

What 1000 Hours of L2 Media Exposure Did to a Complete Beginner

Zsombor Kiss

June 11, 2022

Abstract

This is a case study inspired by Foley's (2018) autoethnography. Like him, I exposed myself to around 1000 hours of media in the TL, which in my case is Spanish. But, in contrast to him, I did not avoid reading, I chose my content more carefully and I had no previous exposure to my TL. Throughout the study I mainly focused on the development of my perceived comprehension, and the effect that the above changes would have on it, providing descriptions of how well I perceived to understand select pieces of content every week. After stating my results, I move on to compare them with Foley's and conclude that my changes had a net positive effect. A heuristic argument is presented to try to gauge the extent of the ambiguity that could arise in such comparisons. The implications of this ambiguity for a theory capable of explaining the discussed differences are explored.

1 Introduction

One of the most important questions in SLA is whether input is sufficient for L2 acquisition, and I believe that studying the effects of L2 media on learners could lead to major advancements because of how easy it makes quantifying exposure: hours of content consumed gives a much better idea of the amount of input received than the length of residence abroad. Despite this, relatively little research has been done on the area, especially on audiovisual media, (see for example Vanderplank, 2010).

The only longitudinal study that focuses on L2 media (that I am aware of) is an autoethnography by Foley (2018), which served as an inspiration for this case study. In the first phase of his experiment, he exposed himself to approximately 1300 hours of French audio intended for native speakers, without using a dictionary or receiving any form of instruction. His results indicated that he had clearly acquired a considerable amount of French, although extremely slowly. Some of the details of his design raise very deep questions:

1. He completely avoided reading. Would the quicker learning of vocabulary through reading over short periods of time (Brown et al, 2008; Vidal, 2011) translate to faster acquisition of competence in the long run?
2. He chose his content with little regard for comprehensibility, for example he started by going through a list of French television programs from Wikipedia and listening to the radio. Would more comprehensible content, as Krashen's theory would predict, accelerate acquisition?
3. For the first 83 days, he noted down every 'consciously observed comprehension of French,' what he called 'mental events.' Could it be that this made him focus more on form than content, which according to Krashen's theory could reduce his rate of acquisition?
4. He had already received some form of French instruction in school, although he claims to have forgotten it. This casts some doubt on whether someone with no prior instruction could reproduce his results. That is: is input sufficient?

Conclusively answering and providing explanations for these questions would more or less require a theory of SLA and is of course beyond the scope of any single study. Nevertheless, I believed that a series of case studies of acquisition under extreme conditions could prove useful for the development of such a theory. This study aimed to be second in line after Foley, recreating his conditions with changes motivated by the above questions: not avoiding reading, choosing content based on comprehensibility, not taking notes immediately and not having previous exposure to the TL.

With initial plans of continuing said series of case studies, I decided to make all these changes at once, and not one by one, for if they did not turn out to have a net positive effect, then it would have been very likely that none of them had an effect, eliminating the need for several other studies. Therefore the immediate goal of this one was to see if there would indeed be such change.

2 Research Design

2.1 The Subject

Due to the ethical concerns of forcing someone to learn using a method that was probable to be very slow, based on Foley's (2018) results, I decided to serve as the subject of this study. I am a 21-year-old native speaker of Hungarian. I am also fluent in English, have a level of Japanese similar to what was achieved in this study, and I have taken 4 years of German classes, but I have gained practically no competence from them.

I chose Spanish as the TL, based on personal preference, it was the language I was most motivated to learn. In fact, I had already attempted to start learning this language in this way: this first attempt lasted 4 days, during which I consumed 30.5 hours of Spanish content, most of which was audiovisual. Roughly

5 months passed between this and my second attempt, with no exposure to the TL.

Besides this, I had no background in Spanish, other than the bits I heard in English TV shows, and some things that are common knowledge for English speakers like ‘bueno.’ I have never been to a Spanish-speaking country before or during the study, and never received any form of instruction.

2.2 Study Parameters

According to the Foreign Service Institute’s (hereafter FSI) data, it takes around 23 to 24 weeks for native speakers of English to reach Speaking/Reading 3 on the IRL scale, with 25 hours of classes a week (Jackson & Kaplan, 2001). Unfortunately, on its own this is not very useful, since I planned to spend 0 hours in the classroom, but if we take into account that they are expected to spend a minimum of three hours a day outside of class to succeed (Berbeco, 2016 p. 185-186), then we get an estimate of 920 hours, which is comparable to Foley’s 1300 hours, considering that French can be learned at approximately the same rate as Spanish (Jackson & Kaplan 2001).

However, since English is not my first language, and Hungarian is much more distant from Spanish, it is not clear how much this would impede my learning. If I were to attain results worse than or similar to that of Foley’s, then it could not be deduced if the positive effects of my changes were canceled out by the negative effects of my native language, or if those effects were simply non-existent to begin with. I decided to take this risk and set to aim for this 900–1300-hour range.

Guided by question 3 of the Introduction, I would not document mental events. Instead, I would focus on the only thing that I could without outputting: comprehension. More precisely perceived comprehension. To do this, I planned to give a qualitative description of my perceived comprehension of at least one piece of content every week, and a more comprehensive one at the end of the study.

In order to compactify these descriptions, I would assign the given piece of media a number between 0 and 50, based on the following scale that I developed for myself based on what I remembered of my own acquisition of Japanese:

- 0 - I don’t understand anything.
- 5 - I recognize a word here and there, but I can’t understand any sentence other than maybe the rare 2-3 words long one every-now and then.
- 10 - I regularly recognize words and, rarely, I can even understand simple sentences that are a few words long.
- 15 - I am familiar with most words that I encounter, but I can only understand short and simple sentences regularly, and I only occasionally understand more complex sentences.

- 20 - I understand a good chunk of the sentences I encounter, but I also don't understand a good chunk of them. Understanding a sentence is 'not the norm' Most of my ability to follow the piece of content still comes from other means (e.g., guess from the visuals.)
- 25 - I understand the majority of sentences I see, but it feels like I still don't understand the important ones, I can follow the piece of content to some extent, but I lose track very often and regularly.
- 30 - I understand most sentences, understanding them is 'the norm', and I can follow a good chunk of what is being talked about, but I still lose track often, and sentences I don't understand still show up often.
- 35 - I can follow most of what is being talked about, but I lose track occasionally, and sentences I don't understand show up regularly, but most of the time aren't too much of a problem because of context. A lot of the nuance is lost on me (jokes, cultural and implicit things.)
- 40 - I can follow pretty much all of what is talked about and understand almost all sentences with occasional exceptions, but some of the nuance is still lost on me. I still encounter a lot of words I don't know.
- 45 - I understand virtually all sentences and never lose track, but occasional unknown words still show up, and a bit of the nuance is still lost.
- 50 - I believe I understand every single word and all the nuance.

To rate a piece of content, I would choose two adjacent statements from above, say A and B, that I would find most fitting and then sum a number between 0 and 5 to the number corresponding to A based on the following:

- 0 - I strongly agree with A, but not with B.
- 1 - I agree with A and to a much lesser extent with B.
- 2 - I agree with both A and B but slightly more with A.
- 3 - I agree with both A and B but slightly more with B.
- 4 - I agree with B and to a much lesser extent with A.
- 5 - I strongly agree with B, but not with A.

The resulting number, hereafter 'comprehension score,' would be given based on a 30-minute segment of the content being rated, being questioned and adjusted in real time during those 30 minutes, and not just given retrospectively. It must be noted that these numbers are only a shorthand for a description and are not meant to hide their qualitative nature.

Furthermore, I planned to count the hours of exposure organized into the following categories:

- Audio-only
- Audiovisual
- Subtitled (Audiovisual with Spanish subtitles)
- Text-only
- Text with visuals (For example comic books and textbooks that have figures or equations)
- Speaking (I planned to do a minimal amount of speaking towards the end to see if I could at all)

Occasional journal entries were also made, but always near the end of the day, to avoid the possible problem articulated in question 3 of the Introduction.¹

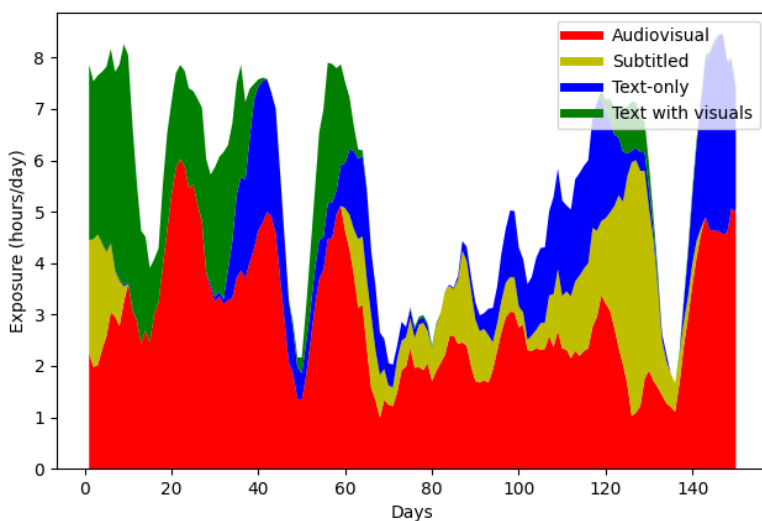


Figure 1: 7-day moving average of exposure

3 Execution

3.1 Time Spent

Before I began testing myself for my final description, and not counting my first attempt, I spent a total of 871 hours consuming media in Spanish, over a period of 156 days.¹ Divided into the predetermined categories we have:

Audio-only	2 hours
Audiovisual	463 hours
Subtitled	125 hours
Text-only	144 hours
Text with visuals	137 hours
Speaking	10 minutes

I only spoke once, after the 800th hour. I retold a fairy tale that I had read in a novel a few days before and confirmed that I could indeed speak to some extent. To get an idea of the proportions of the other categories as the study went on, see Figure 1.

3.2 Conscious Efforts

There is ambiguity in saying I watched/read content. What is it that I was actually doing in my mind? Concisely, I tried my best to pay attention and understand, and although I regularly had trouble with my mind drifting and I did eat occasionally while watching videos, for the most part the content I was consuming had my complete attention.

With the intent of trying to understand, however, I did make some conscious effort to understand the language, which might not be negligible compared to someone who is not so familiar with how languages work. This effort came in the form of guessing what words could mean, how they are read and in keeping an eye out for evidence for a few conjectures that arose during this process. I figured out a lot more than I expected, as noted in the following journal entries (the hours include the first attempt):

Day 13 134.62 hours Week 2 Saturday ...Here is what I figured out so far: The writing is pretty much completely phonetic, the few sounds that I have problems with are ll and y which may or may not be the same sound, and I think I tend to hear two sounds for y one voiced and the other not, this might just be allophonic. One of the few things that I knew about Spanish before this experiment is that it distinguishes between rolled and unrolled r-s which would explain the difference between r and rr, but I speak languages with rolled and unrolled r-s so I think I would be able to hear the difference between the two but I can't which leads me to doubt my memory. I'm also not sure if the h is silent or not most of the time or if it's just a very soft version of j. [...]

About the grammar, I know that nouns have at least two classes el and la, if there are others they must be very rare.[...] The plurals end in an s. Adjectives can come either before or after the noun they modify and conjugate to match its gender and plurality. The verbs work very similar to English with the added conjugations for the subject and the object. The latter can also be expressed by putting the me/te... before the verb. There are continuous and perfect tenses and passive voice...

Day 40 307.19 hours Week 6 Friday ...I noticed that there are two distinct past tenses as in 'estaba' vs 'estuvo' [...] I'm gonna call it [the latter] the recent past [...] And with that I think I got all of the inflected forms of verbs: present, past, recent past, future (e.g. iré), imperative (e.g. escuchen), hypothetical (e.g. podría), infinitive (e.g. caer), perfective (e.g. hecho) and continuous (e.g. haciendo), plus different forms of these conjugated for the subject where applicable. I used to think that there were also different inflections for different objects, but it's not so much of an inflection as it is just appending me/te/lo/la/le/se/nos/os and their plurals to the end or beginning of the verb...

I discovered these in the process of trying to guess what inflected forms of verbs meant, in the beginning they were a 'fuzzy mess' of similar looking words, in the sense described in this entry:

Day 13 134.62 hours Week 2 Saturday ...I don't remember the words exactly, only as jumble of letters or sounds, and I usually don't know what sets apart words that sound similar or refer to similar things for example [...] [a] set that has been bothering me for a while are the c-nouns for head related things. [...]: cara is face, cabeza is head, cabello means hair. I'm not sure about the last one because there is a very similar c??b??ll?? word that probably means horse...

But as I started getting a feel for what extra meaning different inflections carried, the words became clear 'meaning exactly what they say,' and although I believed myself able to recognize different inflections based on this feeling, I never figured out the rules for their conjugations, nor the situations they are used in.

I never thought about grammar for its own sake, only in real time during the above process. I never found out about the word order for example, though had I decided to look at a few example sentences, I could have probably figured it out, but since it was not necessary for guessing an unknown word before me, I never did it.

3.3 Content Selection

For the first 300 hours I watched children's shows that I had seen on the Disney Channel and Nickelodeon when I was around 7 to 13, as well as their continuations and spin-offs that I had not seen before, and I read translations of simple Japanese comic books. I also experimented with shows aimed at a younger audience, like Ben and Holly, which was far easier, but I decided against these because they were less enjoyable.

Around 320 hours in, I started reading the Harry Potter series, which took me the rest of my journey to finish, and I started watching Dragon Ball, an

animated series, which I continued to watch until around 480 hours in. Next to reading Harry Potter, I also kept reading comic books up to the 440th hour.

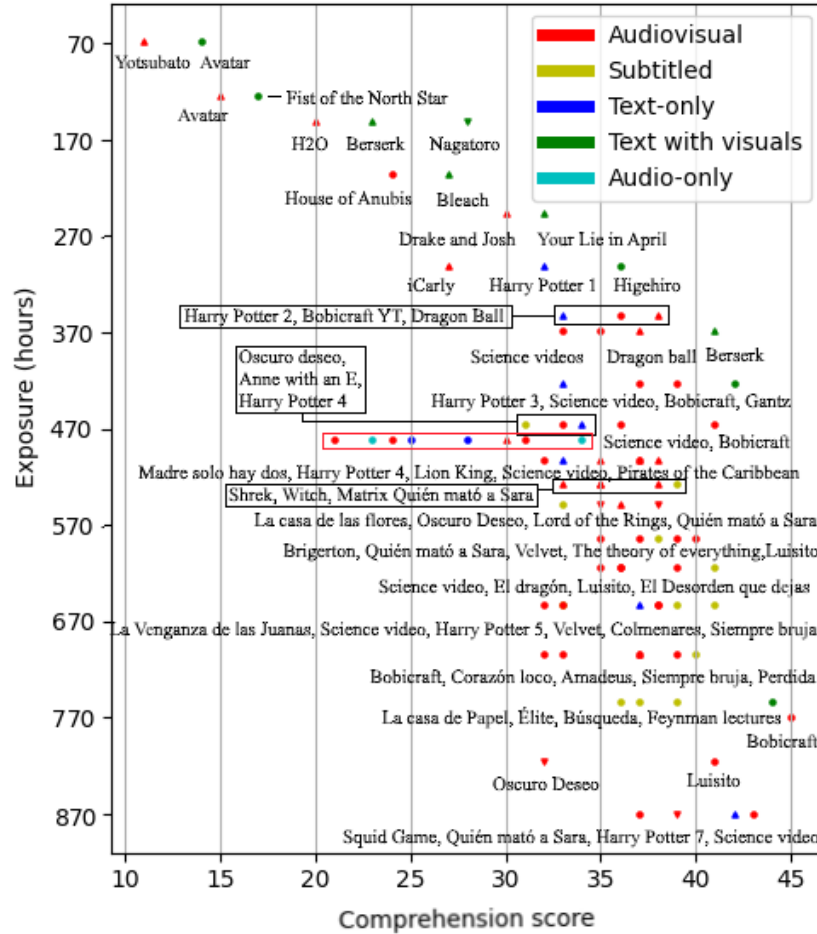


Figure 2: Comprehension scores

Another thing I started at this point, was watching YouTube videos, mostly about the video game Minecraft, and what I call 'science videos:' these were short 10-minute documentary style videos mostly about the natural sciences, but also about math and language learning. These YouTube videos went on to

make up most of my audiovisual exposure for the rest of the study, though I later branched out to travel vlogs and videos about movies.

In addition, I exposed myself to dubbed movies and TV shows between the 480th and 600th hours, and in the final 50 hours; and to original Spanish shows starting from 480 hours in until the end.

The only things that I watched/read several times were H2O, the first season of Avatar, Nagatoro, Oscuro Deseo and Quién mató a Sara.

See Figure 2 for the comprehension scores given throughout the study. They are grouped into lines based on which week they were given, so the times on the left are not exact. They are also concrete examples that are representative of the content consumed that week.

Their shapes represent whether I have seen them before in Spanish, (upside down triangle) another language, (triangle) or neither (circle). Shorthands for their titles can be seen below the datapoints, matching them left to right, except where they are encased in rectangles. The ones encased in the red rectangle are outliers: they are not representative of the content I consumed that week; I was only testing myself. These pieces of content are, from left to right: Somos, YT Podcast, Club de Cuervos, 100 años de soledad, Count of Monte Cristo, House of Cards, House M.D, Multanime podcast.¹

3.4 Complications and Limitations

Not all went as planned: one time, I had forgotten to turn off the timer and another, I did not switch categories when I should have, so I had to estimate with these two cases, which should be negligible, considering that I had a total of 2190 sessions.

In order to search for content, I also had to write some Spanish into search bars, and I also translated some titles, by looking them up on Wikipedia and then switching the language to Spanish, though I only had to do this on a few occasions, most of the time I found the pieces of content searching for the English title plus 'español.'

Once, I came across a picture that intended to make fun of some differences between English and Spanish by pairing the different inflections of 'do' with a cute cartoon figure and the inflections of 'hacer' with a menacing one. This was the closest the consumed content came to explicitly teaching Spanish, so I recorded it in my journal:

Day 62 438.4 hours Week 9 Saturday ...I was mostly aware of all of these [inflections] and have a rough idea of what they mean except for the ones [...] that I recognise by the ie part in them. I thought there was only one with that ie part but it turns out there are actually 3...

Beside these minor complications, there is also the elephant in the room: the limitations of self-assessment. On top of its undisputed inaccuracy, there is the question: do I actually understand what I think I understand? Surely not

in all cases, but considering that by the end I believed to have understood most utterances, consistently for hundreds of hours, is it really plausible that what I believed to understand was widely different from what was actually being said, even though it was consistent discourse in line with the visuals? Here we run into the problem of quantifying this disparity, which is beyond the scope of this study.

It is also worth mentioning that contemplating each score for 30 minutes while consuming the content, instead of just assigning them retrospectively might have improved the assessment's accuracy. For example, looking back at it now, I do not remember the 28 of Nagatoro being different from the 42 of Harry Potter 7.

There is also the question of how well the descriptors of my scale would resemble the actual development of my comprehension. I noted this down in the following entry:

Day 55 383.09 hours Week 8 Saturday ...I like how I made no mention of how well one should be able to follow the piece of content on a whole at the lower levels, because this varied to great extent, depending on how much the visuals helped. But I'm not happy about a few lines I included like: 'understanding them [sentences] is 'the norm.'" Although I do understand and expect to understand most of the sentences if I rate something above this score, it all still feels like a miracle that I can understand them. [...]

Also 'I understand the majority of sentences I see, but it feels like I still don't understand the important ones.' I didn't get this feeling at all, when learning Japanese this might have been the case, because I feel like cognates in Spanish seem to be more frequent in exposition dumps, the absence of which in Japanese lead me to not understand 'the important bits'. That, or I might just remember wrong and this didn't happen with Japanese either.

And finally, I don't like my mentions of nuance, it doesn't seem to correlate in the way I thought with the other parts of my descriptions.

4 Results

After my 901.5 hours of exposure, I proceeded to give the planned description of my level. As opposed to my other comprehension scores, these are given based on the first 15 minutes of the pieces of content, instead of the 30 minutes that could start from anywhere. To write this, I had to expose myself to 7 and a half hours of Spanish over the course of 4 days, bringing my total time spent up to 909 hours.

4.1 Reading

When reading, I can follow most everything that I come across, though the range of my comprehension can vary widely.

The easiest is when the story is accompanied by visuals. The following pieces of content are the first ones that appeared in the 'populares' tab of a site dedicated to Asian comic books translated to Spanish:

- The Death Mage that doesn't want a fourth time: 47
- Shinjiteita Nakama Tachi Ni Dungeon Okuchi De Korosare Kaketa ga Gift...(The full title is too long): 47
- Me Convertí En La Esposa De MI Protagonista Masculino: 46

Technical texts are also surprisingly easy. I chose 3 branches that are pertinent to my interests and within each of these I selected a topic that I had not studied before, but plan to: differential geometry from mathematics, the Python programming language from computer programming and thermodynamics from physics, so I Googled for 'geometria diferencial pdf,' 'python pdf español' and 'termodinamica pdf español' respectively and clicked the first pdf that came up. I skipped forewords and such.

- Introduccion a La Geometría Diferencial - Luis Javier Hernandez Paricio: 48
- Aprenda a Pensar Como un Programador con Python: 47
- Termodinámica - Y. A Çengel, M. A. Boles: 44

Fiction is a lot harder. It is often the case that my comprehension is fairly low, but even then, I usually understand enough to be able to follow the story. The following three books are the 3 bestsellers of 2021 according to the first list that came up on Google after I searched 'libros más vendidos 2021'

- La Bestia - Carmen Mola: 33
- Últimos días en Berlín - Paloma Sánchez-Garnica: 37
- La cuenta atras para el verano - La Vecina Rubia: 38

The next book was chosen because it has a reputation of being a hard one, though I have already read the first half of the first chapter, so I continued from there.

- Cien años de soledad - Gabriel García Márquez: 33

The four books above were originally written in Spanish but translated works are usually easier, even the next one which I picked because I thought it would be difficult, resulted relatively easy. (I had already read its first 2 chapters, so I started rating from chapter 3)

- The Count of Monte Cristo: 39

4.2 Listening

My listening is undoubtedly worse, I can follow most things but the things that I cannot are definitely not rare. I have problems with understanding different accents, with slang and with low clarity audio. Also, I find that my comprehension drops significantly when using speakers instead of headphones, so I must note that the following scores were given using speakers.

The easiest type of audiovisual content is perhaps YouTube videos. To select 'random' videos I used an account with empty history, I chose 3 topics: science, travel and videogames and searched for 'datos científicos curiosos,' 'vlog de viaje' and 'datos curiosos de videojuegos' to get results about these topics. The following videos were the first results from channels I was not familiar with:

- 10 Datos Científicos que no podrás creer: 39
- VLOG un día de viaje *en la Ciudad de México* Comida y turismo: 40
- 50 CURIOSIDADES de los VIDEOJUEGOS: 36

Though in some cases when people's speech has a relatively low volume and clarity compared to that of others, as in some interviews for example, I may not be able to follow what they are saying, as in the following video:

- Este Jugador Tiene Un Survival Que No Podrás Creer!: 29

When it comes to movies and TV series there is something about the audio that makes it a lot less clear, though I do not know what it is. This problem is a lot less significant with dubbed content and I can follow almost all of them for this reason, even if it is at a low level of comprehension. The following shows were the first ones to come up on a newly made Netflix Mexico profile that were both dubbed and new to me.

- The Crown: 35
- You: 41
- Suits: 34

Original Spanish TV shows are even harder. I googled 'best telenovelas on netflix' and selected the first three that I had not seen yet.

- Hache: 35
- Ingobernable: 38
- La Reina del Flow: 24

I am most familiar with European and Mexican Spanish and other accents can completely throw me off, as it is shown by the above series. They were made in Spain, Mexico and Colombia respectively.

Audio only content can be very hard, and I often cannot follow them. I googled 'mejores podcasts' and tried the top 3 from the first list that came up:

- Deforme Semanal Ideal Total - Furia: 28
- Guerra 3 - E01 La noche siria: 27
- Solaris - E18 Algoritmos creativos: 42

4.3 Speaking

My speaking is even worse. I think I can communicate to some extent, but it takes great mental effort on my part, my delivery is not very fluid, and I make tons of mistakes both in terms of pronunciation and grammar. Mistakes that even I notice immediately after I make them.¹

5 Discussion

5.1 Comparison with Foley (2018)

Foley did not describe his comprehension as verbosely as I did, making comparison difficult. From the later journal entries documenting the first phase of his study we get:

JEN593 D264 APR21 It's very disheartening. I don't understand most of what they're saying [in Shrek 3]. In Shrek 1, I had the same thing and Shrek 2 was better. The whole time I'm thinking, maybe everyone is right in that I am wrong. It's a ridiculous way to learn and perhaps impossible. I've been through this many times before. When I understand, I feel encouraged and when not, very disheartened. There are just so many words. At least I know I was wrong about how quickly I would learn. I would have thought that by now I'd be killing it [be very good at it] but shows like this make me realise I'm not. Is it just that this is an adult program language wise? How do I explain the learning I have already done and the programs I can understand... (Foley, 2018 p. 189)

This was written after approximately 850 hours of exposure. Given that I scored Shrek a 33 on my scale after 550 hours, I think it is safe to say that the changes made based on the questions of the introduction had a net positive effect on my comprehension.

5.2 An Attempt at a Comparison with the FSI

As mentioned already, according to Jackson & Kaplan (2001), native speakers of English reach Speaking/Reading 3 on the IRL scale in a similar time frame. Since with such scales, like with that of the CEFR, it is often the case that tests claiming to place the test subject on said scale have little regard in their design for its descriptors; (see Harsch, 2018) therefore the degree to which the results

of the FSI Language Proficiency Test line up with the descriptors of the IRL scale is unclear. Meaning that we cannot use them for a conclusive comparison.

Nevertheless, they are enough to indicate that the disparity between my comprehension and fluency in speaking is not normal:

S/R-3 General professional proficiency: Able to speak accurately and with enough vocabulary to handle social representation and professional discussions within special fields of knowledge; able to read most materials found in daily newspapers. (Jackson & Kaplan, 2001)

I interpret this disparity as evidence that input is not the only factor contributing to the development of speaking skills.

However, considering the ease with which I can ‘read’ technical texts, a case could very well be made that my description comes close to ‘able to read most materials found in daily newspapers,’ so that in terms of reading at least I came close to the level usually achieved at the FSI, but of course, as discussed above, this is nowhere near conclusive.

Although this does make one wonder: supposing that we did have accurate descriptions, just how reliable are such comparisons? In the case of Foley (2018), we had the same concrete piece of content (Shrek) translated into our respective languages, but what about scales with context free qualitative descriptions like those of the IRL and the CEFR? Just how ambiguous is using such qualitative descriptions of competence as a means of comparing the results of different studies?

This potential inaccuracy would pose a major problem for the series of case studies mentioned in the Introduction, so I present a heuristic argument to try to gauge the extent of this ambiguity.

5.3 A Heuristic Argument

Let us look at the example of the CEFR scale with its 6 levels, or 7 if we count not reaching the first level as a level 0. As Harsch (2018) puts it: ‘Perhaps it is not surprising that the three contributions [...] all point toward the non-comparability of different tests, exams, or teacher judgements when reporting student proficiency in terms of CEFR levels.’ Let us be generous and assume that ways of measuring one’s CEFR level (hereafter CL) are never more than 1 level off from their ‘true’ CL and are ‘often,’ say no more than 25% of the time off by 1.

Now let us assume that we have that the measured CL of A is one below that of B, that they are independent, that their tests are equally likely to over- or underestimate their CLs and that their ‘true’ CL-s are equally distributed among the 7 possible values. Then, using Bayes’ theorem, we obtain that the probability that the ‘true’ CL of B is greater than that of A is about 79% which is extremely low compared to the standards of confidence intervals used in science. Thus, it is safe to say, that the measured CL-s of A and B need to be at least two CL-s apart to conclude that one is better than the other.

This might not seem too bad on a scale of 7, but when combined with the logarithmic nature of such scales, we get disastrous results. According to Heathcote et al. (2000), the power law of skill acquisition is, in reality, an exponential law at the individual level. If we assume that this applies to the CEFR scale, say, that the time required to reach the next CL is about half of the time already spent, then we get that it would take $1.5 \times 1.5 = 2.25$ times as long to reach CL $x+2$ than it does to reach CL x . Put another way, if we had two different individuals spend the same amount of time on a language, and then compared the results through qualitative descriptions, in some cases, we could not even tell if one of them performed twice as well as the other!

Of course, this argument rests on a lot of speculation, but if we accept to view qualitative descriptions as ambiguous to this extent, then some strange phenomena in SLA suddenly make sense. For example:

In our half century of language education at FSI, we have moved from “teaching the textbook” to “helping the learner to learn,” [...] Based on reports from overseas, we believe that we are doing a better job of preparing our students now than we ever did before. Yet, the interesting fact remains that Foreign Service officers used to learn their languages to high levels in the 1960s, just as they do today. [...] Why did learners learn almost as successfully in the early days of the long histories of FSI, [...] and other comparable institutions, as they do today, despite the clear increases in the field’s understanding of teaching and learning? Do the curriculum and teaching techniques, in fact, not really matter? (Jackson & Kaplan, 2001).

To which the answer, based on this view, would be that they are likely doing a much better job than they used to, it is only that they cannot tell.

5.4 Some Implications of this Ambiguity

If the definition of competence in an operationalizable way is this difficult, then does the same problem apply to other constructs? I suspect that yes. Take for example comprehensible input. Depending on its interpretation, one could easily twist this study to be a case both for and against comprehensible input: the input that I received as a whole was clearly not comprehensible, so this could be taken to indicate that incomprehensible input is also useful for learning; on the other hand, if we interpret every single utterance comprehended as a chunk of comprehensible input, within the input as a whole, then it could be deduced that my selection of easier content with a higher density of comprehensible utterances, is what lead me to obtain better results than Foley (2018).

Things get even worse when we consider that a theory of SLA, whose constructs are defined by such qualitative descriptions, (i.e., any description ranging from sentence long dictionary-style definitions to long series of descriptors like those of the CEFR) is limited by the ambiguity of said definitions, since they are the only links between different instruments meant to operationalize these constructs. And because, in my opinion, researchers in SLA envision a theory based

on such constructs, I feel the need to point out that there is a very real possibility that such a theory is not possible; and that one of the reasons for which Krashen's Monitor theory was rejected by many: the non-operationalizability of his constructs, is not a problem with his specific definitions, but an inherent limitation of all definitions given by such qualitative descriptions.

What a theory of SLA should actually address has already been a matter of ample discussion, (see Jordan, 2004) but for the purposes of this paper it shall mean any theory capable of predicting the development of a learner's competence depending on factors like the ones described in the Introduction. If the discussed ambiguity is indeed real, it still might not make such a theory impossible, provided that changes in the learner's competence caused by small changes in the learners' method are major in comparison.

However, given the level of competence that I have achieved in this study, (at least in reading) how close it comes to results that can be expected from methods that are at other extremes of the implicit vs. explicit continuum and the massive uncertainty of comparison through qualitative descriptions, I am led to believe that this is not the case and that such a theory is indeed impossible.

5.5 An Alternative

To help clarify the above discussion, (Constructs defined by qualitative descriptions, as opposed to what?) I shall give an example of another way in which a theory of SLA could be searched for: if the problem is the definition of a construct linking two tests attempting to operationalize it, then I propose that we simply omit them from our theories. To replace them with something more rigorous I turned to two principles that I had drawn from physics:

1. The process of measurement of a construct IS the construct. Take for example the construct of length, its process of measurement, grossly speaking, is the stacking of meter sticks, with one meter being defined by a prototype stick. Or at least that is how it used to be. And even though this definition comes very close to our intuitive idea of length, it has some extremely unintuitive properties like length contraction in special relativity. In a similar way, we could define our constructs by a 'prototype test' that resembles our intuitive ideas to some extent, instead of trying to put these ideas into words using qualitative descriptions, and then trying to operationalize them.
2. If we cannot say anything about the individual, resort to statistics. In quantum mechanics, for example, our measurements are not precise enough to be able to say much about a single electron, however by studying the distribution of the result of an experiment performed on many electrons we can still come to understand a lot about them. Similarly in SLA it is unlikely that we could make exact predictions for an individual's performance on our prototype tests, but we could perhaps make predictions about the likelihood of them performing in a certain way.

This raises two big problems: first, it would require experiments that are both longitudinal and have many subjects, but this can be made more feasible by not requiring the experiment to be performed on the subjects simultaneously, since if they are taking the same tests, their results would be comparable. Secondly, one subject could not take the same test twice. This could be remedied by making tests similar to the prototype test, and then establishing with a separate experiment a distribution for how likely it is for someone to obtain result A on the prototype test given that they obtained result B on the second test, figuratively speaking, making a second meter stick by comparing it to the first. As an example, suppose that we have a prototype test for reading comprehension with two outcomes: pass and fail, then a predictive statement might look something like this: If X conditions hold, then the probability of the subject passing the prototype test is a function f of the hours of exposure, for example

$$p = f(t) = \frac{1}{1 + e^{-0.03(t-300)}}$$

An optimal place to start developing such a theory would be performing numerous cross-sectional studies trying to design constructs, that is prototype tests, between which a rich set of correlations can be established, before embarking on longitudinal studies to see how these constructs evolve with time.

For example, suppose that we have a construct for listening comprehension, one for speaking competence and one that tries to predict what stage of acquisition of a set of fixed morphemes the subject is at, in an attempt to gauge their place along the natural order; then could we predict precisely, with distributions, how far along the natural order the subject is, based on the other two tests? Would it depend more on comprehension or speaking competence?

6 Conclusion

Before starting this study, I had naively hoped that it would help work towards a theory that would be capable of answering the questions of the introduction by being one of many case studies with extreme conditions that were yet to be done. I obtained results, that to me at least, were surprisingly good, much better than those of Foley (2018), indicating that my changes had a positive effect. But after trying to compare said results with those of more traditional methods, my views turned more skeptical: I ended up doubting not only my envisioned series of case studies, but also my original idea of how a theory of SLA should be formulated.

Of course, one cannot prove that a method of searching for a theory will never bear fruit, but perhaps the sheer amount of theories (according to Long (1993) there were around 40-60 of them) should serve as an indication that there is some fundamental problem at play. I do not mean to imply that the rough sketch of the previous section is the way it needs to be done, but I am led to believe that without applying a similar level of rigor, not 100 more theories and 1000 more case studies will get us any further.

Notes

¹ The complete journal, starting and ending time for every single session, longer titles for the content rated, should the shorthand be ambiguous, and recordings of my speech can all be found on the project's GitHub: https://github.com/kissZs/spanish_experiment

References

Berbeco S. (2016) Foreign Language Education in America. In: Berbeco S. (eds) Foreign Language Education in America. Palgrave Macmillan, London. https://doi.org/10.1057/9781137528506_1

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, 20, pp.136-163. Hawaii: University of Hawaii.

Foley, Peter. (2018). Picking up a second language from television: an autoethnographic L2 simulation of L1 French learning. <https://doi.org/10.14264/uql.2018.854>

Harsch, Claudia (2018): How Suitable Is the CEFR for Setting University Entrance Standards?, *Language Assessment Quarterly*, <https://doi.org/10.1080/15434303.2017.1420793>

Heathcote, Andrew & Brown, Scott & Mewhort, Douglas. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*. 7. 185-207. <https://doi.org/10.3758/BF03212979>

Jackson, F. H., & Kaplan, M. A. (2001). Lessons learned from fifty years of theory and practice in government language teaching. In J. E. Alatis & A.-H. Tan (Eds.), *Georgetown University Round Table on Languages and Linguistics 1999: Language in Our Time* (pp. 71–87). Washington, DC: Georgetown University Press

Jordan, Geoff (2004). Theory construction in second language acquisition. John Benjamins.

Long, Michael H. (1993). Assessment Strategies for Second Language Acquisition Theories. *Applied Linguistics*, 14(3), 225–249. <https://doi.org/10.1093/applin/14.3.225>

Vanderplank, Robert (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching*, 43, pp 1-37 <https://doi.org/10.1017/S0261444809990267>

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, 61, pp.219-258. Michigan: Wiley-Blackwell.