LA CIENCIA DE LA GRAMÁTICA: MÉTODO Y PRÁCTICA
GUÍA DE CLASE (05-03-2014)

Protocolos de recolección de juicios y sintaxis experimental

Carlos Muñoz Pérez
cmunozperez@filo.uba.ar
Universidad de Buenos Aires

1. Introducción

Cuando discutimos los juicios de aceptabilidad llegamos a dos conclusiones provisorias:

(1) *Conclusiones provisorias (no necesarias) de la clase anterior*
- a. Los juicios de aceptabilidad son graduales.
- b. Queremos mantener la distinción fuerte entre gramaticalidad y agramaticalidad, básicamente para poder elegir entre una G_1 y una G_2.

Tomando estos dos criterios, vamos a ver una serie de protocolos estandarizados para obtener datos a partir de juicios de aceptabilidad. También vamos a discutir un poquito acerca de diseño experimental. Todo esto lo tomamos de Schütze y Sprouse (2012).

Luego vamos a discutir el carácter "informal" de los juicios de aceptabilidad frente a los experimentos basados en estos protocolos. Experimentos POSTA con estadística, control de variables varias, etc.

Y vamos a terminar hablando de la "enfermedad" de los gramáticos (no, no es gripe).

2. Tareas para recolectar datos

Se puede hacer una mínima distinción de las tareas de recolección de datos. Básicamente hay (i) *tareas numéricas* y (ii) *tareas no numéricas*. Las tareas no numéricas están pensadas para detectar *diferencias cualitativas* (e.g, esta oración es buena, esta oración es mala), pero pierden de vista la magnitud de la diferencia. Las tareas numéricas, por el contrario, miden la *magnitud de la diferencia*, pero pierden de vista la diferencia cualitativa.

> It is important to note that at a fundamental level, all of the acceptability judgment tasks are the same: the participants are asked to perform the same cognitive task, that is to report their perceptions of acceptability. Because the cognitive task is the same, the data yielded by each task is likely to be very similar (modulo small differences in the response scale discussed above), especially when the criterion for comparison is the detection of differences between conditions. […] Though there are likely to be differences between tasks with respect to statistical power (e.g., Sprouse and Almeida submitted), when it comes to simply detecting a difference between conditions at relatively large sample sizes (e.g., 25 participants), the fact that the cognitive task is identical across these measures strongly suggests that choice of task is relatively inconsequential (Schütze y Sprouse 2012: 11).

2.1 TAREA DE ELECCIÓN FORZOSA (*FORCED CHOICE TASK*)

La EF es una tarea no numérica que está explícitamente diseñada para comparar cualitativamente dos formas gramaticales.

Se les presentan a los participantes dos (o más) oraciones y se les pide que elijan cual es la más aceptable.

(2)
- a. Todos los ladrones fueron atrapados por la policía. ⊟
- b. Fueron todos los ladrones atrapados por la policía. □

(3) *Ventajas*
- a. Es una tarea muy simple y rápida de presentar. No requiere entrenamiento previo.
- b. Nos informa sobre las diferencias cualitativas entre formas gramaticales.

(4) *Desventajas*
- a. La información cualitativa es "indirecta" (e.g. ¿cómo interpretar que el 65% de los participantes escogió la forma X?).
- b. No sabemos en qué punto de la escala de aceptabilidad se encuentran las formas estudiadas.

2.2 TAREA DE SI/NO (*YES/NO TASK*)

La prueba YN es una tarea no numérica que está diseñada para probar la relación entre una oración y dos categorías: *aceptable* e *inaceptable*. Noten el contraste con la EF, en donde la relación se daba entre dos (o más) oraciones.

Básicamente, se le pide a los participantes que evalúen si una determinada forma gramatical es aceptable o no.

(5) ¿Qué dijo Juan que quién compró? ☐ Si ⛀ No

(6) *Ventaja*
Es una tarea muy simple y rápida de presentar. No requiere entrenamiento previo.

(7) *Desventajas*
a. Es menos sensible que la EF para detectar diferencias/similitudes entre formas gramaticales (es necesario comparar la cantidad de respuestas positivas por cada oración presentada).
b. No sabemos en qué punto de la escala de aceptabilidad se encuentran las formas estudiadas.

2.3. Tarea de escala de Likert (*Likert scale task*)

En la tarea de EL a los participantes se les presenta una escala numérica cuyos extremos se definen como aceptable e inaceptable y se les pide que ubiquen cada oración en un punto de la escala.

Las escalas pueden ser diseñadas con extremos impares (e.g., 1-5 o 1-7) para que haya un punto intermedio (3 o 4), o bien con un extremo impar y el otro par (e.g., 1-4 o 1-6) para forzar, al menos, una elección de polo.

(8) Juan le parece a Luís ser un buen candidato.
☐ 1 ☐ 2 ⛀ 3 ☐ 4 ☐ 5

(9) *Ventajas*
a. Es una tarea muy simple y rápida de presentar.
b. Permite ver la gradación en la aceptabilidad de diversas formas.

(10) *Desventaja*
La escala sugiere que los intervalos son uniformes, pero no es posible saber si todos los hablantes consideran del mismo modo la diferencia entre 4 y 5 que la diferencia entre 2 y 3 (piénsenlo en términos de notas del 1 al 10). Esto puede reducirse si al principio del experimento se presentan ejemplos de oraciones con su valoración en la escala.

2.4. Tarea de estimación de magnitudes (*magnitude estimation task*)

En la tarea de EM a los participantes se les brinda una oración de referencia y se les dice que la aceptabilidad de ducha oración tiene un valor numérico particular (e.g., 100). La oración de referencia recibe el nombre de *Estándar* y su valor numérico se denomina *Módulo*. Los participantes deben asignar valores numéricos a nuevas oraciones en proporción al Módulo (e.g., una oración que es "el doble de aceptable" que la Estándar recibe una puntuación de 200). (Sirve pensarlo como una regla de 3 simple).

(11) a. Algunos niños juegan en el parque. **100**
b. Algunas trampas, los jugadores las conocen. ___

La EM fue especialmente diseñada para subsanar el problema de los intervalos no uniformes que plantean las EL: dado que la Estándar es la única unidad de medida (en tanto proporción) se supone que los intervalos resultantes son mucho más detallados que los de las EL.

(12) *Desventajas*
a. A pesar de la ilimitada cantidad de posibles valores, los participantes se manejan con pocas variables: tienden a elegir múltiplos de 5 y de 10, y eso no tiene nada que ver con el estímulo.
b. Puede que toda la idea de las magnitudes variables esté, lisa y llanamente, mal…

La EM fue originalmente utilizada en *psicofísica*[1], y parece que ahí funciona bastante bien. En psicofísica, también existe la *producción de magnitudes*, que sería la contrapartida de la EM. Por ejemplo, yo emito un grito de "volumen 100" y le pido a alguien que grite lo mismo con "volumen 50". ¿Cómo funcionaría esto para la aceptabilidad"

(13) *Producción de Magnitudes*
Algunos niños juegan en el parque. **100**
*Instrucción: ahora produzca una oración con aceptabilidad de **50*** (¿?)

La idea de "proporcionar" los estímulos no parece funcionar del todo bien con el lenguaje.

2.5. Algunas precisiones sobre el diseño experimental

Empecemos hablando de las **instrucciones** para un experimento/encuesta gramatical. Si bien no hay una manera estandarizada de presentar las "consignas", hay varias cosas

[1] Estudio de la relación entre la magnitud de un estímulo físico y la intensidad con que es percibido

que es necesario aclarar que quedan por fuera de lo que se está investigando. Algunos de estos aspectos son:

- Prescripción gramatical
- Si la oración sería o no utilizada en el mundo real
- Valores de verdad de la oración
- Si es posible "entender" la oración por más que suene mal

Una manera de controlar este último factor es estableciendo como consigna que lo que se entiende por *aceptable* es una oración que podría ser emitida por un hablante nativo.

Algo llamativo es que, a diferencias de otros tipos de experimentos, existe cierto consenso de que las instrucciones para este tipo de experimento no alteran sustancialmente los resultados (básicamente, porque los participantes no les prestan demasiada atención). El cuidado en el diseño experimental debe caer en otros aspectos (materiales, estímulos, ítems de relleno, etc.).

A veces es necesario ofrecerles a los participantes del experimento la posibilidad de que **practiquen** el protocolo. Esto es especialmente cierto para las tareas numéricas. Por ejemplo, En las tareas de *Escala de Likert* es común encontrar anclajes de ciertas formas gramaticales a ciertos puntos de la escala: un ejemplo del tipo de patrón sintáctico con menos o más valor en la escala (o ambos). También suele darse una pequeña práctica sin dar aviso al participante en la que aparezcan construcciones distintas a las estudiadas y que lleven a ubicar valores a lo largo de toda la escala.

En tareas de *Estimación de Magnitudes* es normal incluir una fase de práctica en la cual los participantes entiendan la lógica del experimento. La idea de que se trabaja con proporciones se introduce a través de ayudas gráficas (e.g., una barra larga para 100, la mitad de la barra para 50, etc.).

Otro aspecto importante: si estamos llevando a cabo un experimento (y no simplemente pidiendo un juicio), lo más seguro es que nos estemos preguntando algo más complejo que simplemente evaluar la aceptabilidad o no aceptabilidad de un par mínimo. En esos casos, es recomendable utilizar un ***diseño experimental factorial***. Un experimento factorial mide dos (o más) factores de comportamiento de las unidades experimentales a partir de un conjunto de estímulos que cubre todas las posibles combinaciones de dichos factores.

Imaginemos que queremos estudiar la interacción del llamado *D-Linking* y las restricciones sobre la extracción de nominales complejos.

(14) a. ¿Qué hiciste el comentario de que compró Juan?
b. ¿Qué libro hiciste el comentario de que compró Juan?

¿Podemos a partir de estos datos decir que el *D-Linking* mejora la aceptabilidad de las extracciones desde nominales complejos?

No.

Podría ser que el *D-Linking* mejore la aceptabilidad de ***todas*** las oraciones (y no sólo de las que tienen extracción desde nominales). Para probar esto, necesitamos comparar oraciones que no contengan violaciones de nominales complejos.

(15) a. ¿Qué comentaste que compró Juan?
b. ¿Qué libro comentaste que compró Juan?

La pregunta ahora es si la diferencia entre (14a) y (14b) es igual a la diferencia entre (15a) y (15b). Esto nos dirá si el *D-Linking* mejora (i) sólo extracciones desde un nominal complejo o (ii) construcciones con extracción a nivel general.

Las oraciones de (14) y (15) combinadas forman un experimento con diseño factorial porque hay dos factores en juego (tipo de extracción y tipo de elemento-wh) que tienen cada uno dos posibles valores (±isla y ±D-Linking).

Algo muy importante de incorporar en cualquier encuesta/experimento son los **ítems de relleno** (i.e., oraciones que no se relacionan con la investigación). Estos cumplen tres objetivos:

- Sirven para evitar los efectos de *priming*, y para evitar que el participante note qué se está investigando.
- Sirven para garantizar una mejor distribución de las respuestas a lo largo de la escala.
- Sirven para meter estímulos con otros objetivos investigativos (experimento 2x1).

3. Experimentos informales vs experimentos formales

Como hemos explicado hasta aquí, los juicios de aceptabilidad son "pequeños experimentos". Sin embargo, rara vez se utilizan los métodos de recolección de datos propios de la psicología experimental para reportarlos (protocolos experimentales estructurados, análisis estadístico, etc.). Básicamente, se han estado realizando experimentos psicológicos informalmente.

El problema es cómo sabemos si esta metodología es fiable o no. Pensémoslo así: si los datos son poco fiables (o están mal), la teoría también es poco fiable (o está mal).

Acá es donde entra el paper de Sprouse y Almeida (2012). Ellos revisan los datos de un manual introductorio de gramática generativa (Adger 2003), exactamente 469 oraciones, de forma experimental y controlada para ver si los datos a partir de los que se sostiene la teoría son o no confiables.

¿Qué se quiere decir cuando un dato es "poco fiable"? Hay dos maneras en que un dato puede estar mal: puede ser un *falso positivo* (un experimento que dice que A es distinto de B, lo cual no es cierto) o un *falso negativo* (un experimento dice que A es igual a B, pero en realidad son distintos). Dado que los falsos positivos son más "importantes", Sprouse y Almeida se abocan a ellos.

Los autores clasifican los datos del manual de acuerdo a las siguientes categorías.

Pattern: These are sentences that are reported as part of a group of two or more sentence types that form a pattern of accept-ability as is standard in generative syntax. A pattern always included at least one starred example and one un-starred example.

Existence: These are sentences that were used to demonstrate the existence or inexistence of a given construction in English.

Repeats: As a textbook, some sentences are repeated for expository or pedagogical reasons.

Not English: These are examples that are non-English. We included non-US and non-standard dialects of English (as defined by Adger) in this category because the participant population in our experiments was native US-English speakers.

Untestable: These are sentences that required a task different from acceptability judgment. For example, some data in syntax is based on the availability or unavailability of specific inter-pretations or readings of a potentially ambiguous string of words. These cannot be tested using a standard judgment survey.

	Tokens	Percentage
Pattern	261	29.9%
Existence	250	28.6%
Repeats	124	14.2%
Not English	144	16.5%
Untestable	94	10.8%
Total	873	

Table 1
The distribution of data-like examples in Adger (2003).

De los datos de *pattern,* habían 63 oraciones repetidas. Y 21 de esas oraciones estaban sin la condición de control (el par con el *). Eso suma 219. Las oraciones estudiadas son esas más las de *existence,* lo que suma 469.

Las 219 oraciones de *pattern* se evaluaron con estimación de magnitudes.

Las 250 oraciones de *existence* se evaluaron con tarea SI/NO.

> For the magnitude estimation experiments, we will define ***replication*** as the simple detection of a significant difference in the correct direction be-tween the conditions in a phenomenon. For example, if a phenomenon consisted of two conditions, a replication obtains if the sentence reported as more acceptable by Adger (usually through the lack of a diacritic) is observed to be significantly more acceptable than the sentence reported as less acceptable by Adger (through the presence of a diacritic). Similarly, for the yes–no experiments, we will define replication as the observation of significantly more yes-responses than no-responses for the sentences that were reported as grammatical by Adger, and as the reverse (more no-responses than yes-responses) for the sentences that were reported as ungrammatical by Adger. In other words, for grammatical sentences the proportion of yes-responses must be significantly greater than .5, and for ungrammatical sentences the proportion of no-responses must be significantly greater than .5.

Vamos a saltear algunos aspectos formales del experimento. Las fallas en la replicación del experimento con estimación de magnitudes rondó el 2,5% (dependiendo del método de cálculo), y la tarea SI/NO rondó los mismos números. Incluso asumiendo que ese margen es erróneo (i.e. falsos negativos. También podría ser el caso de que sean datos "neutros" que no digan nada), la replicación general de los datos de Adger es de al menos el 98%.

> It seems to us that the weakest possible conclusion is that there is a coherent set of 469 sentence types (forming 365 phenomena) that are at least 98% replicable. This means that any critic who wishes to claim that syntactic data is unreliable must simultaneously provide an account that explains the re-liability of this data set.

¿A qué se debe que los datos que llegaron al manual (y en los que se basa lo que podríamos denominar una "versión canónica de la teoría") sean tan buenos? Phillips ofrece una suerte de explicación:

> It is not difficult to dig through a few linguistics papers to find a list of questionable acceptability judgments. However, it is less easy to find cases of widely accepted generalizations that are based upon suspect data. Al-though the typical 'armchair linguist' does not systematically test his generalizations using large sets of example sentences and many naïve informants, empirical claims nevertheless undergo extensive vetting before they attain the status of 'widely accepted generalization'. If a key judgment is questionable, this is likely to be pointed out by a colleague, or by audience members in a talk, or reviewers of an abstract or journal article. (Phillips 2009: 3).

¿Podemos concluir algo a partir de todo esto? Se puede, de momento, adoptar una postura salomónica: considerar que los métodos experimentales tienen un lugar complementario al de los juicios informales. Los experimentos tienen algunas ventajas indiscutibles:

- Son mejores para datos en los que es inherente cierta gradiencia, y en donde importa la magnitud de las diferencias.
- Son analizados con herramientas estadísticas, así que resultan datos «más confiables» (i.e., menos debatibles, más defendibles).

Por supuesto, también tienen desventajas:

- Toman tiempo.
- Requieren una inversión monetaria (a veces mínima, pero recordemos que los juicios informales son totalmente gratuitos).
- Requieren participantes (lo que a veces se complica: lingüista de campo)

4. La enfermedad del lingüista

Considere por 15 segundos cada una de estas oraciones:

(16) a. A ninguno de ellos los exámenes le angustian.
b. A nadie sus defectos le avergüenzan.
c. A ninguno de ellos el dinero le obsesiona.
d. A nadie la música le gustó.

Todo gramático ha notado que luego de trabajar mucho con algún tipo de construcción se vuelve insensible a los juicios sobre ella. El efecto se manifiesta generalmente como la progresiva aceptación de formas que suelen ser consideradas agramaticales. Este efecto ha recibido varios nombres informales (e.g., enfermedad del lingüista), pero sólo recientemente (Snyder 2000) ha sido acuñado el término actualmente más difundido: *saciedad sintáctica* (syntactic satiation).

Mientras los lingüistas suelen considerar este efecto como un *gaje del oficio*, su existencia ha sido utilizada por los críticos como una prueba de que los juicios de gramaticalidad no son una fuente fiable de datos.

Snyder (2000) responde: si la inestabilidad de los juicios *es una propiedad de los juicios gramaticales*, y no sólo un síntoma de la labor del lingüista, entonces los no lingüistas deberían manifestarla también. Es más. Si mostramos que sólo una parte de las formas inaceptables produce saciedad, entonces la inestabilidad de los juicios *podría ser una propiedad de ciertas estructuras o restricciones*.

Snyder (2000) llevó a cabo un experimento en el que estudiantes sin formación en lingüística mostraron efectos de saciedad en tres de siete tipos de violaciones sintácticas. Esto no sólo avaló sus hipótesis, sino que hizo de los efectos de saciedad sintáctica un área de estudio dentro de la teoría lingüística.

En particular, Snyder sugirió que las violaciones "saciables" se deben a factores extragramaticales, como restricciones de procesamiento o límite de recursos computacionales: algunas violaciones de aceptabilidad tendrían un origen gramatical (y no pueden ser saciadas), mientras que otras construcciones tendrían una estructura gramatical bien formada pero serían inaceptables por limitaciones propias de los sistemas de actuación lingüísticos.

Los siete tipos de violación testeados por Snyder son los siguientes:

Table 1
Violations tested in Snyder 2000

Violation	Example sentence
Adjunct	Who did John talk with Mary after seeing?
Complex-NP	Who does Mary believe the claim that John likes?
Left-branch	How many did John buy books?
Subject	What does John know that a bottle of fell on the floor?
That-trace	Who does Mary think that likes John?
Want-for	Who does John want for Mary to meet?
Whether	Who does John wonder whether Mary likes?

Un problema que observa Sprouse (2009) es que los resultados de Snyder (2000) son muy difíciles de replicar.

Table 2
Summary of satiation results and the replication problem (√ = significant effect, (√) = marginal effect, — = not tested)

	Snyder 2000	Hiramatsu 2000	Goodall 2005	Sprouse 1	Sprouse 2	Sprouse 3
Adjunct						
Complex-NP	√		√			
Left-branch						
Subject	(√)	√				
That-trace		√				—
Want-for		√	—			—
Whether	√	√	—			

Lógica de Sprouse: si algunas violaciones producen saciedad en algunos experimentos y otras violaciones los producen en otros, hay que empezar a mirar más de cerca los diseños experimentales.

(17) *Diseño experimental de Snyder (2000)*

a. 50 oraciones presentadas en 5 bloques de 10 oraciones.
b. Cada bloque contenía una oración con una de las 7 violaciones (tres ítems de relleno).
c. En resumen: cada participante estuvo expuesto a 35 oraciones malas y 15 buenas.

¿Cómo se calculó la saciedad en el experimento de Snyder?

- Se contó el número de respuestas positivas para cada violación en los dos primeros bloques.
- Se contó el número de respuestas positivas para cada violación en los dos últimos bloques.
- Si el número de respuestas positivas aumentaba, había saciedad; si el número de respuestas positivas disminuía, no había saciedad.

Sprouse (2009) observa que hay un problema en este diseño experimental:

- Supongamos que a los participantes les llama la atención estar respondiendo el 70% del tiempo con NO y preferirían estar respondiendo de forma más pareja (50% SI, 50% NO).
- Al final del tercer bloque (antes de los dos bloques cruciales del experimento), los participantes escucharon 21 oraciones que llevaban al NO y sólo 9 oraciones que llevaban al SI.
- Si adoptan una estrategia de ecualización, van a tender a responder con más SIs en los próximos bloques.

¿Llega Sprouse a probar que Snyder está mal? Más o menos. Realiza 2 experimentos.

- **Experimento 1.** Un intento por replicar los datos de Snyder (2000). Ocho violaciones a lo largo de cinco bloques con dos ítems de relleno.
- **Experimento 2.** Experimento con rebalanceo: 4 violaciones y seis ítems de relleno por bloque.

Si bien en el *experimento 1* no se observó saciedad como la de Snyder, los datos resultan más *inestables* que los del *experimento 2*. En el *experimento 1*, 15 de 25 personas hicieron cambios de juicio; mientras que en el *experimento 2* sólo lo hicieron 2 de 19.