflow had over 10 million registered users[11] and by the mid 2018 around 16 million posted questions. The most discussed topics in Stack Overflow are languages of programming such as JavaScript, Java, Python and HTML among others.

Stack Overflow ranks users according to the reputation earned through interaction and participation in the posts. Reputation indicates that users possess certain level of knowledge, for instance a high reputation implies a high-level knowledge[12]. Statistically most of users interact once or twice, mostly posting questions; some few users have frequent interaction, earning points or badges. When posts or replies are worth for the community, reputation increases among users (phenomenon reflected on the ranking). Expert's opinions or interac-

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

[10] `https://en.wikipedia.org/wiki/Stack_Overflow`

[11] `https://meta.stackoverflow.com/questions/302884/10-000-000th-question-is-here`

[12] Knowledge for programming includes theoretical and pragmatic skills.

tions can be considered as the "gold standard". Programmers contrast information discussed in the site with their own experiences and projects. The knowledge sharing in Stack Overflow circles in the context of a well-informed community. As an extra service Stack Overflow help technology companies to find candidates for their positions by releasing users' databases for consulting.

Apart from the success and popularity, Stack Overflow have improved manners in which users approach to Q&A by using gamification (points, badges, rankings, privileges) which encourages them to participate actively.

2.10.3. YASK

YASK is a Colombian startup created in 2014 by Andrea Higuera and Alejandro Zuleta to learn up to 13 different languages, available in app stores. YASK is a collaborative social network with more than 20.000 registered users who indicate those languages they are proficient in and those expected to learn. The services deal with writing and speaking express translation or correction. Additionally, through collaborative posting, information is shared among users to obtain positive or negative votes.

YASK machine selects automatically users to vote posts, giving three basic options as correct, incorrect, correction (revisor modifies entry). The process pace is entirely given by users, as they collect "XP" or points in the score label system. The advanced users have more privileges, utilities and resources in the APP as they move forward. This gamification system is a trend in educational APPs nowadays trying to follow the trace made by worldwide gaming companies.

2.10.4. PageRank

PageRank is a method for rating web pages objectively and mechanically by measuring the human attention devoted to them (Page et al., 1999). Initially, PageRank authors pointed out some problematic facts for the World Wide Web, such as the massive number of web sites

divided into different categories, topics, and targeted users. Additionally, at that historical moment -the late nineties- a vast number of inexperienced users had serious trouble handling websites.

The World Wide Web sites have a kind of "flat" document collections (Page et al., 1999) very difficult to organize, categorize and rank. The inner structure of the World Wide Web sites is a hypertext containing auxiliary information regarding link structures and link texts (Broder et al., 2000). PageRank uses the link structures of the website to produce a global importance ranking of every web page (Page et al., 1999). PageRank improves two main aspects simultaneously, on one hand, the performance of searching engines; on the other hand, the user's comprehension about the vast heterogeneity of the World Wide Web sites. The link structures of a web usually are represented visually as graphs, composed by pages (nodes) and their links among them (Han et al., 2009) (Han et al., 2012).

According to the citation measures (Garfield, 1996; Pearson and Lumpkin, 2011) highly linked pages are more "important" than pages with few links (Nassiri et al., 2013; Page et al., 1999). Despite this logical measure produced after the number of backlinks in a node, there are a factor to consider, a link coming from a prestigious website is much more valuable in comparison to many links coming from obscure sites (Broder et al., 2000; Movshovitz-Attias et al., 2013; Page et al., 1999). The importance depends directly from the ranks obtained out the sum of the backlinks (many links from obscured sites or few links from well-rank sites).

The intuition of PageRank is that the importance of a web page (a node) is the probability of a random surfer of visiting that page. A random surfer is the one who starts in a random page and, at each time, selects randomly a link on each page to follow. After many random choices some pages have received the visit of that random surfer more times that others. The number of visits received by a page is proportional to the probability of being visited by a random surfer. The PageRank for each node can be computed simulating that random surfer during a long time. However, this method is inefficient and computationally expensive.

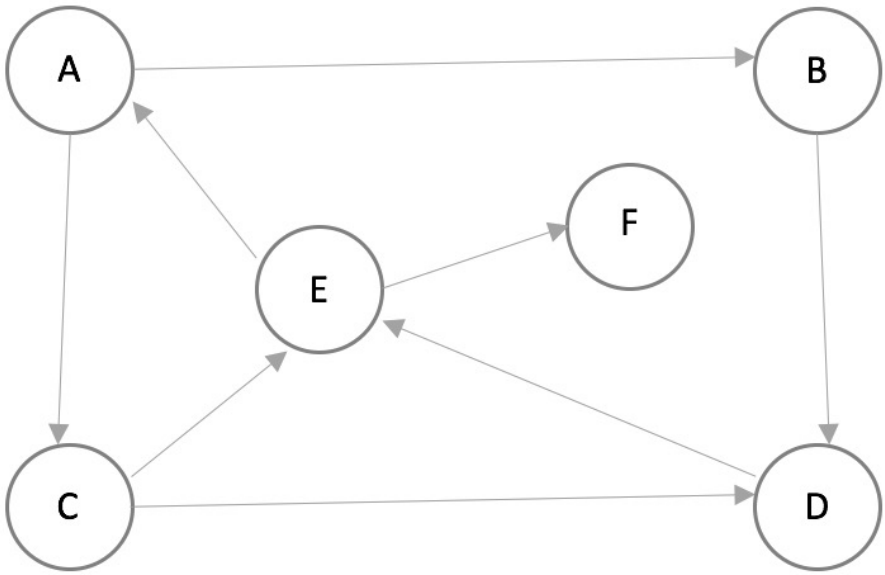


Figure 2-1: PageRank Model graph

To calculate PageRank efficiently, the system's graph is represented as an adjacency matrix. Each PageRank iteration consists of a matrix multiplication of the adjacency matrix with the current vector of nodes' scores to produce the next version of the same vector.. Let us consider an example graph in Figure **2-1**. Nodes are represented with letters A, B, C, D, E and links are represented with arrows following a fixed direction.

That graph can be represented as an adjacency matrix A (Figure **2-2**) where a number 1 in a particular position in a row and a column means, that there is a link starting at the node represented by the column to the node represented by the row, and a 0 otherwise indicating that there is no connection between these nodes. Note, that all the numbers in the diagonal of A are 0s, meaning that there is not allowed a self-link.

Note that node F in Figure **2-2** has only an income link, making impossible to the random surfer to escape of that node. Such types of nodes are called "sink nodes". To deal with sink nodes it is necessary to provide a mechanism to allow the random surfer to move escape from sink nodes. Such a mechanism is controlled by a parameter named "damping factor" d that

$$\boldsymbol{A} = \textit{Link to node} \begin{pmatrix} & \textit{Link from node} & & & & & \\ \textit{NODE} & A & B & C & D & E & F \\ A & 0 & 0 & 0 & 0 & 1 & 0 \\ B & 1 & 0 & 0 & 0 & 0 & 0 \\ C & 1 & 0 & 0 & 0 & 0 & 0 \\ D & 0 & 1 & 1 & 0 & 0 & 0 \\ E & 0 & 0 & 1 & 1 & 0 & 0 \\ F & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 2-2: Adjacency matrix for the graph in Figure **2-1**

assigns a small probability of jumping to any node in the graph when the random surfer is in a sink node. The values of the parameter d ranges between 0 and 1, having values close to 1 a small "damping" effect, and values close to 0 a large "damping" effect. For instance, Page et al. (1999) set d to 0.85 for the web information retrieval application, being a relatively low damping factor. The damping factor is applied to the entries of $a_{i,j}$ in **A** using the following equation: $b_{i,j} = d \times a_{i,j} + \frac{(1-d)}{n}; i, j \in [1 \cdots n]$, where d stands for the damping factor (i.e. $d = 0.85$), $a_{i,j}$ stands for the entry at $i-th$ row and $j-th$ column in **A** matrix, and n stands for the number of nodes in the graph.

The entries $b_{i,j}$ conforms a new matrix **B**, which is the adjacency matrix **A** transformed with the inclusion of the damping factor. For instance, the entry in the adjacency matrix in Figure **2-2** corresponding to the first row and the fifth column (a 1 entry) becomes:
Similarly, applying the damping factor to the entry in the first row and second column (a 0 entry) it becomes:

$$b_{1,5} = 0.85 \times a_{1,2} + \frac{1-0.85}{5}; b_{1,5} = 0 + \frac{0.15}{5}; b_{1,5} = 0.03$$

$$b_{1,6} = 0.85 \times a_{1,2} + \frac{1-0.85}{6}; b_{1,6} = 0 + \frac{0.15}{6}; b_{1,6} = 0.025$$

As a result, all the entries in **B** corresponding to a 1 become replaced by 0.88 and all zero entries by 0.03 as shown in Eq. 2-1.

$$\mathbf{B} = \begin{pmatrix} 0.03 & 0.03 & 0.03 & 0.03 & 0.88 & 0.03 \\ 0.88 & 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.88 & 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.88 & 0.88 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.88 & 0.88 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.88 & 0.03 \end{pmatrix} \tag{2-1}$$

The effect of the damping factor in the adjacency matrix is that the graph becomes fully connected and there is a non-zero, but small possibility for jumping from any node to any other node even if there is not a link in the original graph.

The Page Rank algorithm requires that the columns of B be probability distributions, that is that their values sum up 1. Therefore, it is necessary to normalize the columns in **B** by dividing their entries by the sum of each column, generating a new column-normalized matrix called **M**. Such normalization is shown in Eq. 2-2

$$\mathbf{M} = \begin{pmatrix} 0.016 & 0.030 & 0.016 & 0.030 & 0.468 & 0.167 \\ 0.468 & 0.030 & 0.016 & 0.030 & 0.016 & 0.167 \\ 0.468 & 0.030 & 0.016 & 0.030 & 0.016 & 0.167 \\ 0.016 & 0.850 & 0.468 & 0.030 & 0.016 & 0.167 \\ 0.016 & 0.030 & 0.468 & 0.850 & 0.016 & 0.167 \\ 0.016 & 0.030 & 0.016 & 0.030 & 0.468 & 0.167 \end{pmatrix} \tag{2-2}$$

To obtain the PageRanks for the n nodes of a graph (six nodes in our running example), the algorithm initiates those values with random numbers that sum up 1. These values are presented in a column vector of size n denoted as R_0 meaning the rankings of the nodes at

the iteration No. 0. Eq. 2-3 shows an example of such vector.

$$R_0 = \begin{pmatrix} 0.250 \\ 0.180 \\ 0.200 \\ 0.120 \\ 0.100 \\ 0.150 \end{pmatrix} \tag{2-3}$$

In the next iteration, the values in the vector of rankings R_{t+1} are updated by multiplying the $\mathbf{M}$ matrix by the current rank vector R_t. This recursive relation can be expressed in the following expression: $R_{t+1} = \mathbf{M} \cdot R_t$, where the operator "$\cdot$" is the matrix dot-product, here applied between the matrix $\mathbf{M}$ and R_t. It means that the ranks at the iteration $t+1$ depends only on M and the ranks at the iteration t. Equation 2-4 shows this operation to obtain the ranks at $t = 1$ for our running example.

$$R_1 = \begin{pmatrix} 0.016 & 0.030 & 0.016 & 0.030 & 0.468 & 0.167 \\ 0.468 & 0.030 & 0.016 & 0.030 & 0.016 & 0.167 \\ 0.468 & 0.030 & 0.016 & 0.030 & 0.016 & 0.167 \\ 0.016 & 0.880 & 0.468 & 0.030 & 0.016 & 0.167 \\ 0.016 & 0.030 & 0.468 & 0.880 & 0.016 & 0.167 \\ 0.016 & 0.030 & 0.016 & 0.030 & 0.468 & 0.167 \end{pmatrix} \times \begin{pmatrix} 0.250 \\ 0.180 \\ 0.200 \\ 0.120 \\ 0.100 \\ 0.150 \end{pmatrix} = \begin{pmatrix} 0.088 \\ 0.156 \\ 0.156 \\ 0.281 \\ 0.232 \\ 0.088 \end{pmatrix} \tag{2-4}$$

Similarly, R_2 is obtained from the multiplication between M an resulting ranking scores from Eq. 2-4, as shown in Eq. 2-5.

$$R_2 = \begin{pmatrix} 0.016 & 0.030 & 0.016 & 0.030 & 0.468 & 0.167 \\ 0.468 & 0.030 & 0.016 & 0.030 & 0.016 & 0.167 \\ 0.468 & 0.030 & 0.016 & 0.030 & 0.016 & 0.167 \\ 0.016 & 0.880 & 0.468 & 0.030 & 0.016 & 0.167 \\ 0.016 & 0.030 & 0.468 & 0.880 & 0.016 & 0.167 \\ 0.016 & 0.030 & 0.016 & 0.030 & 0.468 & 0.167 \end{pmatrix} \times \begin{pmatrix} 0.088 \\ 0.156 \\ 0.156 \\ 0.281 \\ 0.232 \\ 0.088 \end{pmatrix} = \begin{pmatrix} 0.140 \\ 0.075 \\ 0.075 \\ 0.234 \\ 0.336 \\ 0.140 \end{pmatrix} \quad (2\text{-}5)$$

This process is performed iteratively until the differences between the entries of R_t and R_{t+1} are small. When that happen the algorithm has converged to the final set of ranking scores for the nodes. In our example the values obtained at the iterations 17th and 18th are shown in Eq. 2-6.

$$R_{18} = \begin{pmatrix} 0.016 & 0.030 & 0.016 & 0.030 & 0.468 & 0.167 \\ 0.468 & 0.030 & 0.016 & 0.030 & 0.016 & 0.167 \\ 0.468 & 0.030 & 0.016 & 0.030 & 0.016 & 0.167 \\ 0.016 & 0.880 & 0.468 & 0.030 & 0.016 & 0.167 \\ 0.016 & 0.030 & 0.468 & 0.880 & 0.016 & 0.167 \\ 0.016 & 0.030 & 0.016 & 0.030 & 0.468 & 0.167 \end{pmatrix} \times \begin{pmatrix} 0.159 \\ 0.116 \\ 0.116 \\ 0.193 \\ 0.255 \\ 0.159 \end{pmatrix} = \begin{pmatrix} 0.160 \\ 0.116 \\ 0.116 \\ 0.192 \\ 0.255 \\ 0.160 \end{pmatrix} \quad (2\text{-}6)$$

As the importance of a node in a graph system under the PageRank algorithm depends on the sum up from incoming links. Prestige of a node depends from the values it obtains from incoming links.

Thus node E can be considered as the "most popular or important" due to the two incoming links deliver coming from two different nodes. Nodes E and D only have one outcoming link each, both addressing to node E. They put all the importance towards a unique node. The next less important node is D having a strong incoming link from D, but a link from C with . The other nodes have relations of incoming and outcoming links in different proportions, none of them is as strong as E and D.

2.10.5. Gamification

Gamification is a concept forged in the first decade of the years 2000 and extensively adopted since 2010 in various disciplines and academic fields (Rodrigues et al., 2019). The basic assumption of gamification indicates the use of computer games features in non-gaming environments.

In education online resources (Flores, 2015), some gamification features are comprised of game elements like:

- Badges: Achievement graphical representations.
- Points: Numeric scale representing actions done.
- Progress bars: Bars showing initial, current, and target position.
- Performance graphs: performance evidence
- Rewards: Merits awarding player.

The use in education is made to motivate and facilitate students to understand exercises and progress (meaningful process)

3 Automatically Assessing L2 Writing Proficiency and Expertise in Social Networks- State of the Art

As mentioned in the introduction, this paper aims to investigate an automated method to determine the level of written proficiency of L2 learners through the use of collaborative social networks. These task had only been addressed in the past through the use of Artificial Intelligence approaches that analyze the texts written by learners and produce proficiency scores. These approaches were based on training data consisting of a large number of texts evaluated by human raters in combination with expert knowledge of teaching curricula or predefined proficiency frameworks such as CERF and TOEFL. Although, these approaches are fundamentally different from the one proposed in this dissertation, it is important to review them because these are the only existing attempts to tackle the task. Another reason is that they are represented in this dissertation by the "CERF Baseline", which is used to compare the performance of the proposed method with previous approaches. For these reasons, that group of approaches is reviewed in section 3.1.

Regarding the use of social media to assess L2 proficiency in formal or informal educational settings, as far as we know, it has not been addressed in the past. The closest approach is that of Movshovitz-Attias et al. (2013) in the domain of computer programming using the Stack Overflow collaborative social network. This approach somehow inspired this dissertation, so it is reviewed in section 3.2. Finally, in section 3.3 some links between this state-of-the-art

and this dissertation are provided and briefly discussed.

3.1. Text-based Artificial Intelligence approaches

3.1.1. Lu (2017) approach

Lu (2017) reviews the importance of syntactic complexity in the construction of a more accurate definition of L2. Syntax deals traditionally with the organization of elements in the sentence level, but syntactic complexity upgrades and extends the concept to a sense of sophistication in the structures deploy in written production (Bulté and Housen, 2014; Lu, 2011; Ortega, 2003). In Knoch (2011) syntax belongs to the "Four Skills Mode of Communicative competence" and syntactic complexity is enlisted in the assessment scales of most of the international English language proficiency tests (TOEFLiBT, IELTS, CAE). A complex sentence level usually is taking into consideration, in medium or advanced levels of proficiency, just as modeled, in test's scales. Certain factors such as syntactic complexity, fluency, and accuracy are for second language acquisitions researchers, complementary dimensions in the L2 quality and proficiency (Norris and Ortega, 2009; Wolfe-Quintero et al., 1998).

Delving into written production, complexity turns into a linguistic feature illustrating manners, in which language tasks are produced, "elaborated and varied" (Ellis et al., 2003; Housen and Kuiken, 2009). Several subsequent constructs compound linguistic complexity: lexical, grammatical/syntactic, propositional, and interactional aspects (Bulté and Housen, 2014).

Attempts to measure syntactic complexity, have migrated to computer technology. (Lu, 2017) highlights three specific works, the Biber Tagger (Biber et al., 1999), Coh-Metrix (McNamara et al., 2014), and L2 Syntactic Complexity Analyzer (Lu, 2010).

The Biber tagger

Initially developed as a tool for multidimensional register variation and text (Biber, 1991; Biber et al., 1999) it is been recently used to inspect grammatical complexity in L2 writing (Biber et al., 2011,1), through this research projects, it has been proved that grammatical spectrum, covers a wide range of lexico-grammatical features associated to syntactic complexity, especially in writing for L2. Biber's analysis surveyed over 23 features, especially grammar complexity patterns like word length, clauses, and phrasal structures. Biber performed a factorial analysis to predict the use statistically. As Biber's analysis considered speech and writing, one of his first conclusion claimed that writing includes more integrate tasks than speech, thus this makes writing quite more syntactically complex (Lu, 2017). Biber and Gray (2013) also reported a set of procedures designed to ensure accuracy in the tagging of a corpus from TOEFL iBT exam responses, thus first stages comprise these steps:

1. Several iterations of tagging.
2. Manual checking.
3. Data cleaning.
4. Tagger revision and retagging.

   The second stage develops as:

5. Use of Perl scripts to correct lexically governed tagging errors.
6. computer-aided manual checking, and correction of selected features
7. Reliability evaluation was performed twice, after finishing the first stage

Coh-Metrix

Originally designed for the assessment of cohesion features in texts, and cohesion of mental representation of texts (McNamara et al., 2014), the system contains two types of measure i) complexity measures, and ii) syntactic pattern density. McNamara holds the idea that syntactic complexity indicates appropriately writing quality.

Coh-Metrix uses a generative tree-model set up by the Charniak (2000) parser. The initial assumption of a tree-model analysis on sentences is the embeddedness of some elements -they are invisible in the shallow surface-. Mcnamara identifies measure items SYNLE (left embeddedness) and SYNNP (embeddedness of noun phrases) the higher degree of embeddedness the higher syntactic complexity.

Coh-Metrix additionally uses the notion of Minimal Edit Distance (McCarthy et al., 2009) to estimate dissimilarity from one string to another by computing the minimum of operations to convert one string into another. the MED measure evaluates what a sentence requires to be edited and then have the same POS tags, words, or lemmas as the following sentence (SYNMEDpos, SYNMEDwrd, and SYNMEDlem); the higher value for editing required the higher syntactic complexity (Lu, 2017).

Relating syntax similarity can be defined as the proportion of intersecting nodes between their generative parse trees (Lu, 2017; McNamara et al., 2014). The common generative parse tree is obtained, after eliminating subtrees. McNamara et al. (2014) proposes the following similarity formula:

$$sim = \frac{T_{a \cap b}}{T_a + T_b - T_{a \cap b}}$$

Here $T_{a \cap b}$ is the number of common nodes between parse tree a and b, T_a is the number of nodes in parse tree a, and T_b the same for b.

To illustrate syntax similarity, Lu (2017) proposed the following examples: “The man cam” and “He entered the door”. The two parse trees are represented in Figure **3-1** (from Figure 1 in Lu (2017)).

The two generative parse trees have eight and 10 nodes, they have six nodes in common (marked by #). Applying the formula, the syntax similarity between the two generative parse trees is: 6/(8+10–6)=0.5. The higher values indicate a higher degree of similarity and a lower degree of syntactic complexity. Similarity opposes to syntactic complexity.

As a conclusion from Biber and McNamara's proposals, the higher the syntactic complexity, the harder to the user to understand a text.

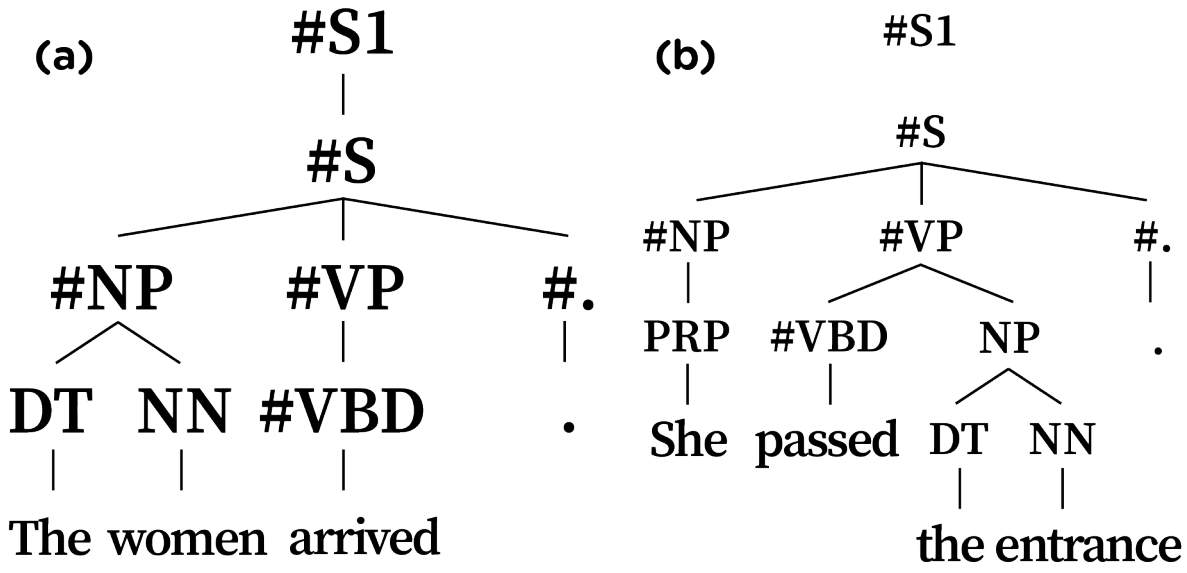


Figure 3-1: Parse trees examples taken and modified from Lu (2017)

The assumptions underlying conclusions, about measures obtained after computing with Coh-Metrix McNamara et al. (2014) highlights that, for quantifying the degree of sophistication in at least two types of embedded structures, measures of syntactic complexity can be used. In the this sense, measures of syntactic pattern density can be used to assess the frequency of complex patterns. The syntax similarity measures allow us to quantify the extent to which varied structures are used across different sentences. Even though the reliability rates have not been published yet, precision should be around 90 percent (Lu, 2017).

L2 Syntactic Complexity Analyzer (L2SCA))

This tool (Lu, 2010) designed "to automate syntactic complexity analysis of L2 English textsüses 14 different measures divided into 5 categories as follows: i) length of production unit, ii) amount of subordination, iii)amount of coordination, iv) degree of phrasal sophisti-

cation, and v) overall sentence complexity. L2SCA is similar to Coh-Metrix, but it is more systematic. Measure items are highly coordinated to categories. Operationalization used manually data collection, during two phases of experimental implementation, considering about 50 essays from Chinese learners of English as an EFL (Lu, 2010; Yoon and Polio, 2017). The reported correlation coefficients were about .81 between human rating and L2SCA.

Findings and conclusions

As these computational tools have been proved in different experimental settings some sort of conclusions can be mentioned:

The Biber Tagger found markers for distinguishing high and low scored essays, the number of clauses, and phrase-level (Taguchi et al., 2013). (Friginal and Weigle, 2014). The contrast analysis was made comparing rubric and scales proposed by Jacobs et al. (1981). In the case of Coh-Metrix, McNamara et al. (2014) concluded that better quality is obtained by human raters. It also reported items identify themselves with holistic ratings. L2SCA also offered some predictive markers, related to the holistic rating. In terms of assessment syntactic complexity is an item offering quality information for defining proficiency in users. Additionally, experiments proved the operationalization and correspondence to rubrics and scales.

3.1.2. Pilán (2018) approach

A complete and rather innovative proposal Pilán (2018), Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning. This work focused on the Swedish language. The cornerstone concept, in this case, is linguistic complexity which, as Pilán (2018) assures, can be used to determine both L2 proficiency and readability. Proficiency comprehends a level of analysis in the users' set of skills at using L2, while readability is understood as L1 skills in users with low reading levels or/and cognitive impairment. The proposed approach implies the use of machine learning and the implication of various similar dimensions.

Linguistic Complexity

Linguistic complexity has multiple assumptions and meaning, Palotti (2014) defines it as the aspect of language that makes communication quite simpler or easier at speaking or describe languages inner features. From a philosophical perspective, linguistic complexity refers to revision and study of language over itself. The role of linguistic complexity in relevant in efficiently processing and conveying information and, meaningful communication, besides influences tasks' performance (Tomanek et al., 2010).

Agent objective complexity enhance complexity to the object, language, or communication; **agent objective complexity** places its focus on the difficulty experienced by a performer who tries to use language. **System complexity** understands the whole language as a unitary system, characterized by certain difficult-to-understand aspects. Indicates how much a person requires to learn to be proficient. **Structural complexity** indicates a feature or particular aspect of language that can be considered more complex than the rest i.e utterances and sentences[13].

Readability

Readability measures started in the 1940s, Dale and Chall (1949), defined as the sum of elements within printed pieces, which affects readers' understanding. Thus, in the readability concept properties from both text and readers gather. Among text properties, there are morphosyntactic- structures and semantics concepts. Properties related to readers deal with life experience, educational level, and motivation. The use of the quantitative measurement introduced new tools such as formulas, first exploring surface or shallow layers in the texts. These formulas identified words and sentences to tokens. Most of the analysis was made on the binary distinction (Pilán, 2018). In 1975, a popular formula for readability was proposed by (Kincaid et al., 1975) this formula, with an educational goal, indicates a U.S. school grade level or the length of education (in years) necessary to understand a given text (Pilán, 2018).

[13] `https://blogs.ntu.edu.sg/hg3040-2014-5/?page_id=142`

$$FK = 0.39 \times \left(\frac{N_w}{N_{sent}}\right) + 11.8 \times \left(\frac{N_{syll}}{N_w}\right) - 15.59$$

In the previous formula N_w represents the number of words in the text, N_{sent} the number of sentences, and N_{syll} the total number of syllables.

A similar formula, as a readability index for Swedish (Björnsson, 1968) replaced syllables word from the percentage of long words (N_{longw} is the number of long words having more than 6 letters)

$$LIX = \frac{N_w}{N_{sent}} + \frac{N_{longw} \times 100}{N_w}$$

Pilán (2018) includes other readability formulas proposals, like Nominal ratio (NR) based on morphological information density capturing (Hultman and Westman, 1977). Studies have evolved further from technical operationalization to readability models for a wide set of languages counting English (Collins-Thompson and Callan, 2004; Feng, 2010; Miltsakaki and Troutt, 2008; Schwarm and Ostendorf, 2005; Vajjala and Meurers, 2012). According to literature revision, (Pilán, 2018) states the transition of readability uses and applicability from L1 education level scale measurement to serve as a reflector of proficiency. Data for research comes from, in most of the cases, from simple coherent pieces of text (Pilán, 2018), but also can include complete coursebooks and exams. A relevant finding after revision is that two thirds from machine learning research based on the CEFR.

According to literature revision, Pilan (2018) states the transition from readability uses and applicability from L1 education level scale measurement to serve as a reflector of proficiency. Data for research comes from simple coherent pieces of text (Pilan (2018), but also can include complete coursebooks and exams. A relevant finding after revision is that two thirds from machine learning research based on the CEFR. Smaller units of analysis (i.e sentences) have been deployed in studies regarding linguistic complexity (Karpov et al., 2014; Pilán et al., 2014). Proposals are aligned to the CEFR. Both studies have made binary distinctions of sentences in three possible scale-levels: bellow B1, B1, or above B1. Ströbel et al. (2016)

introduced Cocogen (Complexity Contour Generator) a system that builds up complexity contour of text by running a comparison to contours from expert-authored texts and L2 high proficient users text.

Proficiency level prediction for learner texts

Barrot (2015) highlights the imbalance between receptive linguistic complexity and productive linguistic complexity. Thus, many theoretical researchers take advantage of this and just study reading comprehension skills. The CEFR-level annotated corpora is a well-structured material available for automatization, conversely, few projects have been developed (Hancke and Meurers, 2013; Nicholls, 2003; Tenfjord et al., 2006; Wisniewski et al., 2013). The methodological differences in assessing receptive and productive are, that the first tends to be relatively error-free. On the contrary, production in L2 environments usually contents error (as a normal process in the acquisition of linguistic skills) which could affect negatively values estimation.

Sentence selection of corpora

Sentences are valuable for illustrating the authentic language, but a careful selection is required to fulfill the criteria of appropriateness in the expected use (O'keeffe et al., 2007). One of the main problems regarding associated use of isolated sentences is context-dependent, that is the confusion arising when there is at least one reference to a concept placed outside the sample.

Pilán (2018) reviews GDEX, Good Dictionary Examples (Husák, 2010; Kilgarriff et al., 2008). This resource makes operative factors such as typicality, informativity, and intelligibility. Besides, it includes in the interface aspects such as sentence length, word frequency, pronouns, anaphors as well as proper sentence beginning and end (capital letter and punctuations).

Corpora

The work developed by Pilan (2018) focused on Swedish language proficiency. Thus, she described the corpora available for Swedish at the moment of her writing. The description was made under two criteria, corpus build on coursebooks and corpus-based on L2 learners essays.

Korp - a corpus infrastructure: available on the internet, build on the infrastructure of Språkbanken. It is a constant-growing database, containing diverse types and genres of texts (informal, scientific, literary). Korp presents tools easy-to-read LäsBaRT (Heimann Mühlenbock 2013) and the newspaper Åtta sidor "eight pages". The utilities of Korp are search and extraction of statistical information, concordance, or keyword in context.

Sparv, an annotation pipeline is an automatic linguistic annotation system for Korp (Borin et al., 2016). Sparv includes lemmatization, part of speech tagging and dependency parsing. It includes other informatics tools like SALDO lexicon HunPos (Halácsy et al., 2007) and MaltParser (Nivre et al., 2006).

COCTAILL (Corpus of CEFR-based Textbooks As Input for Learner Level modeling) is a corpus of Swedish L2 coursebooks for CEFR levels from A1 to C1 (Volodina et al., 2014). this corpus was designed as a subsidiary tool for research on language-related issues. The corpus presents two annotated coursebooks, including classic sections such as lessons, reading texts, and exercises. Each content item indicates the goals and skills involved.

The SweLL (Swedish Learner Language) is a pilot corpus containing three different types of subcorpora. Two of these subcorpus consist of written essays, within preparatory language courses and exams for university studies. The third has essays written by newly arrived immigrants in schools. The corpus contains 144 essays and 144.000 tokens. Annotations were mainly done by teachers, but the CEFR level is unknown. As a plus, the SweLL contains metadata about personal information and time residing in Sweden.

A teacher-evaluated dataset of sentences or HitEx a small databased develop in (Pilán, 2018) the proposal presents generic corpora automatic assessed in the CEFR scales, under some criteria such as well-formedness, independence from the rest in the textual context and some lexical, grammar and syntactic features. HitEx is composed of 330 sentences and 4.060 tokens.

Lexical resources

Pilán (2018) makes an additional description of the lexical resources deploy in her proposal KELLY, SVALex, SweLLex, and SALDO. KELLY is a project developed by the EU, which has the purpose to provide language learning resources available for nine main languages used worldwide including Swedish. KELLY allows estimating use frequencies by web texts (SweWaC composed of 114 million tokens). The performance is possible through SketchEngine (Kilgarriff et al., 2014). The Swedish KELLY encompasses categories like headword lemma, word class, identification under numeric frequency decreasing position, raw frequency, normalized word by million, and CEFR level. The Swedish KELLY has 8,425 entries classified into the CEFR levels based on frequencies.

SVALex and SweLLex is a list intended for learners, teachers, researchers created on L2 data. the estimation for these lists are under the CEFR levels and are estimated using COCTAILL. The entrances of SVALex and SweLLex are denominated as lengram, a combination of a lemma, its part of speech, and an index number, identifying a table of inflectional and compound forms (Pilán, 2018). An important difference from other methods is that SVALex and SweLLex are not based on Raw Frequencies, but on a similar idea, the term frequency (TF) and the document frequency (DF). Carroll et al. (1971); **?** dispersion index was used in calculations.

The last lexical resource revised in Pilán (2018) is SALDO (Borin et al., 2013) is a tool based on word senses' associations, an alternative to Wordnet (Fellbaum, 1998) for the Swedish language. This proposal expects to be computationally easier, as well as, cover almost all parts

of speech and their intricate hierarchy relationships. SALDO offers a sense descriptor analysis possible after entries's centrality is determined in terms of frequency, undermarked style, semantic map relations, and morphological setting. Occasionally syntagmatic relationships among terms are considered.

Method

The method presented in this paper, Pilán (2018) uses machine learning techniques under the trends from Natural Language Processing (NLP). Machine learning holds the scenario of taking a set of data as an example to make predictions about unknown data (Witten and Frank, 2002). The algorithms used in the research are WEKA and scikit-learn (Pedregosa et al., 2011).

Pilán (2018) also includes a linear regression to predict numerical outputs. The vector products have some specific values as i.e weight vectors. Logistic prediction serves to perform classification and to make binary predictions for an instance (sample in a data set). The use of support vector machines (SVM) works on the mapping of non-linear data (use of kernel data) showing a higher dimension that permits linear limits identification (Witten and Frank, 2002). In evaluating the method for measuring, it is noted the multiple available tools under supervised machine learning algorithms. In readability (Witten and Frank, 2002) four instances can be determined, true negative, true positive, false positive and false negative. In evaluating the method for measuring, it is noted the multiple available tools under supervised machine learning algorithms. In readability (Witten et al. 2011) four instances can be determined, true negative, true positive, false positive and false negative. The sum of true and false positives divided the total number of predictions estimate the accuracy of the system. Precision is calculated as TP/(TP+FP); recall as TP/(TP + FN). Pilan also mentions other measures like adjacent accuracy, quadratic weight kappa, and cross-validation.

Domain adaptation

Due to the data sparsity problem in the creation of annotated corpora, which is difficult to obtain in terms of time and effort, some transfer learning methods can be applied. This type of strategic focus in targeting towards research domain directly. Pan et al. (2011) consider relevant the categorization of learning methods based on human's ability to solve tasks faster from similar activities available -domain adaptation-.

Method findings

Pilán (2018) concentrates her explanation first in the two main interests regarding prediction for Intelligent Computer-Assisted Language Learning (ICALL), being i) the size of the linguistic unit (sentences vs. texts), and ii) the type of data (expert-written texts as opposed to learner-produced texts). Pilán (2018) studies both from linguistic complexity and readability.

In the analysis of complexity, some categorizations constrain (Menn et al., 2014) were revised, and the decision made was a division into five groups: count-based, lexical, morphological, syntactic, and semantic. The total number of feature included: 61.

Count-based features are based on readability measures by Dale and Chall (1949) and include sentences and token length that can show the syntactic difficulty. The average token length is not longer than 13 characters (very long in the Swedish language). This analysis proposes a Swedish reading formula in which sentence length is an average of six characters. Also, the type-token ratio (TTR) becomes an indicator of lexical richness; in the reviewed thesis, a bi-logarithmic and a square root TTR to decrease the effect of text and sentence length Vajjala and Meurers (2012).

Word-list based lexical features indicate that word frequencies influence lexical complexity. The lexical entrenchment hypothesis (Diependaele et al., 2013) shows that frequency words are especially demanding to understand and produce initial users/learners. High proficiency level users/learners are typically characterized by low-frequency entries (utterances, words).The use of CEFR for each lemma in KELLY list concedes to extract information.

There are no absolute counts, but the distribution of tokens per CEFR level. They compute incidence scores (INCSC) by dividing 1000 with the total number of tokens (Nt) and multiply that with the count of a certain category of tokens (Nc) in the text or sentence as shown in (10) (Pilán, 2018).

Morphological features include INCSC of different morpho-syntactic categories and variation scores, i.e. nouns (N), verbs (V), adjectives (ADJ), and adverbs (ADV). The verb cases in Swedish are also included. The INCSC besides includes punctuation marks in the analysis. For the syntactic features, Pilán (2018) uses MaltParser (Nivre et al., 2006). The results are indicated with tags above the words in the sentence. The measure of embedded items was considered in clauses pre and pos modifiers Heimann Mühlenbock (2013). Semantic features were restricted to basic disambiguation.

Receptive linguistic complexity analysis

The WEKA implementation comprises a regression classifier. 867 datasets instances for texts and 1874 for sentences, spread across CEFR levels from A1-C1, collected from COCTAILL corpus. The performance achieved shows a correct classification of 8 out of 10 texts and 6 out of 10 sentences, meanwhile, human performance was between 50 percent to 67 percent. Lexical features are predictors of the text level and at the sentence level. CEFR text contains a large number of words from lower levels.

HitEx

HitEx is a corpus example selection system, designed following the previous data from dictionaries. Complexity samples search requires a selection of sentences with a well-formed and appropriate degree of complexity isolated from the context. Revision, brought conclusions like “context-dependence based on referring expressions of a different kind such as pronominal and adverbial anaphora and those sentences containing structural connectives, where the first clause referred to remains outside of the sentence boundaries” Pilán (2018). Lexical

aspects of the sentences samples, must avoid pedagogically inappropriate words, abbreviations, and proper names. The pedagogical relevance of HitEx was tested by teachers (330 sentences taken from generic Swedish corpora, which met three basic requirements linguistic complexity, context independence, and overall suitability). Teachers found the system satisfied the three criteria (between 3.05 and 3.18), disagreement remained within the CEFR level distance.

Overview of the experiments on learner texts

The differences in linguistic complexity vary according to proficiency levels, which influence reading and writing. Anyhow, underlying complexity distribution is different, users usually understand complex structures but they are not able to produce them. Pilán (2018) claims the need to develop a constant training on essay construction accompanied by coursebook data to get better results. For all their experiments of classifying essays written by L2 learners of Swedish into the CEFR levels. Their datasets were error-prone essays written by learners and error-free texts (coursebooks) written by experts, both manually labeled with CEFR levels. The best approach obtained an F1 of .747 and a 2 of .890, which is the weighted combination of L2 coursebook texts and 60 % of Swedish L2 learners' essays. Lexical features were the most predictive measuring the proportion of tokens per CEFR level in the texts.

3.1.3. Vajjala and Loo (2013) approach

Automated assessment (AA) has been a type of technological development used as a method for scoring language skills, including writing. AA originally developed for the English language in stu(Burstein and Chodorow, 2010; Burstein et al., 2003; Crossley et al., 2011; Williamson et al., 2012; Yannakoudakis et al., 2011; Zhang and Liu, 2008). Beyond the mere prediction of learner's proficiency, scholars like Crossley et al. (2014) implemented lexical sophistication indices to define qualitative analysis on proficiency features. National language enterprises, like Swedish corpus classification (Östling et al., 2013) came after English morphological featured approaches.

Previously, Vajjala and Loo (2013) presented a proficiency classification approach for Estonian learner corpus. The percentage of accuracy is 66 percent, and it is divided into three scales A, B, and C. The corpus contained 350 texts per level and used a collection of POS (Part-Of-Speech) and morphological items. A year after, the work was extended adding more featured items and fine-grained corpus. Also, problem modeling used both classification and regression to compare performances.

Corpus and features

The corpus for this research is the Estonian Interlanguage Corpus (EIC), released online by Talinn University. The EIC is a corpus learner of Estonian as a foreign language, obtained from examinations developed by the governmental offices. About 12.000 documents, composed this corpus. The reviewed research took into consideration 879 texts comprising CEFR levels from A2 to C1. Primary code modifications were made in HTMLUnit and Xpath expressions. For the basic statistics, two considerations were adopted, on one hand, take an unbalanced version of the dataset (number of docs and average words); on the other hand, to balance the dataset before experimenting.

Corpus preprocessing

This process POS-tag the texts using TreeTragger (Schmid, 1994), in other words, align information according to parameters of the Estonian data. The morphological disambiguation features from elements within the sentences. A corpus can be divided into the number of words, the number of sentences, mean word, sentence length among others. Morphologically, Vajjala and Loo (2014) describe the complexity of Estonian including, number of nouns and adjectives, the average number of verbs (in mode, tense, declination and conjugations). No syntactic analysis was applied.

Experiments

Vajjala and Loo (2014) highlight the problems at defining proficiency, due to the varying degree of difference between scale levels. Skills and abilities differ in complexity and sophistication without following any measure, even among individuals under similar circumstances.

Evaluation measures

In their research Vajjala and Loo (2014) used multiple evaluation measures depending on choices of learning approaches were used. Among the evaluated aspects, can be mentioned: accuracy prediction, the performance comparison between unbalanced datasets, the Pearson correlation, and Root Mean Square Error (RMSE). Regression prediction showed the percentage of exact matches, the percentage of instances where the prediction is within one-level of the actual value, and the percentage of errors where the prediction is higher than the actual level (Vajjala and Loo, 2014).

Modeling as classification

To compare results from previous experiments in WEKA was implemented the Sequential Minimal Optimization (SMO). Despite this project, kept relation with Vajjala and Loo (2013), its extend is quite different due to the restriction to make a direct comparison with results. Thus, the definition of a baseline was mandatory, they propose a balanced dataset of 92 texts per category. The fully-featured method application on the balanced baseline accuracy of 79percentage, improving from the initial unbalanced dataset. This result forced somehow Vajjala and Loo (2014) to consider using both datasets. The binary classification showed an increase in some scales of CEFR.

Modeling as a regression

As mention by Vajjala and Loo (2013) the CEFR proficiency levels are discrete. Through regression, proficiency prediction can be placed on a scale, is also possible to observe prediction laying between discrete levels. Going further, proficiency prediction is itself modeled as a regression. For this paper, Vajjala and Loo (2013) trained and model a linear regression

in WEKA including some default settings (M5 attribute selection, eliminateCollinearAttributes option set to TRUE) his regression model achieved a Pearson correlation of 0.85 and an RMSE of 0.49. Results could not be compared, due to the lack of this type of research, it has been focused only on classification.

Comparing classification and regression

Vajjala and Loo (2013) presented six comparison items, and they conclude that despite the good performance of classification and regression measuring proficiency, classification has a slight advantage of better performance.

Feature selection

The main question for the authors was how much can we predict with how few interpretable features? Three feature selection methods were chosen: i) information gain, ii) CfsSubsetEval (Hall, 1998), and iii) ReliefFAttributeEval (Kira and Rendell, 1992; Kononenko, 1994). After applying the features to the unbalance dataset (due to its higher accuracy) results showed accuracy levels up to 70percent (information gain 73,5, CfsSubsetEval 78,3percent, and ReliefFAttributeEval 74,5percent). Correlation between features is reported as a paired phenomenon. The highest correlation is between features CTTR and RTTR 0,999. The lowest correlation is between numConj and numInterj -0.623.

Conclusions

Authors finally draw some conclusions as follow:

- Estonian language linguistic model of proficiency showed a prediction accuracy of 79percentage, higher than reports from other languages (German and Swedish). The chosen direction of this research should be continued.
- In analyzing language proficiency two computer mechanisms can be used, classification and regression. Nevertheless, classification has better results.

- As it was proved the high correlation among features, it is mandatory continuing the study about this phenomenon.

- This experiment only considered one single variable in language proficiency, excluding aspects of syntax, discourse, learner errors, the relation of the text to the question asked. Findings must be carefully treated due to this methodological limitation, nonetheless, the proposal proved its value.

3.1.4. E-RATER

TOEFL® is an international widespread test that measures linguistic abilities and skills mostly used in university admissions and different types of migration procedures such as work and residence visas for countries such as USA and Canada.

The exam tries to be simple and consequent by testing four skills divided into four parts: reading, listening, writing, and speaking. However, some sections can be combined. Test-takers may need to read, write, listen, and answer in oral or written form. The exam lasts four hours and a half. Each skill's section proposes some specific tasks to test-takers. In the case of this specific paper, only written skill TOEFL structure will be considered.

ETS[14] has designed two basic exam formats: The TOEFL iBT® test Internet-administered comprises the four academic skills. The revised TOEFL® Paper-delivered Test is the one used in those places where the TOEFL iBT® test can not be taken due to technical limitations. The revised TOEFL® Paper-delivered Test does not include the speaking section because capturing voice technology may not available either. Nevertheless, the writing section does not change between formats, and there is only one structure (See Table **3-1**).

[14]ETS is an American non-profit organization of education experts, researches, and assessment developers interested in assessment designing following an industry-leading insight and an uncompromising commitment to quality. The goal of ETS is to advance quality and equity, pushing institutions towards excellence. `https://www.ets.org/about/who/`

Writing Section	**2 tasks** **1** integrated task based on what is read an heard **1** independent task to support an opinion on a topic **Time**: 50 minutes **20** minutes for integrated task **30** minutes for integrated task Score scale: **0–30 points**

Table 3-1: TOEFL iBT® writing section

ETS points out as design aim for TOEFL iBT® test writing section: "Test-taker can communicate effectively in writing in English language for academic environments" ETS also considers that "essay as the cornerstone text for academic purposes in universities".

Enright and Quinlan (2010) describe E-RATER which is an automated essay scoring system applied specifically to TOEFL iBT® test. ETS has used traditionally two human assessment scores provided by two very well-trained professional raters who read writing tasks independently and score it under rubric assessing parameters.

Enright and Quinlan (2010) point out that ETS remarks about the importance of test quality maintenance expressed by three factors as follows:

1. Scoring consistency or the same scoring criteria for both raters.

2. Scoring reliability (number of tasks, number of ratings per task response and raters' qualifications).

3. Efficiency in the results' delivery as soon as needed.

Certain problems have pushed ETS with creation of an automated tool. Efficiency in rating properly is affected by lack of trained rating personnel, increasing number of test to assess in a short period of time and test quality maintenance. E-RATER uses Natural Language

Processing (NLP) techniques to extract features from writing tasks (essays) to make statistical modeling of human holistic ratings. ETS proposes the use of E-RATER as one of the two ratings for writing tasks for TOEFL iBT® test replacing one of the human raters. This proposal implies that E-RATER takes into consideration the following elements:

1. Evaluation: The score provides the evidence of the targeted skills by the writing task.
2. Generalization: Estimates the expected scores over parallel version of tasks and across raters.
3. Extrapolation: Consistently of scores from others measures of writing ability.
4. Utilization: The use of assessment according to specific educational purposes. Enright and Quinlan (2010)

The analysis of the writing samples using E-RATER focuses on some linguistic aspects as follows: lexical features, grammatical features, syntactic features, and the pragmatic feature about discourse quality (Attali and Burstein, 2006) plus other microfeatures. E-RATER scores the text using statistical procedures e.g.: regression models. Both rating methods, human and automated use the same rubric items, including characteristics from ETS' TOEFL iBT® test aims and goals. (Attali and Burstein, 2006) consider this as a "generic" method due to having the same assessment parameters for human and independent prompts, the similarity in models for psychometric performance and, simplicity in implementation of the method.

The idea of the ETS is that human rating is "ideal" because, trained and expert raters can observe and trace writing features in a natural way, this assumption is supported by a long history of research on this respect, despite automated methods that have not been proved enough. Nevertheless, both rating methods human and automated have advantages and disadvantages dealing with accuracy and efficiency, although the goal is to have a balance between the two methods by having similar rating measures and findings.

E-RATER was applied in the second task of the TOEFL iBT® test (independent task to support an opinion on a topic) a 30-minutes task. Traditionally two human raters score it independently. The experiment replaces one of the human raters by E-RATER. Rubric assessing items remained for both ratings. Similarities and differences were measured. Statistically, some conclusion can be drawn after analysis:

1. Human rating tends to have a higher variability than automated rating, due to individual's discrepancies in the way rubrics are applied.

2. Automated rating with E-RATER have more stable or consistent results at applying rubrics.

3. Human rating focuses more in "wider or macro" features of the writing such quality of ideas and content.

4. E-RATER is quite similar to human raters (similar scores and analysis of rubric's items) in this sense, it has been demonstrated that with the adequate researching effort this method would go beyond any human rating.

5. E-RATER focuses in consistent analysis of the same writing features for all examinees.

6. E-RATER is a complementing tool to human rating.

3.1.5. Other approaches

In this section, other related approaches are reviewed in less detail because they have many commonalities with the approaches described before.

Tack et al. (2017) approach

Tack et al. (2017), (Human and automated CEFR-based grading of short answers) collected a corpus of English short answers question based on the CEFR levels and implemented an approach via a soft-voting classifier integrating a panel of five traditional models: Gaussian Naive Bayes classifier, a CART Decision Tree, a kNN classifier, a one-vs.-rest (OvR) Logistic

Regressor and a OvR polynomial LibSVM Support Vector Machine. They used 695 individual features grouped into 18 different families, among which are: lexical features, syntactic features, discursive features, number of psycholinguistic norms. The best approach obtained an F1 of .495 and an adjacent accuracy of .978. Sentence and word length, lexical features and information about the age of acquisition of words had a strong positive correlation with the assessed CEFR level. Also, the system did not have any particular difficulties in correctly predicting the lowest CEFR levels.

Farag et al. (2018) approach

Farag et al. (2018), (Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input) proposed a method consisting of a corpus of 2, 312 English texts with their CEFR scores, which were assigned by a human expert, five feature types (character sequences, Parts of Speech sequences, hybrid word and Parts of Speech sequences, phrase structure rules, errors, and error rate) and a classification algorithm. The best approach obtained a Pearson r of 0.7654, a Spearman of 0.773 and a of 0.738, with 0.026 of standard error of . This model used the test set (consisting of 260 texts) and the PoS feature.

"My Tailor is rich!" challenge at CAp2018

"My Tailor is rich!" was a Machine learning level prediction competition held in conjunction with CAp2018[15] (Conférence sur l'Apprentissage Automatique), whose task was to predict the English level, according to the 6 reference levels of the CEFR, of written texts between 20 and 300 words. The Organizing Committee provided fifty-nine feature variables, mainly shallow features based on the state of the art of stylometry and language readability. The participating systems produced very high accurated predictions. Despite, a further analysis of the results revealed that the texts contained lexical features that made the classification trivial for some systems (Ballier et al., 2020). For instance, texts labeled with C1 level were prompted by the instruction "Write a movie review". Therefore, the simple identification of words such as "movie", "film", "actor", etc., were accurate predictors of the level. They

[15] http://cap2018.litislab.fr/competition-en.html

concluded that different data is needed and more research is necessary to tackle the problem.

3.2. Assessment of computer programming skills in online forums

Programming computers is an activity that requires several years of study and practice to be mastered. Somehow, learning a computer language resembles learning a second language as in both cases ideas need to be expressed in a new language to achieve a goal. Also, in both cases, learners have at their disposal learning material that guides them through the topics in an increasing degree of difficulty, which was estimated and established by the creators of the materials. Nonetheless, real acquisition of such topics comes through practice. At some point, it is desirable to assess quantitatively the degree of mastery of the new language of the learners, either a computational language or a human language. If the learners are a sizable population, the common answer for that need is to design a test based on learning material.

Because of practical reasons, topics related to computing have pioneered online forums, which can be traced back until USENET newsgroups in the early 80s (Spafford, 1990). These online forums have evolved to collaborative social networks, where skills and competences of the users can be assessed by analyzing the large amount of data obtained from users' interactions (Papoutsoglou et al., 2017).

ExpertiseRank by Zhang et al. (2007)

A seminal work in that topic is the study of Zhang et al. (2007), who aimed to determine if the degree of expertise in the Java programming language can be determined from the dynamics of questions and answers in the Microsoft TechNet newsgroup. To obtain a gold standard for evaluation, the authors selected randomly 135 users with more than 10 posts and hired two independent consultants with high expertise in Java, to rate all questions and answers posted by the selected users. The levels of expertise in which the users were categorized were 5: Top Java expert, 4:Java professional, 3: Java user, 2: Java learner, and

1: Newbie. The inter-agreement of the two raters was $\rho = 0.832$ measured by the Spearman rank correlation. The used data contained 13,739 users (nodes) and 55,761 arcs or edges. An arc from user A to user B means that B answered a question posted by A, that is, B used his/her expertise to help A. They compared different approaches to determine the expertise degree of each user. Two of them used the entire graph, i.e. PageRank (Page et al., 1999) (named ExpertiseRank in this application) and the HITS ("Hypertext induced topic selection") algorithm (Kleinberg, 1999), while the others were simple measures local to each node such as the number of answers a, and the $Z\_number = \frac{a-q}{\sqrt{a+q}}$ (q is the number of questions posted by a user).

In spite of the extensive computational resources used by PageRank and HITS, the simple *Z_number* outperformed them. This rather unexpected result leads them to propose further analysis using simulated data. They found that the TechNet network has a "best preferred" structure, which corresponds to a network where questions are answered preferably by the user with the larger expertise in the topic. Using this heuristic for generating simulated data, they obtained similar results as those obtained in the real data. They also tried a "just better" structure where a question is more probable to be answered by any user with just more expertise than the user that posted the question and less probable by more expert users. In that scenario, ExpertiseRank (i.e. PageRank) outperformed considerably the other approaches, including HITS.

Movshovitz-Attias et al. (2013) approach in Stack Overflow

Stack Overflow (SO) is another question/answer community in the domain of computer programming. This social network has a particular feature that allows users to label answers as "accepted" if the user who posted the question considers the question as appropriate answered. In addition, the community is able to up-vote or down-vote answers as an indicator of their usefulness for the community. Movshovitz-Attias et al. (2013) performed an analysis by comparing the SO's reputation schema based on rules against the PageRank of the users in a setting similar to the proposed by Zhang et al. (2007). They found that allows the

identification of anomalous users having high acquired reputation but low performance, or in other cases suspended users exhibiting "problematic behavior". They concluded that the ranking produced by PageRank is not well correlated with the SO's reputation schema, but it is useful for detecting anomalies. Such a result suggests that PageRank produces more accurate measures of expertise than other approaches based on simple counts of "accepted" labels and differences between up-voting and down-voting counts.

3.3. Contributions from the state of art to this dissertation

The well-structured development of the multidisciplinary academic research has improved innovative approaches to long-term problems in science. In language studies, aims have expanded to different aspects of the learning or the acquisition process, like proficiency (a measurement label) containing abilities and skills from users, learners, or speakers whose L2 is assessed.

The articles and books reviewed are characterized by aiming the language from a quantitative perspective, in the use of statistical and computer technology calculus. In the particular case of the chosen texts conforming the state of the art for this thesis, there are some common factors among them:

The Corpus-Based

The papers revised Lu (2017), Pilán (2018), Vajjala and Loo (2014) and Enright and Quinlan (2010) make use of corpus-based data from a set of different public and private institutions, and the analysis goes directly on the information extracted from the corpus i.e morphological marks or length. In our case, (Silva, 2020) analysis does not rely on the corpus directly but on the interaction among participants of a social network.

The CEFR scales alignment

Three of four revised papers Lu (2017), Pilán (2018), Vajjala and Loo (2014) are aligned to CEFR, considering the description of corpus, analysis, and results in scales from the CEFR. In our case, (Silva, 2020) the CEFR serves as a material for understanding the whole assessment system.

International English proficiency test

Respecting the direct reference of the work to an international English proficiency exam, the only work that makes an explicit reference is the one from Enright and Quinlan (2010). The other proposals restrict the reference to the CEFR or maximum to the simple mention of a specific type of exam. In the present thesis paper, no single examination of exams has been performed.

Assessment criteria

Assessment criteria can be understood in two senses, first, the research describes and analyses assessment methodologies based on previous criteria, i.e international examinations, aligned to the CEFR scales. The second possible sense deals with an alternative proposal, regarding criteria for a new type of assessment criteria. The proposals from Lu (2017),Pilán (2018),and Vajjala and Loo (2014) present statistical criteria for the assessment automatization. On the contrary, Enright Enright and Quinlan (2010) assumes the ETS rubrics as a valid methodology for TOEFLiBT related revisions. In the analysis of YASK (Silva, 2020) some critics of the traditional criteria are presented. Discrepancies deal with the artificiality and distance assessment situations from real contexts of English as an EL2.

Rubric-based analysis

In the CEFR most of the language involved abilities are interconnected with each other. Conversely, for each, a scale is proposed indicating features of reception, production, interaction, and mediation. At a certain point, these scales are prestigious overseas -institutional

prestige from the Council of Europe- then, the international language assessment institutions present their adapted scales or rubrics to compare their descriptors to the test-taker's performance evidence. In the American top-version and wide-spread TOEFLiBT, the CEFR has no direct relation, but also the assessment mechanisms imply the active use of rubrics i.e Enright and Quinlan (2010) work on written proficiency assessment declaring explicitly the use of rubrics in the rating whether human or automatic.

Rating consistency

The language assessment demands humans raters, a particular set of linguistic, educational, and pragmatic skills, to assess test-takers' linguistic performance. In contexts with relatively few numbers of exams, the ratings do not show almost any deviation. However, in the massive tests, rating personnel have to assess a large number of exams in a short period. This heavy load influences negatively rating performance provoking deviation in scores Enright and Quinlan (2010). The use of automated rating methods, eliminates deviation, no matter the number of exams. Additionally, with an appropriate machine learning training or the pre-established route, some embedded and deep structures and phenomena may appear to be studied (Enright and Quinlan, 2010; Lu, 2017; Pilán, 2018; Tack et al., 2017; Yannakoudakis et al., 2011).

Computer technology-based tools

The contributions made by these texts in the description and exploration of diverse computer tools applied to proficiency prediction and measurement, as well as the creation of super-specialized software, to automatized former exclusive human procedures. In this respect, only Pilán (2018) and Vajjala and Loo (2014), used the same tool WEKA.

NLP

The advance in Natural Language Processing (NLP) is mentioned in each text from the state of art. In all these cases, new methodologies to improve and enhance language hidden phenomena take place.

Machine learning

The use of trained computational systems has been constant through the chosen texts. The applicability requires in most cases, the analysis of the huge amount of information, requiring recognition, description, and tagging. Thus, machine learning is relevant to supply those requirements. My thesis also follows the trend and uses machine learning.

Statistical Procedures

Among the procedures, the ones with more mentions in these texts are:

- Factorial analysis
- Minimal edit distance
- Readability index
- SVM
- Spearman Correlation
- Pearson Correlation
- Sequential minimal optimization

Due to the nature of our approach, my thesis has in common with the state of the art the Spearman and Pearson correlations and minimal edit distance, three being very discrete.

Language content-based

All the reviewed texts are language content-based. They revise the interaction among elements in the word or sentence. Hence, morphological, syntactic, grammatical, pragmatic, and discursive tend to be analyzed. On the contrary, my proposal deals with the social network as the legitimation mechanism to consider plausible a written construction.

Word and sentence levels of analysis

In the state of art, there are explicit qualifications expressing both word and sentence as a valuable level of computer-based analysis. In my proposal, the nature of the corpus is sentence level.

Writing quality

The written proficiency is an extended ability that implies in users the performance of different tasks coordinated towards coherence and correctness. The way some particular text should be is prescribed by factors such as style and format, country linguistic variety. etc. The corpora resources used in three ((Enright and Quinlan, 2010; Lu, 2017; Pilán, 2018) of four review texts, the quality of writing is examined, by the computing of data and the use of experimental tools. Besides, the aim of the research includes assessing.

Expertise in questions-answers networks

The approaches presented in section 3.2 provide enough evidence to the hypothesis that recursive measures based-on social graphs such as PageRank can produce meaningful rankings of expertise in collaborative communities focused in a knowledge area. In particular, Zhang et al. (2007) show that the PageRank approach is considerably better in a network where the dynamics of the answers are not exclusively provided by the top-expert users. When this result is translated to an educational setting, a top-expert user corresponds to the tutors, while users of intermediate expertise could be associated with peer learners. This result jointly with the suggested analogy support our decision of using PageRank as the basis for the new method for proficiency assessment presented in this dissertation.

Another particular contribution of the works reviewed in section 3.2 is the fact that the current approaches have shown effectiveness in the identification of users in the higher levels of expertise. This situation when is transferred to an educational scenario becomes an important limitation because the identification of all levels of expertise/ability in a subject is a must-have feature. Therefore, the new method presented in this dissertation includes

modifications to the PageRank algorithm aimed to address this issue.

Additional clarifications

As a whole, the state of art allows the understanding of some conditions of research on automated assessment and scoring. The influence in the supranational policies (European CEFR, Australian Threshold Standards, and American ETS standards) conceives a system in which language-related knowledge is the key to a set of social requirements. Production systems and industries force people to look for the acquisition of certain skills aligned with their interests. Proficiency measures the alignment of personal skills to the institutional expected ones. The English language standards, in writing, for instance, are not permitted to evolve progressively, because adaptation from institutions to innovations is slow, and to preserve their authority as prestigious centers. Any revision is made then, on authorized language self-contain principles and rules. letting behind innovations from users whether native or non-native.

The main utility from this state of the art is to prove the novelty of this dissertation and the fact that the proposed approach offers an alternative to the language self-contented assessment, returning on the speech community concept presented in subsection 2.2.

4 Problem statement

This thesis focuses on the discussion about English language proficiency prediction as an automated process. Nevertheless, the proficiency concept moves across the English as a discipline, through educational policies, language acquisition, teaching (methodologies, resources, strategies), learning, scoring, and testing. As a matter of fact, proficiency goes beyond language disciplines to academic and professional levels.

4.1. Problematic situations

4.1.1. Language policies

The worldwide English mainstream is based on the CEFR, a public policy document for the European Union (COUNCIL, 2018). In the European political, economical, and educational environment, a proposal with these features is highly probable to be accomplished. Despite a few countries in Europe [16] Notwithstanding, in most countries of South America, Africa, South Asia conditions are dissimilar from the conditions related to the CEFR. These poor countries face a lot of obstacles in their education systems. The CEFR is applied in these countries without taking any further consideration, examination centers continue scheduling tests, and courses.

The adaptation of the European education principles in alien environments without minimal

[16] In Europe, some countries, like Greece and Estonia, are poor in comparison to the European average, then the CEFR and other policies' agreements could be more difficult to implement.

applicability conditions could drive the efforts to a waste of resources.

4.1.2. EFL vs ESL

As mentioned above, English as a Foreign Language (EFL) and English as a Second Language (ESL) are different teaching and learning environments, conditioning the acquisition mechanisms (Schauer, 2006). Differences deal with contexts of use (real contexts vs classroom contexts), frequency of use, and amount of knowledge and skills developed. The communicative competence of a person living in an EFL context is lesser than the communicative competence whose second language was English. The English international tests, do not consider differences between ESL and EFL for presenting different versions of the tests, classifying test-takers or their results, which is an evidence of the inadequacy of the English international tests.

4.1.3. The artificiality of language tests

Language proficiency (the ability of non-native users) develops gradually by using the language in real contexts. Real-life situations bring language to users' minds naturally. In ESL contexts students usually practice English outside the classroom, during the daily activities. In the EFL contexts must be artificially recreated in the classroom, auditorium, or virtually. The results are different in both groups. EFL processes tend to take longer and focus only on certain language aspects, while ESL tends to spread attention over several aspects of language.

The use of internet-based technologies for sharing all type of information, facilitate users to get in touch with native users, or at least with higher proficiency level users. By means of these interactions, users can sharpen their skills improving and compensating access to real-English environments.

Authenticity also is linked to those uses of language-related skills at developing roles and activities belonging to personal, professional, and academic life. For instance, a linguist doing her job probably writes papers (essays, reviews, e-mails) on dialect varieties in South Ame-

rica, listening to audio recordings from fieldwork or interviews (ethnography), read papers about phonology, and speak to colleagues or students about new tendencies in the study of linguistics geography. Artificiality, in the former example, could ask the linguist to present a language proficiency test using materials from molecular biology in the reading part, a listening sample about the roman sculptures preserved in German museums, writing about global warming influence in Canadian crops, and a speaking discussion about aircraft crisis post-COVID-19.

4.1.4. Applications (APPs) for learning foreign languages

Internet services have a huge impact on education, connecting institutions and students worldwide. Besides in language learning, many applications (APPs) were produced intended to allow users to practice multiple aspects of languages (grammar, reading comprehension, writing short and long texts, speaking, and listening) individually throughout artificial intelligence interfaces or in collaborative communities.

It has been a tendency, in constant growing, the use of APPs to study and practice languages, usually freely. APPs like Duolingo (Von Ahn, 2013) and Busuu have millions of users around the world. These companies have been updating their APPs' interface constantly, adding new utilities and services. The educational APPs use gamification [17] routes for addressing users in a "learning route"just like traditional games do with their iconic Italian plumber or the giant Gorilla (Pardoel and Athanasiou, 2019). These interfaces looks are more likely to internet average user, due to the entertainment visual interfaces (games, social networks, blogs, and web pages). As the industry has built a standard in English language business, education research should take into consideration its impact and develop innovations from this point on.

[17] Gamification is the process of developing activities of non-game contexts, using gaming design elements (Sailer et al., 2017)

4.1.5. Online English Tests

The current trends in assessment deal with computer-based and online tests, which enable test-takers to present exams easily. International English language testing top-leaders institutions have created alternative versions of their paper-based exams into computer-based (to maintain accessibility and control over materials). Still Online education companies focused on APPs have introduced complete online tests, free or underpayment [18], accessible everywhere. These types of exams are already accepted in different renowned universities all around the world[19]. In the years to come, the English language business will grow even stronger on the development of automated assessment to cover the growing demand for these kinds of proficiency certification tests.

4.1.6. Rubric Based-assessment

In the international language tests, assessment is a standardized process, dealing with a rubric system of language abilities descriptors. The parameters are classified into levels of proficiency, from beginners to advanced users. Wind (2020) reviews the use of different methods in the evaluation of rating scale functioning, like the Rating Scale Model (Andrich, 1978), Partial Credit Model (Masters, 1982), and Many-Facet Rasch (Linacre, 1990). All these evaluation methods proposed models to illustrate relationships among raters' judgment of test-takers' abilities. Wind (2020) highlights the overlapping of language aspects and analysis units in rubric assessment systems. There are intricate relationships among the abilities, thus they can not be considered separately. Rating an aspect affects the rest. An effective assessment requires very well-trained personnel managing those rubric's standards (Enright and Quinlan, 2010).

Positive aspects of rubrics in educational processes have been reviewed by Panadero and Jonsson (2013), transparency in the assessment, provide students with exact feedback, encouraging them to learn self-regulation, and to reduce anxiety. These findings imply full

[18] `https://englishtest.duolingo.com/home`

[19] `https://englishtest.duolingo.com/institutions`

knowledge of criteria in the process by students and teachers as well. Nevertheless, in the international language exams, due to the protection of test-related materials, this information is reserved. Conversely, transparency can not be claimed.

Sadler (2014) pointed out the problem of codification in rubrics (i.e. A1, B2, C1), which are confusing to students' understanding [20], plus the multiple factors composing such as category. Panadero and Jonsson (2013) listed three main criticisms in the use of rubrics as follows: i) Standardization of assessment through rubrics, ii) Rubrics narrowing the curriculum, and iii) The reduction in variability of scores. The standardization of assessment through rubrics implies a narrowing of learning environments, a learning aims readdressing to the categories included in the rubrics. The displacement of contents and skills to fulfill new standardized needs.

Rubrics narrowing the curriculum has been understood as the cycle in which curriculum and rubrics feed each other in a continuum, contents, and skills developed in processes are the same ones in the rubrics and vice-versa. The reduction in variability of scores deals with a limitation in the scope of the variability of scores (Mabry, 1999). Humphry and Heldsinger (2014) mentioned a halo effect produced by an equal number of performance levels in a rubric (results could increase or decrease fewer levels).

4.1.7. Online speech communities

In traditional linguistics, language tends to evolve through speech communities, which usually are shaped by geographical, economic, ethnic, and cultural ties. Speakers build an identity according to their environment. The speech communities have two basic effects on language, on one hand, they could serve as barriers against external influences, preserving their linguistic variety. On the other hand, they could serve as ignitors to linguistic change by

[20] According to Sadler (2014), the terminology used to describe overall performances i.e: integration, accuracy, consistency, are words with multiple and wide semantic interpretation. Students barely fully understand the aim of the descriptions.

influencing weaker surrounding speech communities.

Nowadays, besides the traditional speech communities, the internet has brought to life the concept of an online speech community (Morgan, 2014) where speakers gather through internet environments under special circumstances of mutual identity (profession, education, social groups among others). Milburn (2015) presented some addressing questions which are expected to confirm the existence of online speech communities, which can be applied to the specific case of YASK's community:

- Is the community interested in examining their communication? (Yes)
- Is the community interested in language use? (Yes)
- Are the community participants using or developing a specific set of rules for their interactions? (Yes)
- Do participants have to share a common language or way of speaking, considered appropriate? (Yes)
- Is the community sharing a common goal or objective? (Yes)

YASK speech community is interested in developing collaborative strategies to learn an L2 (English for this case). Users participate by posting texts, while other participants assess texts by using likes or dislikes. Also, the posts can be corrected if they are wrong. Thus, YASK community performs a consistent examination of commoners' language. The goal of YASK is to develop L2 abilities in the users by the permanent interaction. The community is composed of native speakers, advanced speakers, proficient speakers, and beginners. Everyone interacts according to his/her abilities. Interactions can be direct or indirect but, they are expected to follow the international standard of written English.

4.2. Justification

The factors listed above affect methods and procedures to obtain a proficiency measure or rating. In the case of writing skills, the presence of some specific descriptors, predicts a profi-

ciency scaled-level. This paper presents some basic assumptions about the measure or rating procedures and methods by considering the online social networks (comprising online speech community/community of practice) as legitimated groups able to judge and rate participants writing pieces. An innovative perspective in which group participants evaluate the accuracy of language as a whole, having the option to vote positively or negatively [21]. In the articulation of the proposal, some theoretical concepts regarding communities' linguistic behaviors, their ability to make decisions, and predict in a smarter way that individual experts and prestige rankings with communities networks.

The further desire of this document expects to encourage researchers to find new alternatives to develop assessment and rating proficiency in an automated way by considering additional methodologies apart from rubric-based ones. As in traditional Linguistics, social networks, speech communities, and communities of practice legitimate, enhance and even rule language variation uses, online counterparts could provide real assessment and rating on linguistic performance in the given community.

4.3. Research questions

The current study was conducted to answer the following research questions:

1. What is the extent of the relationship between the degree of the reputation of users in a collaborative social network dedicated to the practice of languages (online speech community), with their language proficiency (if it exists)?

2. Which is the relative importance of positive/negative and implicit/explicit information (votes) extracted from the social graph to assess proficiency?

3. How does social media-based proficiency assessment approach collate to the traditional CEFR approach?

[21] YASK app interface and utilities are in constant updating, it has introduced extra functions comprising text operations, listening skills and speaking recording.

4.4. Main Objective

- Determine a method of assessing writing proficiency with an appropriate level of reliability from votes positive/negative and implicit/explicit votes including concepts such as the wisdom of the crowd, social networks, speech community, and community of practice.

4.5. Specific Objectives

- Look into the importance of positive/negative and implicit/explicit votes.
- Compare traditional writing proficiency assessment models to a model based on the PageRank algorithm.

5 Methodology

5.0.1. Method route

From a pedagogic perspective, current approaches in the language proficiency assessment and rating require a well-structured organization divided into stages as follow: i) instrument preparation (questionnaires and exam construction), ii) test-taking date (recollection of test-takers feedback); iii) questionnaires and exams rating; iv)scoring consolidation of assessment according to scopes. Finally, v) results in alignment with international proficiency macro-scales (see section 2.6). Scoring and assessment use statistical and computer-based methods (see 2.6.6 and 3.1.4) under a rubric-based approach, made by professional personnel.

Conversely, the method proposed through this paper presents an innovative linguistic approach (see sections 2.1,2.2 and 2.3) since proposes a separation from traditional approaches of rating and assessing language proficiency. My method understands the online social network as the linguistic development environment, where the speech communities and the communities of practice (depending on the participants' background) monitor the linguistic practice interaction, addressing and judging written posts made by the rest learners/users. These collaborative exercises have had a long tradition in the QA sites on the internet, such as Stack overflow (see section 2.10).

After obtaining data from the online social network (YASK in this case), statistic procedures and the PageRank algorithm are applied to define the relations among users, not to measure the language itself. The major difference in my methodology proposal bases the analysis on the user's prestige within the network. YASK's users are legitimated to rate others' written

posts, due to their linguistic competence and proficiency level. Nonetheless, after ratings, community members can confirm or discard their validity and accuracy.

5.1. Dataset Description

The data used in the experiments was extracted manually from YASK using its application for smartphones. First, a native of the Spanish speaker created a user account in YASK and interacted actively during approximately one month, by posting answers to other users' requests in English. Next, I recorded all the users and votes to all of my answers. Finally, I selected randomly 10 users[22] (among the users recorded in the previous step) having English between beginner to advanced levels and recorded all posts, users, and votes related to their requests. To carry out this process, I obtained authorization from YASK representatives, and the data was anonymized to protect the privacy of users.

The result was a graph with 377 users (nodes), 1,571 positive votes (links), and 490 negative votes (links). The English proficiency level of the users in the data, as established by themselves, is distributed as follows: 69 'Native', 52 'Fluid', 66 'Advanced', 140 'Intermediate', and 50 'Beginner'. The number of requests asked by the users is 179, while the number of answers to those requests is 412. These 412 answers were made by only 107 users, which are the only ones able to receive explicit votes. Approximately, the 50 % of the answers were posted by 10 users, among them are the users labeled as 'Google Translate' and the 'Yask Bot', which is an automatic response of previously answered requests in Yask. This observation confirmed our assumption that users that contribute with answers are a minority in comparison with those that contribute with votes. This scenario is known as "Participation Inequality", where "90 % of users are lurkers (i.e., read or observe, but don't contribute),

[22]The categorization of users in YASK comprises people worldwide who are searching for language skills in daily life environments. As the whole context based on collaborative ties among users, each participant gives support on their native language and receives support from other's language native speakers. These network interactions allow YASK users to acquire skills under native's speaking context perspectives (`https://yask.app/en`).

9 % of users contribute from time to time, but other priorities dominate their time, 1 % of users participate a lot and account for most contributions..." (Nielsen, 2006). The process of extraction was performed in January 2019.

The data also contains the texts of the requests and answers written by the users. The average length of these texts is 42.4 characters with a standard deviation of 26.5 characters.

5.2. Proposed Method: ProficiencyRank

My method, ProficiencyRank, consists in building two adjacency matrices M, one for positive votes and another for negative. Thus, each time an answer posted by a user A receives a vote from a user B, I draw a directed edge from the node B to node A either in the graph of positive votes or in one of the negative votes. Next, I apply the PageRank algorithm separately to each one of the graphs to obtain two ranking vectors r^+ and r^-. Then, when the two PageRank run converge, their results are linearly combined using a parameter that controls the weights of the positive r^+ and negative r^- rankings.When ProficiencyRank, represented by $\alpha = 0$, only the positive votes are taken into account; when $\alpha = 1$, only the negative votes, when $\alpha = 0.5$ both have the same importance, and so on for other intermediate values of α. The rankings obtained from the graph build with positive votes should increase according to language proficiency. Conversely, the rankings obtained from the build graph from negative votes are related inversely with proficiency. Therefore, the linear combination of both rankings must be made with a negative sign. Thus, the resulting ProficiencyRank vector r, containing the ranking values for each user in the network, is defined as:

$$r = (1 - \alpha) \times r^+ - \alpha \times r^-; \alpha \in [0, 1] \tag{5-1}$$

There is a limitation in the use of PageRank, that is, that the rankings can be determined only for those users who post answers, that is, users having incoming votes (Galeotti et al., 2010). In Figure **5-1** that corresponds to users B and D. For the remaining users that only

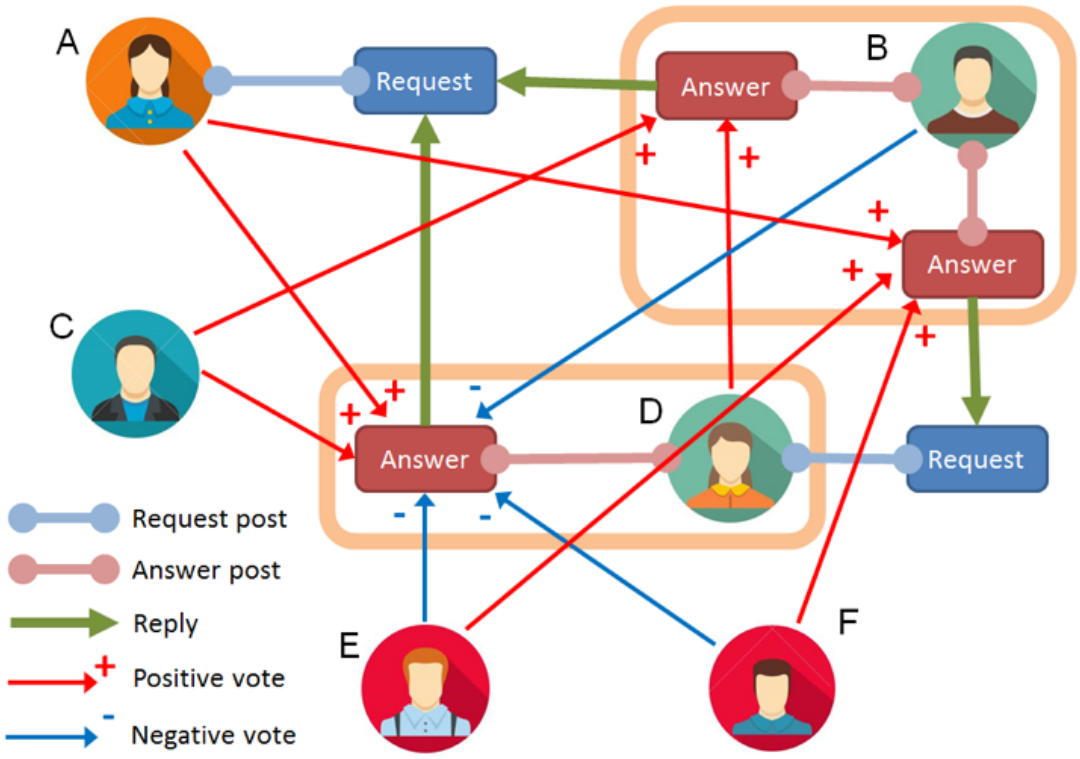
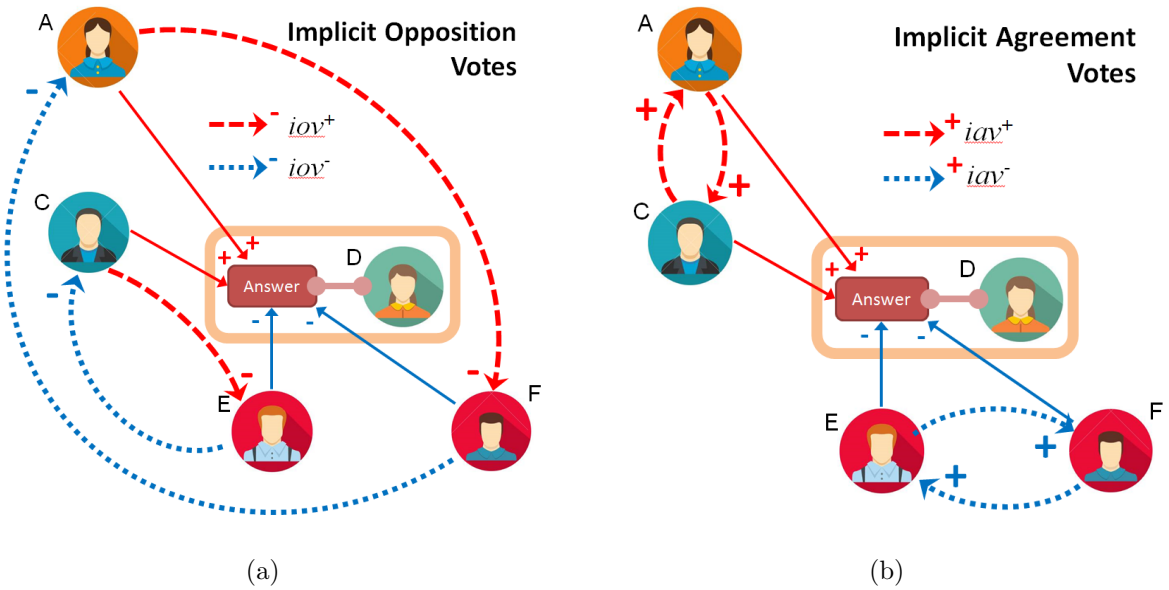


Figure 5-1: Example of a collaborative social network.

vote, the PageRank algorithm assigns a single minimum value. Generally, the users that only make votes outnumber significantly those who post answers making that the rankings can be obtained only for a small subset of users. To overcome this issue, we extracted implicit votes from the graph by inferring new votes from agreements and disagreements between users.

For instance, in Figure **5-1**, user C and E voted contrarily the answer posted by D. Then, aside from the explicit votes of C and E toward D, it can be considered that users C and E mutually oppose producing two implicit negative votes between them. We call these votes Implicit Opposition Votes (iov). We distinguish them as iov^{+}, the implicit negative vote from C to E (given that C voted positively), and as iov^{-} the implicit negative vote from E to C (given that E voted negatively). Figure **5-2a** illustrates this concept in our running example.

I distinguish the implicit positive votes between A and C as iov^{+}, and those between E and F as iov^{-}. Figure **5-2** illustrates that. By considering *iovs* and *iavss*, the number of users in



(a) (b)

Figure 5-2: Examples of some Implicit Opposition Votes and Implicit Agreement Votes inferred from the collaborative social network in Figure **5-1**.

the graph having incoming votes increases considerably, making possible the computation of their ProficiencyRank.

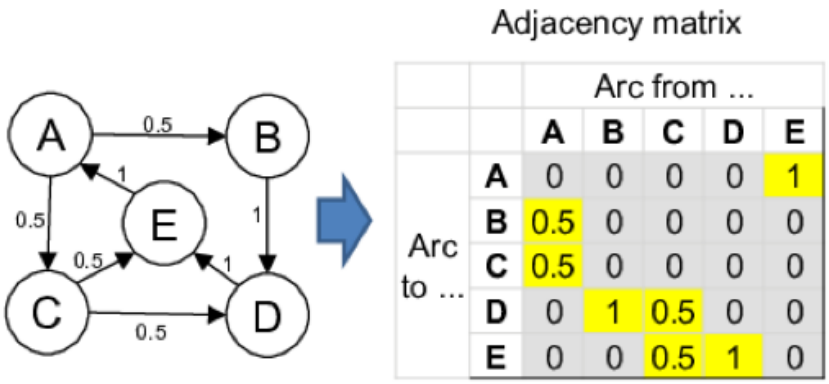


Figure 5-3: An example of a weighted graph and its adjacency matrix.

Similarly, users A and C agree positively in their votes, as E and F agree negatively to the same answer. These agreements produce what we call Implicit Agreement Votes (**iav**), which in this case produce mutual positive votes between A and C, and between E and F.

To integrate implicit votes in the computation of ProficiencyRank, we build the adjacency matrices $\mathbf{M}_+$ and $\mathbf{M}_-$ as combinations of explicit and implicit votes using again weighted linear combinations:

$$\begin{aligned}\mathbf{M}_+ &= (1-\beta) \times \mathbf{M}_{exp^+} + \beta \times \mathbf{M}_{iav};\ \beta \in [0,1]\,; \\ \mathbf{M}_- &= (1-\delta) \times \mathbf{M}_{exp^-} + \delta \times \mathbf{M}_{iov};\ \delta \in [0,1]\,.\end{aligned} \tag{5-2}$$

Here $\mathbf{M}_{exp^+}$ is the adjacency matrix build using the explicit positive votes, and $\mathbf{M}_{exp^-}$ the equivalent with explicit negative votes. Similarly, $\mathbf{M}_{iav}$ is the adjacency matrix build using the *iav*s, and $\mathbf{M}_{iov}$ the equivalent for *iov*s. It is important to note that the entries of all the $\mathbf{M}_*$ matrices contain the total number of votes given by the users indexed by the columns, towards the users indexed by the rows, then the columns are normalized to sum up 1 (see Fig. **5-3**). This differs from the traditional setting of PageRank, where several links from a node A to B are treated as a single one. Parameters β and δ work similarly to α and their values also vary between 0 and 1. In addition, the matrices in Eq. 5-2 need to be transformed for PageRank algorithm by applying the damping factor d using the following equation:

$$\hat{m}_{i,j} = d \times m_{i,j} + \frac{(1-d)}{n} \tag{5-3}$$

In summary, our method has 4 parameters, namely: d the damping factor of PageRank, α the weighting parameter between positive and negative votes, β the weighting parameter between explicit and implicit positive votes, and δ the analogous for negative votes. Figure **5-4** depicts a summary of the computation of ProficiencyRank. Note that for the construction of $\mathbf{M}_{iav}$ it is necessary to determine which combination of iav^+ and iav^- should be used (analogously for $\mathbf{M}_{iov}$).

5.3. CERF Baseline

I provide an additional test based for comparison that reflects the methods of the current lang juage teaching curricula. For that, we used the English Vocabulary Profiles for the

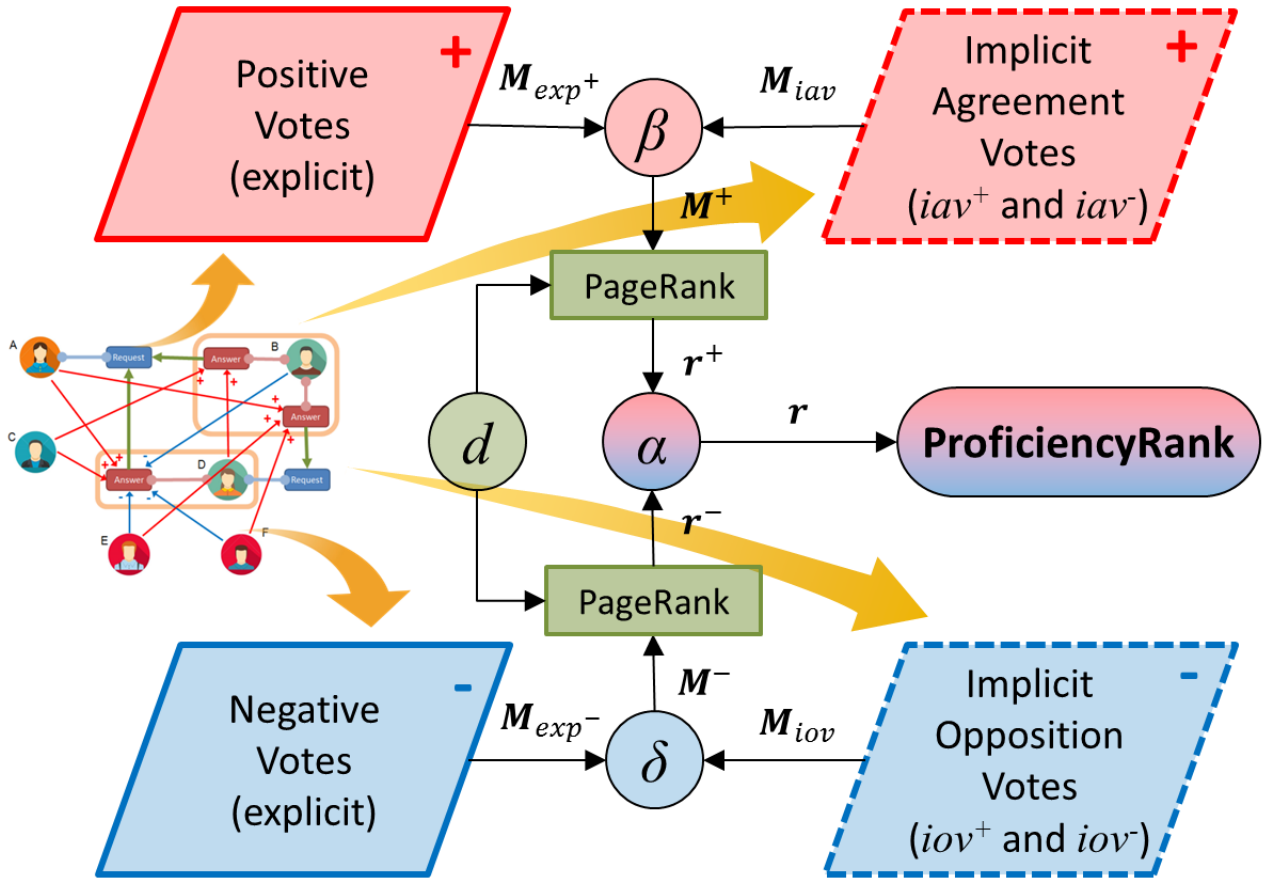


Figure 5-4: Flow chart of the ProficiencyRank method.

Common European Framework of Reference (CEFR) provided by *EnglishProfile*[23], a non-profit organization devoted to produce resources for teaching English aligned with the CEFR levels (i.e. A1, A2, to C2). They provide manually curated word lists obtained from the Cambridge Learner Corpus (Nicholls, 2003), which represent the vocabulary profile for each CEFR level[24].

I used the texts of the answers written by the Yask's users in combination with the CEFR vocabulary profiles to determine the level for each user. For that, I obtained all $w_{u,l}$, which is the number of common words between the set of words derived from the answers written by user u and the vocabulary profile corresponding to the level l. Since, the CEFR levels are meant to correspond to a linear progression of proficiency in English, we assigned increasing

[23] `https://englishprofile.org/wordlists`

[24] We use the lists compiled at `https://www.toe.gr/course/view.php?id=27`

Table 5-1: Number of common words between the English Vocabulary Profiles obtained from the Cambridge Learner Corpus for the CEFR levels.

	A1	A2	B1	B2	C1	C2
A1	541	217	214	188	156	197
A2	-	1038	415	385	262	348
B1	-	-	1806	731	394	533
B2	-	-	-	2495	536	717
C1	-	-	-	-	1701	575
C2	-	-	-	-	-	2182

weights to each level. Thus, the proficiency level P_u for a particular user u is computed with a weighted average as follows:

$$P_u = \frac{\sum_{i=1}^{6} i \cdot w_{u,l_i}}{\sum_{i=1}^{6} w_{u,l_i}} \tag{5-4}$$

Where l_1 corresponds to the level A1, l_2 to A2, until l_6 to C2. Therefore, P_u is a number between 1 and 6. To evaluate the soundness of this baseline, I applied Eq. 5-4 to the 27,306 texts from the CAp2018[25] training dataset[26]. The obtained values were compared against the gold standard levels in the same dataset. We observed a Spearman r rank correlation of 0.65 with $p < 0.0001$. Clearly, the proposed baseline represents the CEFR levels of English proficiency.

[25]meeting of the francophone Machine Learning community.

[26]`http://cap2018.litislab.fr/competition-en.html`

6 Experimental Validation and Discussion

Our experiments aim to address two questions. First, to what extent the ranking methods presented in Chapter 5 are correlated with the English proficiency level of the users in YASK. Second, how much user interaction in YASK is needed to measure adequately the English proficiency level of the users.

6.1. Experimental Setup

The gold standard for comparing and assessing the rankings produced by the presented methods is the English level that users manifested freely when they signed up into YASK, which I assume to be true (see 5.1. This gold standard can be considered as a holistic self-evaluation, which has shown to be accurately correlated with proficiency (Liu and Brantmeier, 2019). I replaced the categorical levels by a simple numerical scale as follows: 5 for 'Native', 4 for 'Fluid', 3 for 'Advanced', 2 for 'Intermediate', and 1 for 'Beginner'. The evaluation measure to compare the degree of agreement between the gold standard and the produced rankings is the Spearman's rank correlation . Apart from Spearman's measure (between -1 and 1) in our work ranges come from 0 (no correlation) to 1 (perfect correlation) and uses ranks instead of values, which means that it reflects to what extent ProficiencyRank can order individuals from "native" to "beginner". Ties were handled by the average of the ranks that would have been assigned to all the tied values by using the implementation of the Spearman's correlation provided by the scipy[27] package.

[27] `https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.spearmanr.html`

Each particular configuration of ProficiencyRank consists of a selection of types of positive and negative votes. Positive votes can be a combination of a selection from $\mathbf{M}_{exp^+}$,$\mathbf{M}_{iav^+}$, and $\mathbf{M}_{iav^-}$. Similarly, negative votes come from $\mathbf{M}_{exp^-}$,$\mathbf{M}_{iov^+}$, and $\mathbf{M}_{iov^-}$. Once the types of votes to use have been established for a configuration, the parameters $[d, \alpha, \beta, \delta]$ are determined in a search grid with a resolution of 0.1. Then, the grid resolution was increased to 0.05 in the vicinity of the current best configuration, and again the resolution is refined until 0.01. The function to optimize is the average of Spearman's r correlations between ProficiencyRank and the gold standard for different subsets of users filtered by a threshold of θ representing the minimum number of incoming votes. θ is incremented from 1 to the maximum number of votes obtained by the top-voted user. Clearly, as θ increases the number of users that surpass what threshold reduces. For the average calculation, I considered only significant correlations with $p < 0.01$.

Table **6-1** shows seven possible configurations that I consider interesting to discuss. The last row reports the total number of votes on each category of the type of votes found in the data. The baseline method consists of the total number of incoming votes per user for each configuration. This measure quantifies the possible undesired effect that the amount of activity from users in the network being correlated with their language proficiency. The results for all the 377 users for each configuration can be seen in the last column in Table **6-1**.

With regard to the reliability of the judgments, I verify the global agreement between raters in the data using the Krippendorff alpha measure due to its robustness against missing values, which are very common in our data. For instance, in conf6, which has the least number of missing values, the rate of non-missing entries is 0.84 % (i.e. $100 \times 11,937/377^2$). For each configuration, I computed Krippendroff's alpha in the weighted matrix of votes M defined by: $\mathbf{M} = (1 - \alpha) \times \mathbf{M}_+ + \alpha \times \mathbf{M}_-$.

Since $\mathbf{M}$ is a square matrix, each rater (a YASK's user) is represented by a row, which con-

		Positive Votes					Negative Votes					
Conf.	d	exp^+	β	iav^+	iav^-	α	exp^-	δ	iov^+	iov^-	Votes	Baseline
conf1	0.86					0.79					2,061	0.186*
conf2	0.80		0.90			0.78		0.40			3,873	0.011
conf3	0.85					0.85		0.15			3,393	-0.085
conf4	0.98		0.40			0.39		0.74			10,125	-0.147
conf5	0.90		0.10			0.65					10,605	-0.094
conf6	0.85		0.14			0.66		0.15			11,937	-0.126*
conf7	0.89		0.53			0.85		0.20			4,539	0.005
Num. of votes:		1,571		7,398	1,146		490		666	666	11,937	

Table 6-1: The seven configurations of ProficiencyRank used in the experiments with their optimal set of parameters.

* significant $p <= 0.05$.

tains the weighted votes given by the rater to the other users. Table **6-2** shows the results for each configuration. The highest degree of agreement is obtained by $conf1$, which is a weighted combination of positive and negative raw votes. Since the remaining configurations manipulate agreements and disagreements through implicit votes, the agreement between raters measured by the Krippendorf's alpha is expected to vary considerably. In fact, ProficiencyRank exploits disagreements to discriminate experts from novices, where experts tend to agree with each other, while novices disagree because of their lack of proficiency. Therefore, a relatively low score of inter-rater agreement on the raw data ($\alpha = 0.456$) seems convenient for our purposes.

	conf1	*conf2*	*conf3*	*conf4*	*conf5*	*conf6*	*conf7*
Krippendorf-alpha	0.456	−0.054	0.283	0.011	−0.268	0.283	0.328

Table 6-2: Inter-rater reliability measured using Krippendorf's alpha.

6.2. Results

Figure **6-1** shows the results obtained by the Proficiency Rank configurations from *conf1* to *conf7*. The vertical axis corresponds to the Spearman's rank correlation r between the ProficiencyRank produced by each configuration versus the gold standard measured in a subset of the users filtered by θ (horizontal axis). In figure **6-1**, all seven configurations increase as θ increases. The total number of votes considered for applying the threshold θ varies on each configuration. For instance, configuration *conf1* considered only 2,061 votes $(1,571 + 490)$, and *conf6* used all the 11,937 available explicit and implicit votes. As θ increases, the number of users that fulfill what threshold decreases. Figure **6-2** shows the same results but replacing the abscissa θ by the number of users, producing a decreasing tendency for all configurations. In general, the best configurations are those that shape the upper-bound in both figures. Note that all configurations outperformed their corresponding baselines (shown in the last column in **6-1**) by a wide margin.

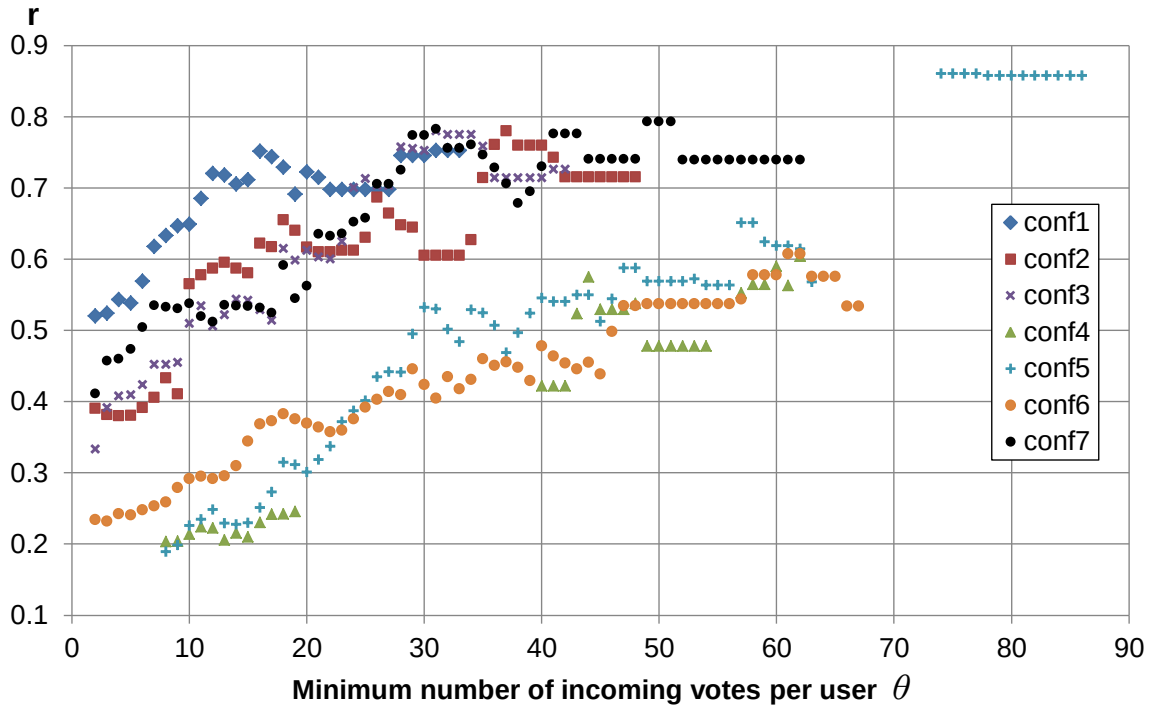


Figure 6-1: Results of the tested Proficiency Rank configurations for different sets of users having at least θ incoming votes.

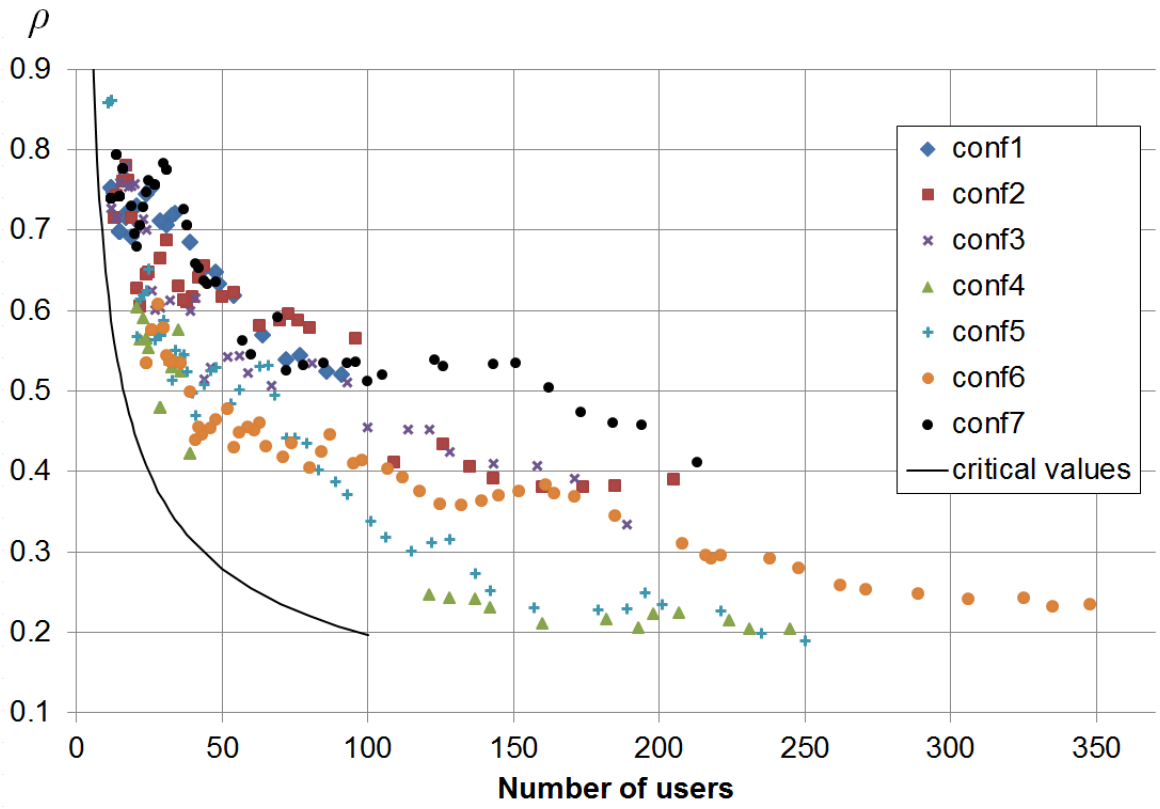


Figure 6-2: Results of the tested ProficiencyRank configurations for different sizes of sets of users. The "critical values" series depicts the critical values for the Spearman's rank correlation for nondirectional $alpha = 0.05$ levels computed by Ramsey (1989).

Figure **6-3** shows a comparison of the results obtained by Proficiency Rank and the CEFR baseline proposed in Eq.5-4 (see 5.3). This comparison is only possible against *conf1* because the users who received votes are the only ones who wrote answers. Thus, the P baseline for each user (Eq.5-4) is computed by aggregating all the answers written by each user. In addition, this figure includes a line with the critical values for the Spearman's r rank coefficient for $\rho = 0.05$. This figure also depicts the effect size of our results showing that as the number of subjects considered in the analysis increases the correlations between the measured proficiency and the gold standard keep a considerable, and rather constant, margin from their critical values.

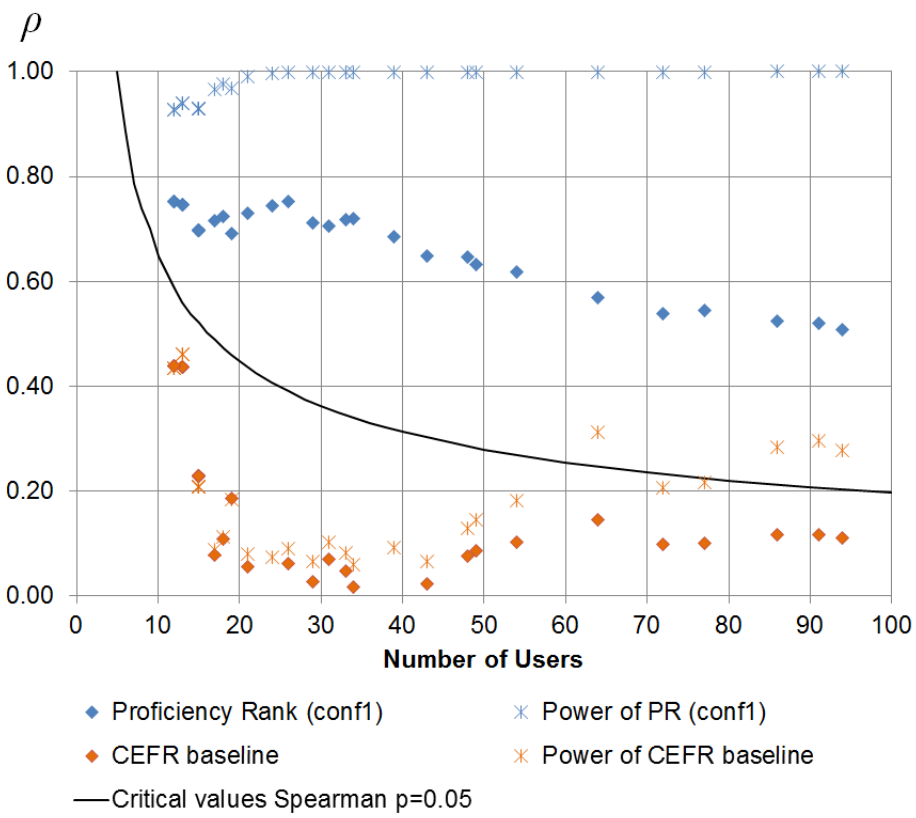


Figure 6-3: Results of Proficiency Rank *conf1* in comparison with the CEFR baseline (see 5.3). The power of the test of statistical significance is paired vertically to each result depicted by an asterisk in the same color. The power of the statistics was computed using G*Power software[28]. The value of the probability of an error type-I error is $\alpha = 0.05$.

6.3. Discussion

6.3.1. Analysis based on experiment findings

In general terms, the margins between the results of all studied configurations and the critical values in Figure **6-2** is a strong signal that ProficiencyRank is considerably related to the self-assessed proficiency level of the users in a collaborative social network like YASK. This provides a clear affirmative answer to our first research question.

[28] `http://www.gpower.hhu.de/`

Regarding the second research question, let us discuss the results obtained by *conf1*, a configuration composed only of explicit votes, therefore the one with the least number of votes. *Conf1* achieved the best results when varies between 1 and 25. This result indicates that explicit votes are the strongest signal in the social graph. Nonetheless, explicit votes are in short supply producing only significant Proficiency Ranks for only a maximum of 91 users out of 379. Figure **6-2** shows that other configurations produce Proficiency Ranks for far more users. The optimal value for the parameter d, when using *conf1*, coincides with the default value for the equivalent damping parameter for PageRank ($d = 0.85$) (Page et al., 1999). Parameter $\alpha = 0.79$ indicates that, despite the number of explicit negative votes is relatively small ($|exp^{-}| = 490$), they weight much more than the explicit positive votes ($|exp^{+}| = 1571$).

In addition to the shown experiments, I tested the results of PageRank for positive and negative votes separately. Nonetheless, no significant correlation was observed ($\rho < 0.01$) for any possible value of θ. In the first test (only votes in exp^{+}), the number of explicit positive votes seemed to be sufficient for 377 users, but their lack of informativeness could explain the poor results. Contrarily, the explicit negative votes are highly informative, but only 490 edges for 377 nodes produce a very sparse graph. This result proves that our method for combining the positive and negative votes using a linear combination controlled by is effective in comparison with PageRank using alternatively positive or negative votes.

Figure **6-1** shows that configurations *conf1*, *conf2*, *conf3*, and *conf7* outperformed the others. These configurations have in common the use of iov^{-} and the disregard of iov^{+}. Again, this result shows the preponderance of a few negative signals versus a large number of positive signals. I consider that the best configuration is *conf7*. Figure **6-2** shows that *conf7* performs among the best configurations in the range from 10 to 100 users, and it is the best for more than 100 users. Although, *conf5* and *conf6* produce ProficiencyRanks for more than 200 users, their correlations are considerably lower than those of *conf7*. In general terms, all optimal values of the parameter α for all considered configurations are larger than 0.5 (see **6-1**), indicating that the information obtained from negative votes (explicit and implicit)

is preponderant in the computation of meaningful ProficiencyRank scores. Similarly, *conf7* serves to determine the ideal balance between explicit and implicit information by observing the optimal values of β and δ parameters (see last row in **6-1**). The value of $\delta = 0.20$ indicates that explicit votes are more indicative of language proficiency than Implicit Opposition Votes. This imbalance is considerable given that the obtained ideal weighted combination was 80 % for explicit votes versus 20 % for explicit. In contrast, the value of $\beta = 0.53$ indicates that explicit positive votes are roughly balanced versus Implicit Agreement Votes, but only iav^{-}, that is the implicit positive votes extracted from agreements between negative votes. Note that *conf4*, *conf5*, and *conf6*, which make use of iav^{+} obtained smaller values of β and comparatively lower performance versus *conf7*. This means that the inclusion of iav^{+} produces an increased importance of explicit votes and a lower performance, indicating that their informativeness is poor.

The top result (ρ=0.86) was obtained using *conf5* and $\theta > 75$. That result corresponds to a set of 12 users having each one more than 75 incoming votes. In spite of being a small subset of users, the observed correlation was highly significant, $_{value} = 0.000323$. This result is somewhat unexpected given the poor performance of *conf5* for larger sets of users. Nevertheless, this result suggests that high correlations can be achieved if there is enough information (votes) associated with each user. To provide conclusive proof of that point it would be necessary to carry out experiments with a considerably larger dataset.

Regarding baseline results (see the last column in **6-1**), it is clear that the effect of the amount of user activity in the collaborative social network is poorly correlated with language proficiency. Only configurations *conf1* and *conf6* obtained significant correlations ($_{value} < 0.05$) but with a considerable margin to the lowest Proficiency Rank results. This result indicates that the measure of proficiency is mostly independent of the amount of activity of the users.

Regarding the third research question, the results of the comparison between ProficiencyRank and a CEFR baseline (see section 5.3) show that there is a possible misalignment

between the targets in each measure aim. Figure **6-1** shows that for all subsets of users (controlled by θ) Proficiency Rank produces significant correlations, while the CEFR baseline does not. The fact that the CEFR baseline is strongly correlated with a large corpus produced by learners in a curriculum aligned with the CEFR (see section 5.3), but poorly correlated with our gold standard, provides empirical evidence of the misalignment of that curriculum with the language proficiency perceived by the YASK's users. It means that there is only a loose relation between the CEFR vocabulary profiles and the self-perceived proficiency in the written modality. Therefore, using the lexical profiles of the CEFR as a point of comparison, these are aligned with the corpus of Cambridge learners ($\rho = 0.65$) on a scale ranging from A1, A2 to C2, but they are not with the texts posted by users in YASK on a scale ranging from "Beginner" to "Native‘' ($r < 0.18$ for more than 20 YASK's users in Figure **6-1**). Although some academics criticize current approaches used in standardized tests (Alderson, 2007; Hulstijn, 2007), our results seem to be the first empirical evidence of the disagreement between the "Wisdom of the crowd" and the written language proficiency tests based on current curricula.

At this point, the question arises of how it is possible that from ProficiencyRank scores emerge proficiency measures that could be more precise than traditional methods? One possible answer is that traditional methods are indirect since they are based on the inference of proficiency from observable traits in learners' written and verbal production. Moreover, the success of said inference depends on the accuracy of the known linguistic, pedagogical and cognitive models used. In contrast, the information used by ProficiencyRank is the collective aggregation of non-expert judgments that directly and naturally judge the communicative capacity of the language used, which helps, along with the voting information, to identify the proficiency of each user. Certainly, an advantage of the proposed approach, is its independence from domain and language. While this facilitates its implementation, the availability of languages and domains depends heavily on the collective will of the users.

I believe that a form of assessment equal or similar to ours should be taken into account as

an alternative way to measure the proficiency level of a person in a language. On one hand, the user who answers questions and earns votes for it demonstrates knowledge in certain topics. This user would show that s/he has high proficiency. On the other hand, the user who asks gains knowledge after answering her/his question, without differentiating whether this answer came as a formative assessment, assessment for learning or peer feedback. Both cases would imply that these users could use the respective language structures learned or taught in real contexts and note that they work. Therefore, regardless of when users ask or answer, these questions or answers would show the specific place of proficiency of a user with respect to other users within the same social network. However, our approach inherits the intrinsic limitation of ranking-based systems versus score-based systems, which makes the assessment relative to the community.

Although ProficiencyRank can only be used in particular technological environments, our results provide an alternative perspective to traditional language assessment and probably reveal the misalignment of these approaches with the fundamental goal of measuring the real proficiency of the individuals. This result also confirms the difficulty of the construction of valid language assessment tests and the potential of the Social Computing technologies in that area. In particular, a Collaborative Social Network to practice/learn a particular skill, knowledge or ability, which makes use of ProficiencyRank could be considered as a Self-Improving Intelligent Educational System (Brusilovsky and Rus, 2019).

6.3.2. Analysis from second Language Acquisition

Theoretical approaches in psycholinguistics and language teaching have drawn a line between L1 and L2 acquisition. Despite age and difference in physiological processes conducted in the brain, there are some common issues to consider, for example, the cues (see 2.3.1) and their frequency in both learning processes [29]. In the Usage-Based Theory (UBT) see 2.3.2 distinguishes stages in language development like pre-linguistic communication, utterances

[29] YASK makes evident that in some cases, people repeat same words, structures, and expressions in different contexts, adapting and adjusting language items

and words, schemas and constructions, and abstract constructions. Keeping the logic from easier to harder. Tomasello (2009) proposes this in children from L1, but through YASK is possible to check these stages in users from beginner to advanced. Probably both acquisition processes run by different systems, but the theoretic proposal could also identify concordance in processes from a pragmatic perspective of learning. YASK users post participation from their constructicon (see 2.4.3) consciously or unconsciously, interaction and validation improve the strength of constructicon. In initial levels of proficiency, users make clear they use the L1 constructicon adjusting into the L2.

6.3.3. Analysis from the educational framework

ProficiencyRank method uses an approach that can be called a social network-based approach, where individuals developed their language abilities and skills by participating actively in a collaborative social network comprised of speech communities and communities of practice holding a common enterprise which is to learn an L2. ProficiencyRank enhances practice through the use of different methods for written competence, which is also applicable to the other competencies, due to collaborative social networks like YASK can be adapted to any type of activity (see Table 2-1).

Analysis from EL2

The use of YASK or any other collaborative social network, displace the roles and context from traditional EL2. In EFL countries this type of resource offers users closer contact with native speakers and natural interaction. Teachers in YASK are not necessary unless the teacher assumes the role of another user

6.3.4. Analysis based on collaborative social networks

Wisdom of the crowd influence

The new method presented in Chapter 5 considers the very first time an alternative approach to written rating, assessment, and automated proficiency prediction. The ProficiencyRank

validity lays on the concept of social interaction and group judgment, which as mentioned previously by (Galton, 1907) and (Surowiecki, 2005), the wisdom of the crowd (WoC) or the collective opinion exceed an expert's opinion. WoC has been widely used in special types of question and answers websites ((Q&A) [30], in the case of Stack Overflow (see 2.10.2), a highly prestige QA focused on computer programming, users participate for reason beyond the shallow quest for information on facts, but also a matter of identity (Gazan, 2011). There is also a sociologic dimension enhancing QA members by taking an active role in the maintenance of worth information system and credibility balance (Shachaf et al., 2009). In this sense, the collaborative online social networks could include strategies to make their user active participants in the content-creation. YASK which is an educative-driven application for mobile gadgets, bases its content on posts made by users (from a list of utilities offered by the APP). The system for the post validation is simple and common through like and dislike voting system.

The criteria for an accurate opinion given by a crowd, consider three items (See 2.10.2):

1. Diversity in the crowd: YASK community crowd is highly diverse coming from many different nationalities in America, Europe, and Asia.[31]

2. The crowd must be decentralized: YASK users do not belong to a well-established hierarchy, there a social distancing among them, and YASK allows a little social interaction, related exclusively to learning processes. Ratings are difficult to access and compile by the rest of the members.

3. Member should be independent: YASK users follow advice from the APP and rate post according to their know competence and knowledge, there is little room to be influenced by other users.

[30](Q&A) are online resources created to serve as users-interactive repositories of specific-related topics. Internet surfers access these sites to find out authorities' source-based information not available on internet queries nor the web. One of the main features of (QA) is high-quality information (Jurczyk and Agichtein, 2007)

[31]YASK's users come from countries as distant as Russia and India.

Management of the wisdom of the crowd within a online social network generates rankings, usually prestige rankings, related to the position of a user in a community. In social groups individuals, follow rules given and enhanced by abilities and skills, the higher abilities and skills, the higher prestige. In Stack Overflow posting indicates knowledge and skills of members of specific programming issues, after validation from other users . In YASK prestige is obtained by revising and judging properly most of the post. Nevertheless, through calculations and analysis in 6.2 prestige based on voting accuracy can be explicit and implicit. This phenomenon is similar to the prestige rankings occurring within linguistic communities, where some users have a higher ranking among commoners due to the age, social status, profession or knowledge.

Linguistic social networks and online social networks

In section 2.1 linguistic social network is defined as a model represented as a huge spider web comprising the entire society. This web is made of ties, some are stronger than others. Direct or closer ties known as first-order network ties and indirect or distant ties known as second-order network ties. Ties highlight relationships among members within networks. Thus, individuals' family and closest friends represent first-order network ties, while neighbors and high school classmates are second and even third-order network ties. Language considers as a social interaction product that evolves within social networks than can be considered as a series of overlapped ties. Despite linguistic social networks still present a vehicle to language development in geographical locations, the internet related technologies have been developing the (online) social network as a virtual environment for interaction among people worldwide.

Garton et al. (1997) considered that online networks connecting people as social networks. Computer-mediated communication extended the panorama to virtual communities, online network collaborative work among others. In a certain way, the transfer of social interaction, including language from face-to-face contexts to online communication, released new paradigms in strategies, methods, and approaches of every aspect including interaction.

Nevertheless, common features go beyond, the community cohesion, confidence, and formation of group identity. When cohesion and identity are strong enough to determine rules and particular linguistic behaviors it can be considered as a speech community. Conversely, community of practice is general understood as the group following a common enterprise, they share conventionalization and meaning construction, due to language features were developed already.

YASK environment follows trends of most of APP in terms of visual interfaces and participation strategies, following gamification tools and resources. Users follow a "personal" route, that exhibits the progress. YASK asks users -each time they enter- to revise and vote for other users' posts. Criteria used for such task assignments depend on the language proficiency level each participant indicates at the registration survey format. Each user is a perfect rater in the own native language, but it is also a suitable rater in those languages with high proficiency as well. In a detailed observation of its features, YASK itself is the network replacing simultaneously geographical location and social factors -neighborhood, tribe, ghetto-. Active participants get involved in a speech community where their posts, votes, corrections, and additional activities chase particular language standards. As YASK is an overseas impact APP -users worldwide- language interactions could include several linguistic registers from different countries, YASK language standard is international English, which is the type of language learned and taught in language courses.

A great extend of users access YASK to practice language and improve their level. On one hand, they rate written samples post on their native language, but on the other hand, users learn sophisticated vocabulary, structures, and expressions from native users. Thus YASK is a community of practice of languages, in our study case, an international English language community of practice.

A requested clarify about previous works developed by sociolinguistic and ethnographers (see sections 2.1 and 2.2) in the applied methodology for this research in data recollection is the ethnographical insertion inside the community, I interacted as part of the speech community (English language and German Language) and in a community of practice (when learning words from Chilean and Mexican Spanish varieties).

The random users and usual users (Dorian, 1982) perceived in description and anlysis of data collide in participation, whether is active (posting) or passive (rating). Participation implies (Hymes et al., 1974) a kind of concurrence of use of grammar and performance rules, as well multiple types of knowledge forms, coherence and constructions. (see 2.2)

6.3.5. Analysis from the CEFR

The CEFR (see 2.6) is a European proposal for curriculum and assessment of languages. Initially, the CEFR was a European Policy document, but it spread worldwide as one of the major referents in education and pedagogic manuals. Despite the CEFR is not the only model (ETS and American Education bureau could differ in few aspects) they all collide in pre-set curriculum, skills and abilities description, educational aims, and assessment criteria. These standards consider language as a railroad, where learners as wagons travel among train stations. Each train station represents a specific skill or ability.

From A1 to C2 there are six scales, each scale comprises many descriptors indicating what language users can do or should do, to be classified at that particular level. The CEFR is a fixed model that creates well-structured curricula and on this, the assessment contents are proposed. Teachers, learners, and users, in general, expecting to work on this model, should follow the development of materials, contents, exercises preparing in advance the future assessment certification. Under this perspective, two problems appear, first, develop through pedagogical resources, and strategies outside the model could have problems in the official assessments. Second, there is no difference between EFL and ESL learning contexts (see 4.1.2). The CEFR appeals to be wide and general, with an advice character, recommending constant updating and revision of communicative categories (see 2.6.2). International institutions adapt their formats to these commands. Designing materials and examinations require high-quality standards based on multiple revisions. International language testing systems, have a lot of advantages in measurement procedures and rating strategies. Notwithstanding, language testing systems are artificial, created by professional staff, under assessment intentions, distant from the user's reality. As a teacher I can confess that my experience test-takers consider the examinations instruments as hard, confusing, and quite different from

their communities of learning (speech communities or communities of practice).
YASK is a non-prescriptive proposal, thus it encourages users to interact by practicing the language, similar to the face-to-face interaction in children. Message can be understood despite errors, and errors are seen as natural according to the linguistic development stage. ProficiencyRank is a coherent method for proficiency assessment prediction under natural-style learning. ProficiencyRank bases on social links among individuals (who are the real raters and assessors) measurement their prestige within the speech community.

Written production and interaction

The overall written production descriptors of CEFR (see table **??**) lay on two main key concepts the i) type of message, and the ii) type of language. As proficiency level highs more complex are types of messages and type of language. In real contexts of application, this can not measure language proficiency nor predict it [32]. Intentions of the speaker could influence the use of a type of vocabulary, intonation, and syntax expressing alternative meanings. YASK does not mind the sophistication as a mandatory marker of high language proficiency but in the consideration made by the user about coherence and accuracy in a post.

Online interaction

The CEFR included in 2018 updating a new descriptor for written performance, online interaction (see 2.6.5). In the CEFR online interaction requires redundancy to ensure understanding. The majority of online social networks are natives or at least proficient in digital communication, knowing codes and written variations. Online interaction offers a wide range of tools for communication, from written instant messaging systems to video chatting. In terms of this method proposal, I consider ProficiencyRank could be considered in future, and after more research as a predictor and assessment proficiency resource aligned to the CEFR.

[32] the main principle in communication is situational and pragmatic, thus in informal communication, words used are not academic nor sophisticated, but slang expressions (usually excluded of language courses). Intentions of the speaker could influence the use of a type of vocabulary, intonation, and syntax expressing alternative meanings.

7 Conclusions

I presented ProficiencyRank, a method for measuring user importance in a collaborative social network. By testing ProficiencyRank in a dataset obtained from a sample of the Yask community, I observed that the rankings obtained by this method are highly correlated with the language proficiency of the users. We carried out experiments that revealed how different amounts of user interactions (postings and votes) produce different intensities of that correlation. In addition, I observed that the most informative signal in a collaborative social network is the one from negative votes. Similarly, we found that between explicit and implicit signals, the former are the most informative. However, the best configurations of ProficiencyRank were obtained by linear combinations of positive, negative, explicit, and implicit votes.

In general, it is possible to say that the results provided by ProficiencyRank are significantly correlated with user self-assessments, becoming a promising tool for developing predictive evaluation tools in collaborative social networks similar to Yask. That is, the users ordered by ProficiencyRank are arranged from "Native" to "Beginner" level meaningfully. It is important to note that the method is independent of the English language and it is not related to any learning curriculum, considered above rubric-based approach, instead, The method for ProficiencyRank is Online-interaction-based approach. The construction of an assessment tool based on this discovery requires more research. For instance, in our experiments, YASK's users did not expect to be evaluated, therefore fraud or artificial preparations are not considered issues. The control of these issues in proficiency evaluation based on social interaction is an interesting research perspective. Similarly, the degree of complexity and difficulty of

the requests in Yask is distributed accordingly with the proficiency of the users, which is roughly uniform. In a potential evaluation scenario based on ProficiencyRank, the requests should be provided by an evaluation authority as a teacher and their difficulty should be controlled. Clearly, extremely difficult or trivial requests can hinder the overall evaluation of the users. In addition, those artificial requests should promote divided voting polls to produce enough negative votes for ProficiencyRank. The determination of an appropriate set of initial requests is also an interesting research topic.

It is also important to note that our method is independent of the domain and modality of the requests. Therefore, the requests could include any type of media opening the perspective of constructing requests based on listening, pronunciation, conversation, translation, etc. Eventually, in other domains, ProficiencyRank could be used to assess other hard-to-evaluate skills, such as pattern recognition, critical thinking, problem solving, etc. We envision MOOCs as an ideal setting to implement an assessment framework based on ProficiencyRank.

7.0.1. Further Research Perspectives

ProficiencyRank is a cutting-edge proposal for assessing collaborative educational processes in any discipline or field. From a pedagogical perspective, it fixes directly to create ties among people sharing common learning goals. Teachers could switch traditional roles towards becoming real addressers and guides who encourage and motivate interactions, opinions, reflection, and knowledge enhancement inside their communities (whether face-to-face or virtual). Also, in my professional environment as an -English language teacher at a public school- ProficiencyRank piloting can be developed by encouraging students to work collaboratively. The assessment becomes an essential part of the interaction, a kind of game in which students have fun while learning.

Related to the core of this paper, a logical extension to this research would be a revision and application of ProficiencyRank to the Spanish language by using accessible data from Yask

or by implementing data recollection from courses in the Instituto Caro y Cuervo. Later, the research could include other languages such as Colombian indigenous ones.

References

Abrahamsson, N. and Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language learning*, 59(2):249–306.

ACTFL, J. and Portal, M. A. (2012). Actfl proficiency guidelines 2012.

Alderson, J. C. (1991). Language testing in the 1990s: How far have we come? how much further have we to go?.

Alderson, J. C. (2007). The cefr and the need for more research. *The Modern Language Journal*, 91(4):659–663.

Alexander, R. (2008). Culture, dialogue and learning: Notes on an emerging pedagogy. *Exploring talk in school*, 2008:91–114.

Anderson, J. R. (1989). A rational analysis of human memory. In Roediger, H. L. and Craik, F., editors, *Varieties of memory and consciousness: Essays in honour of Endel Tulving*, pages 195–210. Lawrence Erlbaum Associates.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied psychological measurement*, 2(4):581–594.

Andringa, S., Dabrowska, E., et al. (2019). Individual differences in first and second language ultimate attainment and their causes. *Language Learning*, 69(S1):5–12.

Angoff, W. H. (1988). Validity: An evolving concept.

Asher, J. J. (1969). The total physical response approach to second language learning. *The modern language journal*, 53(1):3–17.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Bachman, L. F., Palmer, A. S., et al. (1996). *Language testing in practice: Designing and developing useful language tests*, volume 1. Oxford University Press.

Ballier, N., Canu, S., Petitjean, C., Gasso, G., Balhana, C., Alexopoulou, T., and Gaillat, T. (2020). Machine learning for learner english: A plea for creating learner data challenges. *International Journal of Learner Corpus Research*, 6(1):72–103.

Balog, K., Fang, Y., De Rijke, M., Serdyukov, P., Si, L., et al. (2012). Expertise retrieval. *Foundations and Trends® in Information Retrieval*, 6(2–3):127–256.

Barrot, J. S. (2015). Comparing the linguistic complexity in receptive and productive modes. *GEMA Online® Journal of Language Studies*, 15(2).

Basri, H., Hashim, H., and Yunus, M. M. (2019). Using google apps as learning strategy to enhance esl writing. *Creative Education*, 10(12):2649–2657.

Bates, E. and MacWhinney, B. (1982). Functionalist approaches to grammar. In Gleitman, E. W. . L., editor, *Language acquisition: The state of the art*, pages 173–218. Cambridge University Press.

Bennett, S. and Marsh, D. (2002). Are we expecting online tutors to run before they can walk? *Innovations in Education and Teaching International*, 39(1):14–20.

Bialystok, E., Craik, F. I., and Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in cognitive sciences*, 16(4):240–250.

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Biber, D. and Gray, B. (2013). Discourse characteristics of writing and speaking task types on the toefl ibt® test: a lexico-grammatical analysis. *ETS Research Report Series*, 2013(1):i–128.

Biber, D., Gray, B., and Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in l2 writing development? *Tesol Quarterly*, 45(1):5–35.

Biber, D., Gray, B., and Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 37(5):639–668.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., and Quirk, R. (1999). *Longman*

Grammar of Spoken and Written English. Longman.

Birdsong, D. (2005). Interpreting age effects in second language acquisition. *Handbook of bilingualism: Psycholinguistic approaches*, 109:127.

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language learning*, 56:9–49.

Birdsong, D. and Vanhove, J. (2016). Age of second language acquisition: Critical periods and social concerns.

Björnsson, C.-H. (1968). *Lesbarkeit durch Lix*. Pedagogiskt centrum, Stockholms skolförvaltn.

Black, P. and Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, 21(1):5.

Black, T. R. (1999). *Doing quantitative research in the social sciences: An integrated approach to research design, measurement and statistics*. Sage.

Bloomfield, L. (1922). Review of sapir's language. *The Classical Weekly*, 18.

Bloomfield, L. (1933). *Language*.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., and Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18.

Borin, L., Forsberg, M., and Lönngren, L. (2013). Saldo: a touch of yin to wordnet's yang. *Language resources and evaluation*, 47(4):1191–1211.

Bourhis, R. Y. and Marshall, D. F. (1999). The united states and canada. In Fishman, J. A., editor, *Handbook of language and ethnic identity*, pages 244–264. New York: Oxford University Press.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Computer networks*, 33(1-6):309–320.

Brown, D. (2007). Principles of language learning & teaching.(5th eds.).

Brown, J. D. (2014). The future of world englishes in language testing. *Language Assessment*

Quarterly, 11(1):5–26.

Brown, J. D. (2019). World englishes and international standardized english proficiency tests. *The Handbook of World Englishes*, pages 703–724.

Brusilovsky, P. and Rus, V. (2019). Social navigation for self-improving intelligent educational systems. In Sinatra, A. M., Graesser, A. c., Hu, X., Brawner, K., and Rus, V., editors, *Design Recommendations for Intelligent Tutoring Systems*, pages 131–146. US Army Research Laboratory.

Bulté, B. and Housen, A. (2014). Conceptualizing and measuring short-term changes in l2 writing complexity. *Journal of second language writing*, 26:42–65.

Burstein, J. and Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In *The Oxford handbook of applied linguistics*.

Burstein, J., Chodorow, M., and Leacock, C. (2003). Criterionsm online essay evaluation: An application for automated evaluation of student essays. In *IAAI*, pages 3–10.

Canagarajah, A. S. (2006). The place of world englishes in composition: Pluralization continued. *College composition and communication*, pages 586–619.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, 1(1):1–47.

Carroll, J. B. (1961). Fundamental considerations in testing for english language proficiency of foreign students. *Testing the English proficiency of foreign students*, 36.

Carroll, J. B., Davies, P., and Richman, B. (1971). *The American Heritage word frequency book*. Houghton Mifflin.

Castaneda, D. A. and Cho, M.-H. (2016). Use of a game-like application on a mobile device to improve accuracy in conjugating spanish verbs. *Computer Assisted Language Learning*, 29(7):1195–1204.

Chambers, J. K. (1995). *Sociolinguistic theory: Linguistic variation and its social significance*. Blackwell Publishers.

Chapelle, C. and Hegelheimer, V. (2004). The language teacher in the 21st century. *New perspectives on CALL for second language classrooms*, pages 299–316.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *1st Meeting of the North*

American Chapter of the Association for Computational Linguistics.

Chile, C. (2018). Lanzan novedosa app que facilita la comunicación entre chilenos y haitianos. Retrieved: 2019-01-30, https://www.cnnchile.com/tendencias/lanzan-novedosa-app-que-facilita-la-comunicacion-entre-chilenos-y-haitianos_20180329/.

Chomsky, N. (2006). *Language and mind*. Cambridge University Press.

Chomsky, N. (2014). *Aspects of the Theory of Syntax*, volume 11. MIT press.

Clyne, M. and Sharifian, F. (2008). English as an international language: Challenges and possibilities. *Australian Review of Applied Linguistics*, 31(3):28–1.

Cochran, M., Larner, M., Riley, D., and Henderson Jr, C. R. (1993). *Extending families: The social networks of parents and their children*. Cambridge University Press.

Cohen, A. P. (1982). Belonging: the experience of culture. *Belonging: Identity and social organisation in British rural cultures*, pages 1–17.

Collins-Thompson, K. and Callan, J. P. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200.

Compton, L. K. (2009). Preparing language teachers to teach language online: A look at skills, roles, and responsibilities. *Computer Assisted Language Learning*, 22(1):73–99.

COUNCIL, O. E. (2018). Common european framework of reference for languages: learning, teaching, assessment–companion volume with new descriptors. *Strasbourg: Council of Europe*.

Creese, A. and Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching? *The modern language journal*, 94(1):103–115.

Cronbach, L. (1984). How to judge tests. *Essentials of Psychological Testing. 4th ed. New York, Harper & Row*.

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., and McNamara, D. S. (2014). Linguistic microfeatures to predict l2 writing proficiency: A case study in automated writing evaluation. *Grantee Submission*, 7(1).

Crossley, S. A., Salsbury, T., McNamara, D. S., and Jarvis, S. (2011). Predicting lexical

proficiency in language learner texts using computational indices. *Language Testing*, 28(4):561–580.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, (19):121–129.

Dabène, L. and Moore, D. (1995). Bilingual speech of migrant people. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 17–44.

Dale, E. and Chall, J. S. (1949). The concept of readability. *Elementary English*, 26(1):19–26.

Davidson, F. (2006). World englishes and test construction. *The handbook of world Englishes*, pages 709–717.

Davies, A. (1984). Validating three tests of english language proficiency. *Language testing*, 1(1):50–69.

Davies, A. (2003). *The native speaker: Myth and reality*, volume 38. Multilingual Matters.

Dede, C., Richards, J., and Saxberg, B. (2018). *Learning Engineering for Online Education: Theoretical Contexts and Design-based Examples*. Routledge.

Diependaele, K., Lemhöfer, K., and Brysbaert, M. (2013). The word frequency effect in first-and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, 66(5):843–863.

Diessel, H. (2004). *The acquisition of complex sentences*, volume 105. Cambridge University Press.

Diessel, H. and Tomasello, M. (2001). The acquisition of finite complement clauses in english: A corpus-based analysis. *Cognitive linguistics*, 12(2):97–142.

Docherty, G. J., Foulkes, P., Milroy, J., Milroy, L., and Walshaw, D. (1997). Descriptive adequacy in phonology: a variationist perspective. *Journal of Linguistics*, pages 275–310.

Dorian, N. C. (1982). Defining the speech community to include its working margins. *Sociolinguistic variation in speech communities*, pages 25–33.

Douce, C., Livingstone, D., and Orwell, J. (2005). Automatic test-based assessment of programming: A review. *Journal on Educational Resources in Computing (JERIC)*,

5(3):4.

Duranti, A. (1997). Universal and culture-specific properties of greetings. *Journal of linguistic Anthropology*, 7(1):63–97.

Eckert, P. (2000). *Language variation as social practice: The linguistic construction of identity in Belten High*. Wiley-Blackwell.

Ellis, N. C. (1994). *Implicit and explicit learning of languages*. Academic Press Inc.

Ellis, N. C. (2008a). The associative learning of constructions, learned attention, and the limited l2 endstate. In Robinson, P. and Ellis, N. C., editors, *Handbook of cognitive linguistics and second language acquisition*, pages 372–405. Routledge.

Ellis, N. C. (2008b). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The modern language journal*, 92(2):232–249.

Ellis, N. C., Römer, U., and O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Wiley-Blackwell.

Ellis, R. et al. (2003). *Task-based language learning and teaching*. Oxford University Press.

Enright, M. K. and Quinlan, T. (2010). Complementing human judgment of essays written by english language learners with e-rater® scoring. *Language Testing*, 27(3):317–334.

Farag, Y., Yannakoudakis, H., and Briscoe, T. (2018). Neural automated essay scoring and coherence modeling for adversarially crafted input. *arXiv preprint arXiv:1804.06898*.

Fellbaum, C. (1998). Towards a representation of idioms in wordnet. In *Usage of WordNet in Natural Language Processing Systems*.

Feng, L. (2010). Automatic readability assessment.

Ferguson, C. A. (1959). Diglossia. *word*, 15(2):325–340.

Firth, A. and Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in sla research. *The modern language journal*, 81(3):285–300.

Firth, A. and Wagner, J. (1998). Sla territory: No trespassing. *Modern Language Journal*, 72:8–22.

Flores, J. F. F. (2015). Using gamification to enhance second language learning. *Digital*

Education Review, (27):32–54.

Friederici, A. (2009). Brain circuits of syntax: From neurotheoretical considerations to empirical tests. In Bickerton, D. and Szathmáry, E., editors, *Biological foundations and origin of syntax*, pages 239–252. Cambridge, MA: MIT Press.

Friginal, E. and Weigle, S. (2014). Exploring multiple profiles of l2 writing using multidimensional analysis. *Journal of Second Language Writing*, 26:80–95.

Galeotti, A., Goyal, S., Jackson, M. O., Vega-Redondo, F., and Yariv, L. (2010). Network games. *The review of economic studies*, 77(1):218–244.

Galton, F. (1907). The wisdom of crowds. *Nature*, 75(1949):450451.

Gan, H. (1962). *The Urban Villagers: Group and class in the life of Italian-Americans*. New York: Free Press of Glencoe.

Garfield, E. (1996). Fortnightly review: How can impact factors be improved? *Bmj*, 313(7054):411–413.

Garton, L., Haythornthwaite, C., and Wellman, B. (1997). Studying online social networks. *Journal of computer-mediated communication*, 3(1):JCMC313.

Gazan, R. (2011). Social q&a. *Journal of the American Society for Information Science and Technology*, 62(12):2301–2312.

Gentner, D. and Markman, A. B. (1997). Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

Giddens, A. and Sutton, P. W. (1989). *Sociology. Cambridge: Polity*. Cambridge: Polity Press.

Ginther, A. and McIntosh, K. (2018). Language testing and assessment. In *The Palgrave Handbook of Applied Linguistics Research Methodology*, pages 845–867. Springer.

Godwin-Jones, R. (2017). Data-informed language learning. *Language Learning & Technology*, 21(3):9–27.

Gohard-Radenkovic, A., Lussier, D., Penz, H., and Zarate, G. (2004). Reference fields and methodologies. *Cultural mediation and the teaching and learning of languages*, pages 27–58.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*.

Oxford University Press on Demand.

Goldin-Meadow, S. (2009). From gesture to word. In Bavin, E. L., editor, *The Cambridge handbook of child language*, pages 145–160. Cambridge University Press.

Golub, B. and Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49.

González Moncada, A. (2009). On alternative and additional certifications in english language teaching: The case of colombian efl teachers' professional development. *Íkala, revista de lenguaje y cultura*, 14(22):183–209.

Grant, D. M., Malloy, A. D., and Murphy, M. C. (2009). A comparison of student perceptions of their computer skills to their actual abilities. *Journal of Information Technology Education: Research*, 8(1):141–160.

Gu, L. and So, Y. (2015). Voices from stakeholders: What makes an academic english test 'international'? *Journal of English for Academic Purposes*, 18:9–24.

Gumperz, J. J. (1962). Types of linguistic communities. *Anthropological linguistics*, pages 28–40.

Gumperz, J. J. (1996). The linguistic and cultural relativity of conversational inference. *Rethinking linguistic relativity*, pages 374–406.

Gumperz, J. J. and Hernandez-Chavez, E. (1972). Bilingualism, bidialectalism and classroom interaction. *Functions of language in the classroom*, 1.

Hahnel, C., Goldhammer, F., Naumann, J., and Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior*, 55:486–500.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). Hunpos-an open source trigram tagger.

Hall, H. and Graham, D. (2004). Creation and recreation: motivating collaboration to generate knowledge capital in online communities. *International Journal of Information Management*, 24(3):235–246.

Hall, M. A. (1998). Correlation-based feature subset selection for machine learning. *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato.*

Hampel, R. and Stickler, U. (2005). New skills for new classrooms: Training tutors to teach languages online. *Computer assisted language learning*, 18(4):311–326.

Han, Y.-S., Kim, L., and Cha, J.-W. (2009). Evaluation of user reputation on youtube. In Ozok, A. A. and Zaphiris, P., editors, *International Conference on Online Communities and Social Computing*, pages 346–353. Springer.

Han, Y.-S., Kim, L., and Cha, J.-W. (2012). Computing user reputation in a social network of web 2.0. *Computing and Informatics*, 31(2):447–462.

Hancke, J. and Meurers, D. (2013). Exploring cefr classification for german based on rich linguistic modeling. *Learner Corpus Research*, pages 54–56.

Harsch, C. (2016). Proficiency. *ELT Journal*, 71(2):250–253.

Heimann Mühlenbock, K. (2013). I see what you mean.

Herder, J. G. and Scheibe, W. (1949). *Johann Gottfried Herder*. Klett.

Hosio, S., Goncalves, J., Anagnostopoulos, T., and Kostakos, V. (2016). Leveraging wisdom of the crowd for decision support. In *Proceedings of the 30th International BCS Human Computer Interaction Conference 30*, pages 1–12.

Housen, A. and Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics*, 30(4):461–473.

Hudson, R. A. (1996). *Sociolinguistics*. Cambridge university press.

Hudson, T., Detmer, E., and Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics*, volume 2. Natl Foreign Lg Resource Ctr.

Hulstijn, J. H. (2007). The shaky ground beneath the cefr: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, 91(4):663–667.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, 8(3):229–249.

Hultman, T. G. and Westman, M. (1977). *Gymnasistsvenska*. Institutionen för nordiska språk, Univ.

Humboldt, W. (1988). On language: The diversity of human language-structure and its influence on the mental development of mankind.

Humphreys, P., Haugh, M., Fenton-Smith, B., Lobo, A., Michael, R., Walkinshaw, I., et al. (2012). Tracking international students' english proficiency over the first semester of undergraduate study. *IELTS research reports online series*, page 41.

Humphry, S. M. and Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher*, 43(5):253–263.

Husák, M. (2010). Automatic retrieval of good dictionary examples. *Bachelor Thesis, Brno*, 392.

Hymes, D. (1972). On communicative competence. *sociolinguistics*, 269293:269–293.

Hymes, D. et al. (1974). Ways of speaking. *Explorations in the ethnography of speaking*, 1(1974):433–451.

Jacobs, D. T. (2006). *Unlearning the language of conquest: Scholars expose anti-Indianism in America*. University of Texas Press.

Jacobs, H. L. et al. (1981). *Testing ESL Composition: A Practical Approach. English Composition Program.* ERIC.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., and Taylor, C. (2000). Toefl 2000 framework. *Princeton, NJ: Educational Testing Service*.

Jenkins, J. (2006). The spread of eil: A testing time for testers. *ELT journal*, 60(1):42–50.

Johnson, J. C. (1994). Anthropological contributions to the study of social networks: a review. *Advances in social network analysis*, pages 113–151.

Jurczyk, P. and Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 919–922.

Jurkovič, V. (2019). Online informal learning of english through smartphones in slovenia. *System*, 80:27–37.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73.

Karpov, N., Baranova, J., and Vitugin, F. (2014). Single-sentence readability prediction in russian. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 91–100. Springer.

Kerswill, P. et al. (1994). *Dialects converging: Rural speech in urban Norway*. Oxford University Press on Demand.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., and Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychlỳ, P. (2008). Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, pages 425–432. Universitat Pompeu Fabra Barcelona, Spain.

Kim, E., Lin, J.-S., and Sung, Y. (2013). To app or not to app: Engaging consumers via branded mobile apps. *Journal of Interactive Advertising*, 13(1):53–65.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

King, J. R. and Chetty, R. (2014). Codeswitching: Linguistic and literacy understanding of teaching dilemmas in multilingual classrooms. *Linguistics and Education*, 25:40–50.

Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pages 249–256. Elsevier.

Klein, W. (1998). The contribution of second language acquisition research. *Language learning*, 48(4):527–549.

Kleinberg, J. M. (1999). Hubs, authorities, and communities. *ACM computing surveys (CSUR)*, 31(4es):5–es.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, 16(2):81–96.

Knoch, U. and Elder, C. (2010). Validity and fairness implications of varying time conditions on a diagnostic test of academic english writing proficiency. *System*, 38(1):63–74.

Komlodi, A., Soergel, D., and Marchionini, G. (2006). Search histories for user support in user interfaces. *Journal of the American Society for Information Science and Technology*, 57(6):803–807.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief. In *European conference on machine learning*, pages 171–182. Springer.

Kramsch, C. (1993). *Context and culture in language teaching*. Oxford university press.

Kramsch, C. (2011). The symbolic dimensions of the intercultural. *Language teaching*, 44(3):354.

Kukulska-Hulme, A. and Shield, L. (2008). An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL*, 20(3):271–289.

Labov, W. (1966). The social stratification of english in new york city. *Center for Applied Linguistics*.

Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular*. Number 3. University of Pennsylvania Press.

Labov, W. (1989). Exact description of the speech community: Short a in philadelphia. *Language change and variation*, pages 1–57.

Lado, R. (1961). Language testing: The construction and use of foreign language tests. a teacher's book.

Lee, P. and Lin, H. (2019). The effect of the inductive and deductive data-driven learning (ddl) on vocabulary acquisition and retention. *System*, 81:14–25.

Lévy, D. and Zarate, G. (2003). La place de la médiation dans lechamp de la didactique des langues et des cultures. *Français dans le monde. Recherches et applications*, (33):186–189.

Lewis, J. L., Ream, R. K., Bocian, K. M., Cardullo, R. A., Hammond, K. A., and Fast, L. A. (2012). Con cariño: Teacher caring, math self-efficacy, and math achievement among hispanic english learners. *Teachers College Record*, 114(7):1–42.

Li, P. and MacWhinney, B. (2012). Competition model. *The encyclopedia of applied linguistics*.

Lieven, E. and Tomasello, M. (2008). *Children's first language acquistion from a usage-based perspective*. Routledge/Taylor & Francis Group.

Lin, S., Hong, W., Wang, D., and Li, T. (2017). A survey on expert finding techniques.

Journal of Intelligent Information Systems, 49(2):255–279.

Linacre, J. M. (1990). Many-faceted rasch measurement.

Lippi, R., Donati, S., Lippi-Green, R., and Donati, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States*. Psychology Press.

Liu, H. and Brantmeier, C. (2019). "i know english": Self-assessment of foreign language reading and writing abilities among young chinese learners of english. *System*, 80:60–72.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *TESOL quarterly*, 45(1):36–62.

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based l2 writing research and implications for writing assessment. *Language Testing*, 34(4):493–511.

Lyons, J. (1970). *New horizons in linguistics*. Penguin.

Lys, F. (2013). Computer-mediated grammar teaching and its effect on language acquisition over time. *Calico Journal*, 30:166–186.

Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan*, 80(9):673.

MacWhinney, B. and O'Grady, W. (2015). *The handbook of language emergence*. John Wiley & Sons.

Manovich, L. (2009). The practice of everyday (media) life: From mass consumption to mass cultural production? *Critical Inquiry*, 35(2):319–331.

Marx, K. (2015). *Capital: A critique of political economy*, volume 1. Arkose Press.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174.

Matthiesen, S. J. (2017). *Essential Words for the TOEFL*. Simon and Schuster.

McCarthy, P. M., Guess, R. H., and McNamara, D. S. (2009). The components of paraphrase evaluations. *Behavior Research Methods*, 41(3):682–690.

Mccloskey, G., Perkins, L., and Diviner, B. (2008). Assessment and intervention for executive function difficulties. *Assessment and Intervention for Executive Function Difficulties*,

pages 1–362.

McKay, S. L. and Brown, J. D. (2015). *Teaching and assessing EIL in local contexts around the world*. Routledge.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Melchers, G., Shaw, P., and Sundkvist, P. (2019). *World Englishes*. Routledge.

Menken, K. (2006). Teaching to the test: How no child left behind impacts language policy, curriculum, and instruction for english language learners. *Bilingual Research Journal*, 30(2):521–546.

Menn, L., Duffield, C. J., Newmeyer, F., and Preston, L. (2014). Looking for a 'gold standard'to measure language complexity: what psycholinguistics and neurolinguistics can (and cannot) offer to formal linguistics. *Measuring grammatical complexity*, pages 281–302.

Mercer, N. and Hodgkinson, S. (2008). *Exploring talk in school: Inspired by the work of Douglas Barnes*. Sage.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, 45(1-3):35–44.

Milardo, R. M. (1988). *Families and social networks*. Sage Publications, Inc.

Milburn, T. (2015). Speech community. *The International Encyclopedia of Language and Social Interaction*, pages 1–5.

Milroy, J. and Milroy, L. (1993). Mechanisms of change in urban dialects: the role of class, social network and gender. *International Journal of Applied Linguistics*, 3(1):57–77.

Milroy, L. (1987). *Language and social networks*. Blackwell.

Milroy, L. (1999). Women as innovators and norm-creators: The sociolinguistics of dialect leveling in a northern english city. In *Engendering Communication: Proceedings of the 5th Berkeley Women and Language Conference. Berkeley, BWLG Publications*, pages 361–376.

Milroy, L. and Llamas, C. (2013). Social networks. *The handbook of language variation and change*, pages 407–427.

Miltsakaki, E. and Troutt, A. (2008). Real time web text classification and analysis of reading difficulty. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 89–97.

Mitchell, J. C. (1986). Network procedures. *The quality of urban life*, 73:92.

Monner, D., Vatz, K., Morini, G., Hwang, S.-O., and DeKeyser, R. (2013). A neural network model of the effects of entrenchment and memory development on grammatical gender learning. *Bilingualism: Language and Cognition*, 16(2):246–265.

Moore, M. G. and Kearsley, G. (1996). *Distance education: A system view*. Wadsworth.

Morgan, M. H. (2014). *Speech communities*. Cambridge University Press.

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., and Faloutsos, C. (2013). Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 886–893. ACM.

Munro, M. J. and Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, 45(1):73–97.

Nakamura, T. (2019). Understanding motivation for learning languages other than english: Life domains of l2 self. *System*, 82:111–121.

Nassiri, I., Masoudi-Nejad, A., Jalili, M., and Moeini, A. (2013). Normalized similarity index: An adjusted index to prioritize article citations. *Journal of Informetrics*, 7(1):91–98.

Neuner, G. and Byram, M. (2003). Intercultural competence. strasbourg: Council of europe. *Language Policy Division*.

Nicholls, D. (2003). The cambridge learner corpus: Error coding and analysis for lexicography and elt. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581.

Nielsen, J. (2006). Participation inequality: Encouraging more users to contribute. *http://www. useit. com/alertbox/participation_inequality. html*.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *LREC*, volume 6, pages 2216–2219.

Noack, R. and Gamio, L. (2015). The world's languages, in 7 maps and charts. *The Wa-*

shington Post, 4(23):65–70.

Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating caf in instructed sla: The case of complexity. *Applied linguistics*, 30(4):555–578.

North, B. and Piccardo, E. (2016). Developing illustrative descriptors of aspects of mediation for the cefr. *Strasbourg, France: Council of Europe. rm. coe. int/common-european-framework-of-reference-for-languages-learning-teaching/168073ff31*.

of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, C. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

O'keeffe, A., McCarthy, M., and Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Oller, J. W. (1979). Language tests at school.

Oller Jr, J. W. (1983). *Issues in language testing research*. ERIC.

Ortega, L. (2003). Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied linguistics*, 24(4):492–518.

Östling, R., Smolentzov, A., Hinnerich, B. T., and Höglin, E. (2013). Automated essay scoring for swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pajak, B. (2016). Which countries study which languages, and what can we learn from it? duolingo.

Pal, A., Farzan, R., Konstan, J. A., and Kraut, R. E. (2011). Early detection of potential experts in question answering communities. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 231–242. Springer.

Palloff, R. M. and Pratt, K. (1999). *Building learning communities in cyberspace*, volume 12. San Francisco: Jossey-Bass.

Palmer, A. S. and Bachman, L. F. (1981). Basic concerns in test validation. In Alderson, J. C. and Hughes, A., editors, *Issues in Language Testing. ELT Documents 111*, pages

135–151. British Council.

Pan, W., Liu, N. N., Xiang, E. W., and Yang, Q. (2011). Transfer learning to predict missing ratings via heterogeneous user feedbacks. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Panadero, E. and Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational research review*, 9:129–144.

Papoutsoglou, M., Mittas, N., and Angelis, L. (2017). Mining people analytics from stackoverflow job advertisements. In *2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 108–115. IEEE.

Pardoel, B. and Athanasiou, A. (2019). Moodle app gamification features and their potential for foreign language learning.

Patrick, P. L. (2001). The speech community. Technical Report 35, Department of Language and Linguistics, University of Essex.

Peacock, C. (2017). *Classroom skills in English teaching: A self-appraisal framework*. Routledge.

Pearson, A. W. and Lumpkin, G. (2011). Measurement in family business research: How do we measure up?

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Piaget, J. (1977). The role of action in the development of thinking. pages 17–42.

Piccardo, E. (2012). Médiation et apprentissage des langues: pourquoi est-il temps de réfléchir à cette notion? *Ela. Études de linguistique appliquée*, (3):285–297.

Pilán, I. (2018). *Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning*.

Pilán, I., Volodina, E., and Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, pages 174–184.

Rachels, J. R. and Rockinson-Szapkiw, A. J. (2018). The effects of a mobile gamification

app on elementary students' spanish achievement and self-efficacy. *Computer Assisted Language Learning*, 31(1-2):72–89.

Ramsey, P. H. (1989). Critical values for spearman's rank order correlation. *Journal of educational statistics*, 14(3):245–253.

Raymond, M. (2015). English grammar in use.

Rebuschat, P. (2015). *Implicit and explicit learning of languages*, volume 48. John Benjamins Publishing Company.

Rodrigues, L. F., Oliveira, A., and Rodrigues, H. (2019). Main gamification concepts: a systematic mapping study. *Heliyon*, 5(7):e01993.

Romaine, S. (1982). *Sociolinguistic variation in speech communities*. Arnold.

Ros Martínez de Lahidalga, I. (2008). Moodle, la plataforma para la enseñanza y organización escolar.

Rosell-Aguilar, F. (2017). State of the app: A taxonomy and framework for evaluating language learning mobile applications. *CALICO journal*, 34(2):243–258.

Sadler, D. R. (2014). The futility of attempting to codify academic achievement standards. *Higher Education*, 67(3):273–288.

Sailer, M., Hense, J. U., Mayr, S. K., and Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, 69:371–380.

Sandberg, K. L. and Reschly, A. L. (2011). English learners: Challenges in assessment and the promise of curriculum-based measurement. *Remedial and Special Education*, 32(2):144–154.

Sapir, E. (1921). *An introduction to the study of speech*. Harcourt, Brace.

Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language learning*, 43(3):313–344.

Saville-Troike, M. (1982). The etnography of communication basil blackwell.

Sawaki, Y., Stricker, L. J., and Oranje, A. H. (2009). Factor structure of the toefl internet-based test. *Language Testing*, 26(1):005–30.

Schauer, G. A. (2006). Pragmatic awareness in esl and efl contexts: Contrast and development. *Language learning*, 56(2):269–318.

Schmid, H. (1994). Treetagger-a language independent part-of-speech tagger. *http://www. ims. uni-stuttgart. de/projekte/corplex/TreeTagger/*.

Schmidt, R. W. (2001). Attention. cognition and second language instruction. In Robinson, P., editor, *Cognition and second language instruction*, pages 3–32. Cambridge University Press.

Schoonen, R., Gelderen, A. v., Glopper, K. d., Hulstijn, J., Simis, A., Snellings, P., and Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language learning*, 53(1):165–202.

Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

Secretariat, C. (1998). University of cambridge local examinations syndicate.

Shachaf, P., Rosenbaum, H., Abels, E., Radford, M., Silipigni Connaway, L., Gazan, R., and Shah, C. (2009). Social reference and digital reference: Online question answering practices in two diverse communities. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–5.

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

Soars, J. and Soars, L. (2001). *American Headway 1-Student book*. Oxford University Press.

Spafford, E. (1990). The usenet. In *The User's Directory of Computer Networks*, pages 386–390. Elsevier.

Spolsky, B. (1995). *Measured words: The development of objetive language testing*. Oxford University Press.

Stewart, W. (1962). An outline of linguistic typology for describing multilingualism. *Study of the role of second languages in Asia, Africa, and Latin America*, pages 15–25.

Ströbel, M., Kerz, E., Wiechmann, D., Neumann, S., et al. (2016). Cocogen-complexity contour generator: Automatic assessment of linguistic complexity using a sliding-window

technique. In *CL4LC@ COLING 2016*, pages 23–31.

Surowiecki, J. (2005). The wisdom of crowds ancor books.

Tack, A., François, T., Roekhaut, S., and Fairon, C. (2017). Human and automated cefr-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179.

Taguchi, N., Crawford, W., and Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? a case of argumentative essays in a college composition program. *Tesol Quarterly*, 47(2):420–430.

Taylor, L. (2006). The changing landscape of english: Implications for language assessment. *ELT journal*, 60(1):51–60.

Tenfjord, K., Meurer, P., and Hofland, K. (2006). The ask corpus-a language learner corpus of norwegian as a second language. In *LREC*, volume 6, pages 1821–1824.

Tesnière, L. (1959). Éléments de syntaxe structurale.

Tomanek, K., Hahn, U., Lohmann, S., and Ziegler, J. (2010). A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1158–1167.

Tomasello, M. (2009). The usage-based theory of language acquisition. In Bavin, E. L., editor, *The Cambridge handbook of child language*, pages 69–87. Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press.

Toulmin, S. E. (2009). *Return to reason*. Harvard University Press.

Tsui, A. B. (2001). Classroom interaction. *The Cambridge guide to teaching English to speakers of other languages*, pages 120–125.

Vajjala, S. and Loo, K. (2013). Role of morpho-syntactic features in estonian proficiency classification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72.

Vajjala, S. and Loo, K. (2014). Automatic cefr level prediction for estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Van Gelderen, A., Schoonen, R., Stoel, R. D., De Glopper, K., and Hulstijn, J. (2007). Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology*, 99(3):477.

Volodina, E., Pilán, I., Eide, S. R., and Heidarsson, H. (2014). You get what you annotate: a pedagogically annotated corpus of coursebooks for swedish as a second language. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 128–144.

Von Ahn, L. (2013). Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 1–2.

Wardhaugh, R. (1998). *Sociolinguistics*.

Weber, M. (2002). *The Protestant ethic and the "spirit" of capitalism and other writings*. Penguin.

Wei, L. (1994). *Three generations, two languages, one family: Language choice and language shift in a Chinese community in Britain*, volume 104. Multilingual Matters.

Weinreich, U. (2010 [1953]). *Languages in contact: Findings and problems*. Number 1. Walter de Gruyter.

Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13.

Wind, S. A. (2020). Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? *Assessing Writing*, 43:100416.

Wisniewski, K., Schöne, K., Nicolas, L., Vettori, C., Boyd, A., Meurers, D., Abel, A., and Hana, J. (2013). Merlin: An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In *ICT for Language Learning 2013, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni*.

Witten, I. H. and Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77.

Wolfe-Quintero, K., Inagaki, S., and Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Number 17. University of Hawaii Press.

Wulff, S. and Ellis, N. C. (2018). Usage-based approaches to second language acquisition. In Miller, D., Bayram, F., Rothman, J., and Serratrice, L., editors, *Bilingual cognition and language: The state of the science across its subfields*, volume 54, pages 37–56. John Benjamins Publishing Company.

Xue, J. and Zuo, W. (2013). English dominance and its influence on international communication. *Theory & Practice in Language Studies*, 3(12).

Yaniv, I. and Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, 103(1):104–120.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Yıldız, B. and Ozger, Z. O. (2012). Generalization of the lee weight to zpk, twms j. *App. &Eng. Math*, 2(2):145–153.

Yoon, H.-J. and Polio, C. (2017). The linguistic development of students of english as a second language in two written genres. *Tesol Quarterly*, 51(2):275–301.

Young, M. and Wilmott, P. (2013). *Family and kinship in East London*. Routledge.

Zarate, G. (2003). *La médiation et la didactique des langues et des cultures*. Number 14. Clé International.

Zentella, A. C. (1997). Latino youth at home, in their communities, and in school: The language link. *Education and Urban Society*, 30(1):122–130.

Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230.

Zhang, W. and Liu, M. (2008). Investigating cognitive and metacognitive strategy use during

an english proficiency test. *Indonesian JELT*, 4(2):32–49.

Zhao, J., Shu, H., Zhang, L., Wang, X., Gong, Q., and Li, P. (2008). Cortical competition during language discrimination. *NeuroImage*, 43(3):624–633.