# Análisis Estadístico de Datos Aplicados al Estudio de Calidad en Servicios de Traducción

## Statistical Data Analysis Applied to the Study of Quality in Translation Services

**Analía Marta Pogliano**
Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Argentina
analiampoglino@gmail.com

## Abstract

This research studies the existing linguistic quality problems in translation of texts, from a leading company in translation services and its Client . The overall objective is to evaluate and estimate the values of the metrics required by the Client to ensure the expected quality.

Using logistic regression models to analyze data to find models that describe the audited linguistic errors in the forms of quality control for both the Client and the Company.

Also, using analysis of Canonical Correlations, the measurements between the two groups are studied, identifying the factors influencing the behavior of the errors.

The results indicate that the form of linguistic control between the Customer and the Company is not the same. The structures of correlation between the variables under study differ in each set of measurements , as well as the models adjusted for each group of measurements do not show the same significant variables.

**Keywords:** logistic regression, canonical correlation, translation services, language errors, linguistic quality control

## Resumen

En esta investigación se estudian los problemas de calidad lingüística existentes en traducciones de textos, entre una Empresa líder en servicio de traducción y su Cliente. El objetivo general es evaluar y estimar los valores de los parámetros de medición exigidos por el Cliente para garantizar la calidad esperada.

Mediante los modelos de Regresión Logística se analizan los datos para encontrar los modelos que mejor describan a los errores lingüísticos auditados en los formularios de control de calidad, tanto para el Cliente como para la Empresa.

Asimismo, empleando el análisis de Correlaciones Canónicas se estudian las asociaciones existentes entre las mediciones de los dos grupos, identificando los factores influyentes en el comportamiento de los errores.

Los resultados obtenidos indican que la modalidad de control lingüístico entre el Cliente y la Empresa no es la misma. Las estructuras de correlación entre las variables bajo estudio difieren en cada grupo de mediciones, como así también, los modelos ajustados para cada grupo de mediciones no presentan las mismas variables significativas.

**Palabras claves:** regresión logística, correlaciones canónicas, servicios de traducción, errores lingüísticos, control de calidad lingüístico.

# 1. INTRODUCCION

La Empresa ha presentado serios problemas en cuanto a la calidad de ciertos pedidos entregados a un Cliente en particular. Al tratarse de traducciones de textos, el control de calidad está basado de acuerdo a las normas estándares internacionales de LISA (Localization Industry Standards Association). Debido a esto, la calidad de las entregas por parte de la Empresa deben alcanzar los valores establecidos como "aceptables", según las normas internacionales.

Cliente emplea un proceso de control de calidad interno, con sus propios revisores y especialistas lingüísticos, quienes van a estar encargados de decidir si el pedido recibido presenta buena o mala calidad. De la misma manera, la Empresa cuenta con su departamento de control de calidad lingüística formado por revisores nativos de cada idioma localizado. Y ellos son los encargados de decidir si la traducción alcanza o no los niveles estándares de calidad.

La Empresa comenzó a recibir mal feedback del Cliente, haciendo referencia a la baja calidad en los pedidos entregados para ciertos idiomas, y que éstos no alcanzaban los valores estándares requeridos.

Esta noticia provocó gran incertidumbre y preocupación dentro de la Empresa, ya que todas las medidas de control lingüístico estaban siendo cumplidas, obteniéndose en la mayoría de los casos resultados positivos, garantizándole al Cliente una buena calidad.

Pero los resultados que la Empresa recibió en los reportes mensuales enviados por el Cliente no coincidían con los que ésta contaba. A pesar de los intentos de mejora por parte de la empresa, ésta no pudo alcanzar los parámetros de calidad exigido por el Cliente.

Por tal motivo se propuso investigar con sumo detalle, mediante la aplicación de métodos estadísticos el porqué de esta inconsistencia en los resultados. Una consideración importante a tener en cuenta para esta investigación es el cumplimiento del siguiente lema: "el Cliente siempre tiene la razón", por más que los datos y resultados demuestren lo contrario. Debido a esto, el feedback recibido y los resultados obtenidos por sus revisores deben ser aceptados y considerados como "lo correcto, lo ideal", es decir, en términos estadísticos, los datos del Cliente se deben tomar como grupo control.

# 2. MATERIALES Y METODOS

## 2.1. Materiales

La información recopilada para la creación de la base de datos proviene de formularios de control de calidad lingüísticos llamados LQA ("Lingusitic Quality Assurance"), pertenecientes a dos grupos: del Cliente y de la Empresa.

Estos formularios son completados por los revisores linguisticos, que, automáticamente, mediante fórmulas de cálculos que ponderan la cantidad de errores con sus correspondientes pesos de error, se determina el resultado final del control de calidad, esto es, si el resultado es aceptable ("Pass") o si es inaceptable ("Fail"). La metodología de evaluación es prácticamente la misma entre el Cliente y la Empresa, pero las variables medidas, en algunos casos, son diferentes. Esto se debe a la continua actualización de los formularios con el fin de mejorar las mediciones y obtener resultados más confiables.

## 2.2. Métodos Estadísticos

### *2.2.1. Análisis de Correlaciones Canónicas*

El análisis de correlaciones canónicas estudia las relaciones existentes entre dos grupos de variables. Investiga en detalle las interdependencias lineales entre dichos conjuntos, que pueden ser tratados simétricamente, o bien desempeñar un rol diferente en el análisis: un grupo de variables predictoras y otro de variables respuesta. Ambos conjuntos no deben ser independientes, se trata de descubrir "relaciones complejas" que reflejan la estructura existente entre ambos grupos de variables. El objetivo de esta técnica es resumir las asociaciones entre estos dos grupos de variables mediante la creación de nuevas variables a partir de las variables de cada grupo.

Sean las variables del primer grupo identificadas por *X* y las del segundo grupo identificadas por *Y*. Estas variables pueden pensarse distribuidas conjuntamente con esperanza nula (sin pérdida de generalidad), y matriz de covariancias, particionadas en cuatro submatrices (2.2.1).

$\underline{Z}' = (\underline{X}' \mid \underline{Y}') = (X_1 X_2 ... X_p \mid Y_1 Y_2 ... Y_q)$ distribuida conjuntamente con $E(\underline{Z}') = \underline{0}$ y $\underline{\underline{\Sigma}} = \left( \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right)$ (1)

Se desean construir dos combinaciones lineales U y V (2), determinando el conjunto de coeficientes de forma que la correlación entre U y V sea máxima. Por lo tanto deberá expresarse dicha correlación como función de los coeficientes.

Esto es:

$$\begin{aligned} U &= \alpha' X = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p \\ V &= \gamma' Y = \gamma_1 y_1 + \gamma_2 y_2 + ... + \gamma_q y_q \end{aligned} \qquad (2)$$

tales que:

$$\rho_{U,V} = \frac{E(U \cdot V)}{\sigma_U \sigma_V} = \frac{E(\alpha' X \cdot \gamma' Y)}{\sqrt{E(\alpha' X)^2} \cdot \sqrt{E(\gamma' Y)^2}} = \frac{\alpha' \Sigma_{XY} \gamma}{(\alpha' \Sigma_{XX} \alpha)(\gamma' \Sigma_{YY} \gamma)} \qquad (3)$$

Suponiendo que $\rho_1$ es la correlación máxima entre U y V, es decir: $\rho_1 = \max_{\underline{\alpha} \neq 0, \underline{\gamma} \neq 0} [corr(a'x, \gamma' y)]$ (4)

Luego, la primera correlación canónica entre X e Y se define por $\rho_1$.

Además, $U_1 = \underline{\alpha_1}' \underline{X}$ y $V_1 = \underline{\gamma_1}' \underline{Y}$ en donde $\alpha_1$ y $\gamma_1$ son los valores de α y γ que producen esta correlación máxima, se conocen como las primeras variables canónicas.

Sin pérdida de generalidad, se pueden elegir $\alpha_1$ y $\gamma_1$ de modo que Var ($U_1$) = Var ($V_1$) = 1

Sean ahora $U_2 = \underline{\alpha_2}' \underline{X}$ y $V_2 = \underline{\gamma_2}' \underline{Y}$ , en donde se eligen $\alpha_2$ y $\gamma_2$ de modo que:

1. $U_2$ y $V_2$ no están correlacionadas con $U_1$ y $V_1$.
2. Var ($U_2$) = Var ($V_2$) = 1 y
3. la correlación entre $\underline{\alpha_2}' \underline{X}$ y $\underline{\gamma_2}' \underline{Y}$ , denotada por $\rho_2$ es un máximo sobre todos los $\alpha_2$ y $\gamma_2$.

Entonces $\rho_2$ es la segunda correlación canónica y $U_2 = \underline{\alpha_2}' \underline{X}$ y $V_2 = \underline{\gamma_2}' \underline{Y}$ reciben el nombre de segundas variables canónicas. La cantidad real de correlaciones canónicas posibles es igual al

mínimo de *q* y *p-q*. La cantidad de correlaciones canónicas diferentes de cero es igual al rango de la matriz $\Sigma_{12}$.

### *2.2.2. Modelos de Regresión Logística Múltiple*

Los métodos de regresión son una componente integral de cualquier análisis de datos asociado con la descripción de la relación entre una variable respuesta y una o más variables explicativas, cuyo objetivo es encontrar el modelo que mejor ajuste los datos y que sea el más parsimonioso.

En modelos de regresión logística, a diferencia de los modelos de regresión lineal, la variable respuesta es binaria o dicotómica. Se desea conocer la relación entre:

- Una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial).
- Una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas.

Por sus características, los modelos de regresión logística permiten dos finalidades:

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente.
- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

Considerando un conjunto de p variables independientes las cuales serán denotadas con el vector $x' = (x_1, x_2, \ldots, x_p)$, la probabilidad condicional de que y tome el valor 1 (presencia de la característica estudiada), en presencia de las covariables X:

$$P(y=1/X)=\pi(x)=\frac{e^{g(x)}}{1+e^{g(x)}} \qquad (5)$$

Siendo $g(x)=\beta_0+\beta_1 x_1+\ldots+\beta_p x_p$ donde

- $\beta_0$ es la constante del modelo o término independiente
- p el número de covariables
- $\beta_i$ los coeficientes de las covariables
- $x_i$ las covariables que forman parte del modelo.

Si se divide la expresión (5) por su complemento, es decir, si se construye su odds se obtiene una expresión de más fácil manejo matemático: $\frac{P(y=1/X)}{1-P(y=1/X)}=\frac{\pi(x)}{1-\pi(x)}=e^{g(x)}$ (6)

Si ahora se realiza su transformación logarítmica con el logaritmo natural, se obtiene una ecuación lineal que es lógicamente de manejo matemático aún más fácil y de mayor comprensión:

$$\log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right)=\log(e^{g(x)})=g(x) \quad (7)$$

En la expresión (7) la primera igualdad es el llamado logit, es decir, el logaritmo natural de la odds de la variable dependiente (esto es, el logaritmo de la razón de proporciones de cometer un error al

traducir). El término a la derecha de la igualdad es la expresión lineal, idéntica a la del modelo general de regresión lineal: $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$ (8)

La importancia de esta transformación es que *g(x)* tiene muchas propiedades deseables de un modelo de regresión lineal. El logit, *g(x),* es lineal en sus parámetros, puede ser continuo, variando entre -∞ y +∞, según del rango de variación de *x*.

# 3. RESULTADOS

## 3.1. Análisis de Correlaciones Canónicas entre las mediciones del Cliente y de la Empresa

Es de interés estudiar si existen relaciones complejas entre los dos grupos de variables medidas por el Cliente y por la Empresa. Para ello se desea investigar las interdependencias lineales entre dichos conjuntos, que serán tratados como un grupo de variables predictoras (variables medidas por la Empresa) y otro de variables respuesta (variables medidas por el Cliente). Así, se podrá ver cómo estos dos conjuntos de mediciones se relacionan.

Las variables que se analizan para el estudio de correlaciones canónicas son: Cumplimiento, Significado, Puntuación, Terminología y Ediciones. Cada una de ellas mide la frecuencia de errores de cada tipo encontrados en los documentos traducidos. Los valores que pueden asumir estas variables son: 0, 1, 2,…etc.

Previo a este análisis, fue necesaria una transformación logarítmica en las variables para lograr que los datos sigan una distribución normal, supuesto requerido para esta técnica. Pero como las variables pueden tomar el valor 0, se recurre a la utilización de la transformación logarítmica: $\ln(x + 0.05)$

Una vez realizadas las transformaciones, se probó la normalidad de las variables, en forma individual, mediante el Test de Shapiro-Wilks y se concluyó que, trabajando con un nivel de significación del 5%, se cumple la normalidad de cada una de las variables transformadas, lo cual es una condición necesaria pero no suficiente de normalidad conjunta.

Se presenta el análisis de los resultados obtenidos mediante el procedimiento Proc Cancorr de SAS para el análisis de correlaciones canónicas:

Según la matriz de correlaciones entre las variables del Cliente, $\boldsymbol{R_{11}}$ (Tabla 1), se observa que existe asociación entre las variables. Es decir, los errores lingüísticos encontrados por el Cliente están correlacionados entre sí.

Tabla 1: Matriz de correlaciones entre las variables del Cliente $\boldsymbol{R_{11}}$

| Correlaciones entre Variables Cliente | | | | | | |
|---|---|---|---|---|---|---|
| | **cumplimiento** | **gramática** | **significado** | **puntuación** | **terminología** | **edición** |
| **cumplimiento** | 1.0000 | **0.3221** | **0.3434** | **0.2287** | **0.3304** | **0.2051** |
| **gramática** | 0.3221 | 1.0000 | **0.4395** | **0.3927** | **0.3005** | **0.2519** |
| **significado** | 0.3434 | 0.4395 | 1.0000 | **0.3046** | **0.3550** | **0.2185** |
| **puntuación** | 0.2287 | 0.3927 | 0.3046 | 1.0000 | **0.3151** | **0.2527** |
| **terminología** | 0.3304 | 0.3005 | 0.3550 | 0.3151 | 1.0000 | **0.1952** |
| **edición** | 0.2051 | 0.2519 | 0.2185 | 0.2527 | 0.1952 | 1.0000 |

En cambio, analizando la matriz de correlaciones entre las variables de la Empresa, $\boldsymbol{R_{22}}$ (Tabla 2), se observa los errores lingüísticos encontrados por la Empresa se podrían pensar que son independientes entre ellos (no hay asociacion).

Tabla 2: Matriz de correlaciones entre las variables de la Empresa $\boldsymbol{R_{22}}$

| | cumplimiento | gramática | significado | puntuación | terminología | edición |
|---|---|---|---|---|---|---|
| **cumplimiento** | 1.0000 | **0.1279** | **0.0214** | **0.0969** | **0.1147** | **0.0707** |
| **gramática** | 0.1279 | 1.0000 | **0.2318** | **0.1859** | **0.1526** | **0.0108** |
| **significado** | 0.0214 | 0.2318 | 1.0000 | **0.0764** | **0.1688** | **-0.0298** |
| **puntuación** | 0.0969 | 0.1859 | 0.0764 | 1.0000 | **0.1733** | **0.0391** |
| **terminología** | 0.1147 | 0.1526 | 0.1688 | 0.1733 | 1.0000 | **0.1775** |
| **edición** | 0.0707 | 0.0108 | -0.0298 | 0.0391 | 0.1775 | 1.0000 |

Si se analiza la matriz de correlaciones entre las variables del Cliente y de la Empresa, $\boldsymbol{R_{12}}$, (Tabla 3) se distingue que la variable Gramática de la Empresa está altamente correlacionada con las variables Significado, Puntuación y Terminología del Cliente. Es decir, los errores gramaticales encontrados por la Empresa están asociados con los errores de significado, de puntuación y de terminología encontrados por el Cliente.

Estos valores indican que las estructuras de las variables entre el Cliente y la Empresa son diferentes, a pesar que ambos grupos de variables estén midiendo lo mismo, sobre la misma muestra. Es por ello que los resultados obtenidos por uno u otro grupo son tan diferentes.

Tabla 3: Matriz de correlaciones entre las variables del Cliente $\boldsymbol{R_{12}}$

| | Cumplimiento_E | Gramática_E | Significado_E | Puntuación_E | Terminología_E | Edicion_E |
|---|---|---|---|---|---|---|
| **Cumplimiento_C** | 0.0760 | 0.2421 | 0.1462 | 0.0974 | 0.2511 | 0.0507 |
| **Gramática_C** | 0.1162 | 0.2395 | 0.1118 | 0.1743 | 0.1024 | 0.1164 |
| **Significado_C** | 0.1645 | **0.2993** | 0.1427 | 0.0730 | 0.0824 | 0.0766 |
| **Puntuación_C** | 0.1438 | **0.3272** | 0.1260 | 0.1369 | 0.0445 | 0.1812 |
| **Terminología_C** | 0.0995 | **0.2971** | 0.0770 | 0.1531 | 0.1174 | 0.1300 |
| **Ediciones_C** | 0.1241 | 0.1386 | 0.0337 | 0.1169 | -0.0068 | 0.0924 |

En base al Test de significación de correlaciones canónicas, presentado en la tabla 4, mediante la estadística de Wilks, se observa que la primera correlación canónica es la única significativa, a un nivel de significación $\alpha = 0.05$. Todas las restantes correlaciones, a partir de la segunda, no fueron significativas, con un valor de p-asociado = 0.3692. Es por ello que se cuenta con un solo par de variables canónicas, cuya interpretación es relevante al estudio.

Tabla 4: Test para la significación de las correlaciones canónicas

| | Razón de verosimilitud | F | Gl numerador | Gl denominador | Pr > F |
|---|---|---|---|---|---|
| **1** | **0.68367539** | **2.75** | **36** | **1096.2** | **<.0001** |
| **2** | 0.89989808 | 1.07 | 25 | 930.21 | 0.3692 |
| **3** | 0.96634623 | 0.54 | 16 | 767.46 | 0.9261 |
| **4** | 0.98958609 | 0.29 | 9 | 613.45 | 0.9765 |
| **5** | 0.99669622 | 0.21 | 4 | 506 | 0.9332 |
| **6** | 0.99950544 | 0.13 | 1 | 254 | 0.7232 |

El primer y único par de variables canónicas significativas está dado por las siguientes combinaciones lineales:

$$U_1 = Cliente1 = 0.0866\ \text{Cumplimiento} + 0.0704\ \text{Gramática} + 0.1367\ \text{Significado} + 0.2514\ \text{Puntuación} + 0.1740\ \text{Terminología} + 0.0352\ \text{Ediciones}$$

$$V_1 = Empresa1 = 0.1692\ \text{Cumplimiento} + 0.3856\ \text{Gramática} + 0.0864\ \text{Significado} + 0.1294\ \text{Puntuación} + 0.0093\ \text{Terminología} + 0.2297\ \text{Ediciones}$$

Observando las cargas canónicas entre las variables del Cliente y sus variables canónicas Cliente1 en la tabla 5, éstas indican que del grupo de variables del Cliente, P*untuación*, T*erminología* y S*ignificado* son las más importantes dentro de la combinación lineal $U_1$.

Tabla 5: Correlaciones entre las variables del Cliente y Empresa y sus respectivas variables canónicas.

| Cliente/Empresa | Cliente1 ($U_1$) | Empresa1 ($V_1$) |
|---|---|---|
| **cumplimiento** | 0.5375 | 0.3800 |
| **gramática** | 0.6123 | **0.8631** |
| **significado** | **0.6705** | 0.3431 |
| **puntuación** | **0.7950** | 0.3775 |
| **terminología** | **0.6841** | 0.2786 |
| **ediciones** | 0.3934 | 0.3756 |

De igual manera, se presentan las cargas canónicas entre las variables de la Empresa y su variable canónica Empresa1 ($V_1$). En este caso, para el grupo de las variables de la Empresa, la única variable que aporta mucha importancia dentro de la combinación lineal es la variable *Gramática*.

Para ver la relación entre una variable de un conjunto, con la variable canónica del otro conjunto, se analizan las cargas cruzadas entre variables originales y variables canónicas (Tabla 6). Nuevamente se observa que, para el grupo de las variables del Cliente, las variables con más peso sobre el grupo de variables canónicas de la Empresa son P*untuación*, T*erminología* y S*ignificado;* y para el grupo de las variables de la Empresa, la variable con más importancia sobre el grupo de las variables canónicas del Cliente es *Gramática.*

Tabla 6: Cargas canónicas cruzadas para Cliente y para Empresa

| **Correlaciones entre las variables del Cliente y la variable canónica de la Empresa** | |
|---|---|
| **Cliente** | **Empresa1** |
| **Cumplimiento_C** | 0.2635 |
| **Gramática_C** | 0.3001 |
| **Significado_C** | 0.3287 |
| **Puntuación_C** | 0.3897 |
| **Terminología_C** | 0.3353 |
| **Edicion_C** | 0.1928 |

| **Correlaciones entre las variables de la Empresa y la variable canónica del Cliente** | |
|---|---|
| **Empresa** | **Cliente1** |
| **Cumplimiento_E** | 0.1863 |
| **Gramática_E** | 0.4231 |
| **Significado_E** | 0.1682 |
| **Puntuación_E** | 0.1850 |
| **Terminología_E** | 0.1366 |
| **Edicion_E** | 0.1841 |

Se determina que sí existe una relación entre las variables del Cliente y la Empresa; la misma está dada entre las variables Puntuación, Terminología y Significado, medidas por el Cliente, y la variable Gramática, medida por la Empresa. Éstas han demostrado ser variables que hacen un aporte mayor a las interdependencias lineales de cada grupo. Así, en los documentos en los que el Cliente distingue más errores de puntuación, terminología y significado, la Empresa encuentra errores gramaticales.

Otro aspecto que deja en evidencia este análisis es que las estructuras de las variables bajo estudio difieren en cada grupo de mediciones. Para el caso de las variables del Cliente, las mediciones de los errores lingüísticos presentan cierta asociación entre ellas. En cambio, para el caso de la Empresa, no existe asociación entre los distintos tipos de errores. Es por ello que los resultados de los controles de calidad lingüística pertenecientes a un mismo documento analizado por la Empresa y el Cliente a menudo difieren sustancialmente.

## 3.2. Modelo de Regresión Logística Múltiple

Mediante el análisis de regresión logística se desea encontrar el modelo que mejor ajuste los datos, cuantificando la importancia de la relación existente entre cada una de las covariables y la variable dependiente.

En base a los datos de la investigación, se cuenta con dos grupos de variables: las medidas por el Cliente y las medidas por la Empresa. Entonces es necesario encontrar dos modelos de regresión logística que mejor describan la variable respuesta dicotómica "Resultado", para luego compararlos y ver cómo difieren entre ellos y qué tan distintas son las estimaciones de los coeficientes de los modelos.

Asimismo, es de interés analizar cuáles son las mediciones realizadas por la Empresa que tienen efecto sobre el resultado del Cliente. Es decir, modelar la probabilidad de rechazo del documento por parte del Cliente ("Fail") en base a las mediciones de la Empresa. La variable respuesta "Resultado" toma valores 0 = "Pass" y 1 = "Fail".

### *3.2.2. Ajuste del modelo para los Datos del Cliente*

Utilizando el procedimiento Logistic de SAS, mediante el método de Selección de variables hacia atrás "Backward", el modelo resultante presenta los efectos de las variables referidas a los errores de terminología, cumplimiento, gramática, significado, año, cantidad de palabras analizadas (conteo) y las interacciones de esta última con los errores de gramática, cumplimiento y significado. (Tabla 6)

Tabla 6: Efectos incluidos en el modelo final y sus significancias

| Efectos | GL | Wald Chi-Cuadrado | Pr > ChiSq |
|---|---|---|---|
| **cumplimiento** | 1 | 7.0214 | 0.0081 |
| **gramática** | 1 | 0.6989 | 0.4032 |
| **significado** | 1 | 3.6531 | 0.0560 |
| **terminología** | 1 | 5.4174 | 0.0199 |
| **conteo** | 1 | 0.8033 | 0.3701 |
| **año** | 1 | 4.8186 | 0.0282 |
| **cumpli*conteo** | 1 | 4.8533 | 0.0276 |
| **grama*conteo** | 1 | 7.7957 | 0.0052 |
| **sig*conteo** | 1 | 9.1734 | 0.0025 |

Además, el valor de la estadística Chi-Cuadrado en la Test de Hipótesis Global, mediante el Test de Razón de Verosimilitud, fue de 109.83, con 9 grados de libertad, la cual resultó ser significativo a un nivel $\alpha$=1%. Esta prueba indica que las variables predictoras que se están usando son variables estadísticamente significativas del resultado de los LQAs analizados.

Las estimaciones de los parámetros del modelo obtenidas por el método de máxima verosimilitud nos permiten obtener la función logit g (x) estimada como:

$$\hat{g}(x) = -4.4352 + 0.2931\,cumplimien\,to + 0.1051\,gramática + 0.1563\,significad\,o + 0.2357\,terminilog\,ía + 0.8389\,conteo + 1.2169\,año + 0.9359\,cumplimien\,to * conteo + 2.9701\,gramática * conteo + 1.4724\,significad\,o * conteo$$

Los valores de dichas estimaciones se ven resumidos en la Tabla 7.

Tabla 7: Estimaciones de los coeficientes del modelo del Cliente

| Parámetro | | GL | Estimador | Error Estándar | Chi-Cuad | Pr > ChiSq |
|---|---|---|---|---|---|---|
| **Intercepto** | | 1 | -4.4352 | 0.6477 | 46.8840 | <.0001 |
| **cumplimiento** | | 1 | 0.2931 | 0.1106 | 7.0214 | 0.0081 |
| **gramática** | | 1 | 0.1051 | 0.1257 | 0.6989 | 0.4032 |
| **significado** | | 1 | 0.1563 | 0.0818 | 3.6531 | 0.0560 |
| **terminología** | | 1 | 0.2357 | 0.1013 | 5.4174 | 0.0199 |
| **conteo** | **A** | 1 | 0.8389 | 0.9360 | 0.8033 | 0.3701 |
| **año** | **2008** | 1 | 1.2169 | 0.5544 | 4.8186 | 0.0282 |
| **cumpli*conteo** | **A** | 1 | 0.9359 | 0.4248 | 4.8533 | 0.0276 |
| **grama*conteo** | **A** | 1 | 2.9701 | 1.0637 | 7.7957 | 0.0052 |
| **sig*conteo** | **A** | 1 | 1.4724 | 0.4861 | 9.1734 | 0.0025 |

La interpretación de los coeficientes del modelo de regresión logística se hace en términos de la razón de odds, cuya expresión es

$$\hat{\theta} = \frac{\left[\frac{\pi(x+1)}{1-\pi(x+1)}\right]}{\left[\frac{\pi(x)}{1-\pi(x)}\right]} \Rightarrow \ln\hat{\theta} = \hat{g}(x+1) - \hat{g}(x)$$

A continuación se presentan las razones de odds correspondientes a cada tipo de error (Tabla 8):

Tabla8: Razones de Odds para Cliente

| Variables | Conteo <1000 | Conteo > 1000 |
|---|---|---|
| Cumplimiento | 3.41 | 1.34 |
| Gramática | 21.65 | 1.11 |
| Significado | 5.10 | 1.17 |
| Terminología | 1.266 | |
| Año | 3.377 | |

Un aspecto relevante que se observa en este análisis es la gran influencia sobre los resultados finales que aporta la cantidad de palabras revisadas (variables cumplimiento, gramática y significado). Esto significa que la importancia de los errores de significado, gramática o cumplimiento se consideran más “severos” cuando se analizan menos palabras.

También se observa que el año marca una gran diferencia en el resultado. Depende el año de realización del QA, la chance aumenta 3 veces de un año a otro. Esto responde a uno de los interrogantes planteados previamente en esta investigación.

### *3.2.3. Ajuste del modelo para los Datos de la Empresa*

Utilizando el procedimiento Logistic de SAS, mediante el método de Selección de variables hacia atrás "Backward", el modelo resultante presenta los efectos de las variables referidas a los errores de terminología, cumplimiento, gramática, significado, conteo, año y la interacción entre significado y la cantidad de palabras (conteo). (Tabla 9)

Tabla 9: Efectos incluidos en el modelo final y sus significancias

| Efectos | GL | Chi-Cuadrado | Pr > ChiSq |
|---|---|---|---|
| **cumplimiento** | 1 | 4.8451 | 0.0277 |
| **gramática** | 1 | 16.0914 | <.0001 |
| **significado** | 1 | 1.5443 | 0.2140 |
| **terminología** | 1 | 16.7958 | <.0001 |
| **conteo** | 1 | 17.0856 | <.0001 |
| **año** | 1 | 4.9855 | 0.0256 |
| **sig*conteo** | 1 | 8.7700 | 0.0031 |

La medida global de Bondad de Ajuste del modelo, mediante el Test de Hipótesis Global indica un buen ajuste del modelo con las variables que fueron seleccionadas, obteniéndose un valor de la estadística Chi-Cuadrada de 99.15, con 7 grados de libertad, cuya probabilidad asociada es menor a 0.001, siendo significativa a un nivel de α del 1%.

Las estimaciones de los parámetros del modelo (ver tabla 10) obtenidas por el método de máxima verosimilitud nos permiten obtener la función logit g (x) estimada como:

$$\hat{g}(x) = -3.6931 + 0.5062 cumplimiento + 0.3000 gramática + 0.1541 significado + 0.6332 terminilogía + 2.1761 conteo + 1.0399 año + 1.1364 significado * conteo$$

Tabla 10: Estimaciones de los coeficientes del modelo de la Empresa

| Parámetro | | DF | Estimador | Error Estándar | Chi-Cuad | Pr > ChiSq |
|---|---|---|---|---|---|---|
| **Intercepto** | | 1 | -3.6931 | 0.4968 | 55.2590 | <.0001 |
| **cumplimiento** | | 1 | 0.5062 | 0.2300 | 4.8451 | 0.0277 |
| **gramática** | | 1 | 0.3000 | 0.0748 | 16.0914 | <.0001 |
| **significado** | | 1 | 0.1541 | 0.1240 | 1.5443 | 0.2140 |
| **terminología** | | 1 | 0.6332 | 0.1545 | 16.7958 | <.0001 |
| **conteo** | **A** | 1 | 2.1761 | 0.5264 | 17.0856 | <.0001 |
| **año** | **2008** | 1 | 1.0399 | 0.4657 | 4.9855 | 0.0256 |
| **sig*conteo** | **A** | 1 | 1.1364 | 0.3837 | 8.7700 | 0.0031 |

De la misma manera que se analizara anteriormente para el modelo del Cliente, la interpretación de los coeficientes del modelo en regresión logística para las mediciones realizadas por la Empresa, se hace en términos de la razón de odds (Tabla11).

Tabla11: Razones de Odds para Empresa

| Variables | Conteo <1000 | Conteo > 1000 |
|---|---|---|
| Cumplimiento | 1.659 | |
| Gramática | 1.350 | |
| Significado | 3.634 | 1.17 |
| Terminología | 1.884 | |
| Año | 2.829 | |

A diferencia del modelo, si bien se observa el mismo tipo de efecto, el conteo no influye con la misma intensidad. Si se comparan las razones de Odds correspondientes a los errores de significado, para el Cliente se observa que la chance de Fail es 5 veces mayor al incrementarse en un error, mientras que para la Empresa es 4 veces mayor. Asimismo, en las mediciones del Cliente la cantidad de palabras interactúa también con los errores de gramática y cumplimiento.Con respecto al Año, se observa lo mismo para ambos grupos. La chance de obtener un rechazo aumenta 3 veces de un año a otro.

### *3.2.4. Análisis de la respuesta del Cliente en función de las mediciones de la Empresa*

Es de interés analizar cuáles son las mediciones realizadas por la Empresa que tienen efecto sobre el resultado del Cliente. El objetivo es modelar la probabilidad de rechazo del documento por parte del Cliente (“Fail”) en base a las mediciones de la Empresa realizadas sobre los documentos.

Se lleva a cabo el mismo procedimiento que en los análisis precedentes obteniéndose un modelo final que incluye los efectos de los errores de gramática, terminología, la cantidad de palabras evaluadas (conteo) y la interacción de estos últimos dos. (Tabla 12).

Tabla 12: Efectos incluidos en el modelo final y sus significancias

| Efectos | GL | Chi-Cuadrado | Pr > ChiSq |
|---|---|---|---|
| **gramática** | 1 | 7.4872 | 0.0062 |
| **terminología** | 1 | 0.2827 | 0.5949 |
| **conteo** | 1 | 1.7142 | 0.1904 |
| **term*conteo** | 1 | 4.3206 | 0.0377 |

Además, el valor de la estadística Chi-Cuadrado en la Test de Hipótesis Global, mediante el Test de Razón de Verosimilitud, fue de 16.21, con 4 grados de libertad, la cual resultó ser significativa a un nivel α=1%. Esta prueba indica que las variables predictoras que se están usando son variables estadísticamente significativas del resultado de los LQAs analizados.

Las estimaciones de los parámetros del modelo (Tabla 13) obtenidas por el método de máxima verosimilitud nos permiten obtener la función logit g (x) estimada como:

$$\hat{g}(x) = -2.2167 + 0.1551\,gramática\ - 0.0891\,terminolog\,ía + 0.5851\,conteo + 0.7605\,terminolog\,ía * conteo$$

Tabla 13: Estimaciones de los coeficientes del modelo Cliente-Empresa

| Parámetro | DF | Estimador | Error Estándar | Chi-Cuad | Pr > ChiSq |
|---|---|---|---|---|---|
| **Intercepto** | 1 | -2.2167 | 0.3375 | 43.1390 | <.0001 |
| **conteo** | 1 | 0.5851 | 0.4469 | 1.7142 | 0.1904 |
| **gramática** | 1 | 0.1551 | 0.0567 | 7.4872 | 0.0062 |
| **terminología** | 1 | -0.0891 | 0.1675 | 0.2827 | 0.5949 |
| **term*conteo** | 1 | 0.7605 | 0.3659 | 4.3206 | 0.0377 |

La interpretación de los coeficientes del modelo en regresión logística para el resultado del Cliente y las mediciones realizadas por la Empresa, se hace en términos de la razón de odds (Tabla 14).

Tabla 14: Razones de Odds Cliente/Empresa

| Variable | Conteo <1000 | Conteo >1000 |
|---|---|---|
| Gramática | 1.17 | |
| Terminología | 1.92 | 0.92 |

Este modelo pone en evidencia que la aceptación o rechazo de la traducción de un documento por parte del Cliente sólo se relaciona con los errores de gramática y terminología hallados por la Empresa. Por cada error de gramática que se encuentra en el texto evaluado, la chance de ser rechazado por el Cliente se incrementa un 17%. El efecto de un error de terminología es importante sólo cuando el número de palabras evaluadas es inferior a 1000.

# 4. CONCLUSIONES

El análisis de Correlaciones Canónicas determina que existe una relación entre las variables del Cliente y la Empresa; la misma está dada entre las variables Puntuación, Terminología y Significado, medidas por el Cliente, y la variable Gramática, medida por la Empresa.

Éstas han demostrado ser variables que hacen un aporte mayor a las interdependencias lineales de cada grupo. Así, en los documentos en los que el Cliente distingue más errores de puntuación, terminología y significado, la Empresa encuentra errores gramaticales. Otro aspecto que deja en evidencia este análisis es que las estructuras de correlación entre las variables bajo estudio difieren en cada grupo de mediciones. Es por ello que los resultados de los controles de calidad lingüística pertenecientes a un mismo documento analizado por la Empresa y el Cliente a menudo difieren sustancialmente.

Mediante el análisis de Regresión Logística, para los datos del Cliente, un aspecto relevante que se observa es la gran influencia sobre los resultados finales que aporta la cantidad de palabras revisadas.

A diferencia del modelo anterior, para los datos del Empresa, si bien se observa el mismo tipo de efecto, el conteo no influye con la misma intensidad.

Analizando cuáles son las mediciones realizadas por la Empresa que tienen efecto sobre el resultado del Cliente, el modelo obtenido pone en evidencia que la aceptación o rechazo de la traducción de un documento por parte del Cliente sólo se relaciona con los errores de gramática y terminología hallados por la Empresa.

## Referencias

[1] Beltrán, Celina, “Modelización lingüística e información estadística”- 1ra ed.- Rosario-Juglaría, 2009.

[2] Dallas E. Johnson, “Métodos Multivariados aplicados al análisis de datos”. 1998.

[3] Eriksson L., Johansson E., Kettanen N. –Wold and S. Wold, “Multi and Megavariate Data Analysis, Principles and Applications”. 1991-2001.

[4] Grupo de Investigación Traducción, literatura y sociedad, “Ética y política de la traducción literaria”. Volumen de colección Itaca. Miguel Gómez Ediciones 2004.

[5] Hosmer, David W. y Lemeshow, Stanley. “Applied Logistic Regression”. Wisley Series in Probability and Mathematical Statistics.

[6] Khattree, Ravindra y Dayanand N. Naik, “Multivariate Data Reduction and Discrimination with SAS Software, Cary, NC: SAS Institute Inc. 2000.

[7] Lomprecht, James L, “Applied Data Analysis for Process Improvement, A practical Guide to Six Sigma Block Belt Statistics”. 2005.

[8] Mauly, Bryan F.J., “Multivariate Statistical Methods”. 1986-2004.

[9] Waddignton, Christopher, “Estudio Comparativo de Diferentes Métodos de Evaluación de traducción general”. 1999-, Madrid.