Recueillir et exploiter des données en linguistique

Pascal Amsili

ILPGA – Lattice

version présentée le 8 avril 2026

Sorbonne Nouvelle

Plan

Les bases d'une généralisation

Généralisation

Il y a 7 099 langues dans le monde.

Les bases d'une généralisation

Généralisation

Il y a 7 099 langues dans le monde.

1. consensus sur la définition des termes/concepts
2. support empirique : observations à l'appui de la généralisation

Les bases d'une généralisation

Généralisation

Il y a 7 099 langues dans le monde.

1. consensus sur la définition des termes/concepts
2. support empirique : observations à l'appui de la généralisation
   = des données

Les bases d'une généralisation

Généralisation

Il y a 7 099 langues dans le monde.

1. consensus sur la définition des termes/concepts
2. support empirique : observations à l'appui de la généralisation
   = des données

- la théorie doit énoncer des propositions **falsifiables**
- la théorie doit être validée par des observations **reproductibles**

Sorbonne Nouvelle

Falsifiabilité

Proposition 1

Toutes les phrases du français comportent un verbe conjugué.

Falsifiable ?

Falsifiabilité

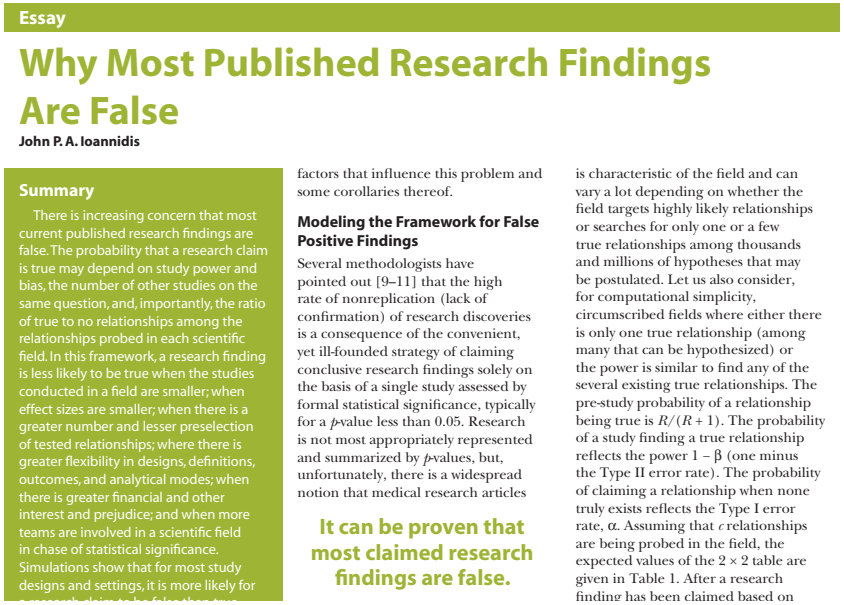
Proposition 1

Toutes les phrases du français comportent un verbe conjugué.

Falsifiable ?

On préfère une proposition falsifiable non (encore) falsifiée à une proposition non falsifiable.

Reproductibilité



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on

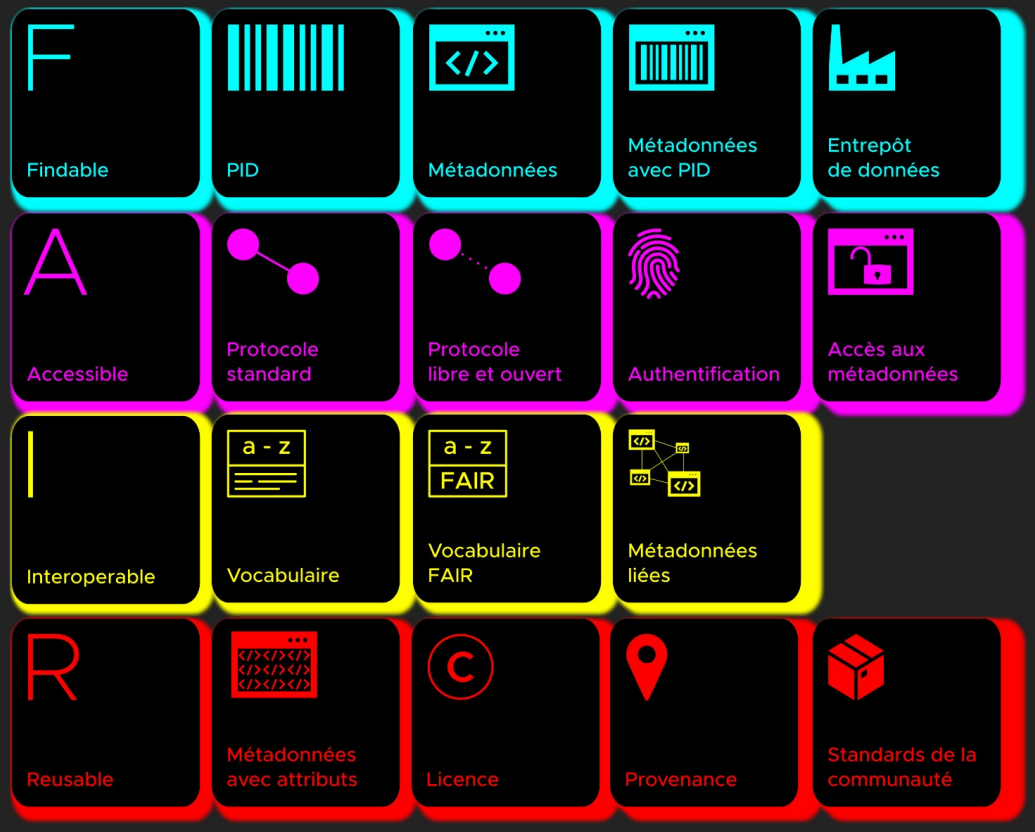
(Ioannidis, 2005)

Selon une étude réalisée auprès de 1 500 scientifiques et publiée par Nature en 2016, plus de 70 % des chercheurs affirment avoir été incapables de reproduire une expérience scientifique d'un autre chercheur et plus de la moitié affirment avoir échoué à reproduire une de leur propre expérience.

Sorbonne Nouvelle

Les principes FAIR

Les chercheurs s'appuient sur les connaissances scientifiques antérieures, notamment sur les résultats publiés dans les articles scientifiques.
La reproductibilité des résultats, ainsi que leur croisement, ne sont cependant envisageables qu'avec des données originelles et leurs conditions d'obtention.
C'est pourquoi la science ouverte vise à faciliter l'accès aux publications scientifiques et aux données de la recherche.
Cette facilitation s'accompagne d'un certain nombre de mesures pour rendre les données scientifiques facilement découvrables, accessibles, interopérables et réutilisables.
Ce sont les principes FAIR : Findable, Accessible, Interoperable, Reusable.



(Urfist Méditerrannée, 2019)

Plan

Sorbonne Nouvelle

Plan

Données introspectives

Jugements produits par les chercheurs ou des informateurs

- Grammaticalité
- Anomalie sémantique
- Relation sémantique (dont la paraphrase)
- Pertinence pragmatique
- Identité/différence

Données de corpus

Productions « naturelles » rassemblées par les chercheurs

- Productions langagières publiées (littérature, presse, discours...)
- Enregistrements oraux ou vidéo (multimodaux)
- Traces langagières d'activité (communication professionnelle, réseaux sociaux...)

Données expérimentales

Données recueillies dans un cadre contrôlé

- Productions langagières induites
- Jugements (grammaticalité, anomalie)
- Mesures psycho-physiques (temps de réaction, mouvements occulaires...)
- Enregistrements d'imagerie cérébrale (eeg, irm)

Données de terrain

Données recueillies dans un contexte spacial et temporel spécifique

- Enregistrements
- Enquêtes
- Entretiens

Plan

Données introspectives

- Introspection des chercheurs, publication et échange entre linguistes
- « Chasse aux papillons »
- Recours à des informateurs
- Fabrication de données critiques

Données de corpus

- Nombreux corpus existants, toutes modalités (science ouverte)
- On constitue fréquemment un (sous) corpus pour une recherche spécifique
- Toute collection de données linguistiques naturelles ne constitue pas un corpus

Qu'est-ce qu'un corpus ?

Un corpus doit être :

- Représentatif
- Fini
- Numérisé
- Standard
- (Finalisé/stabilisé)

Données expérimentales

- Méthodes inspirées de la psychologie cognitive
- Champ de la psycho-linguistique
- Principe général : contrôle des variable d'intérêt ("toutes choses égales par ailleurs")
- Chaque expérience a une portée étroite, mais cumulative
- Pratique récente : *crowdsourcing* (externalisation ouverte, myriadisation)
- Autre pratique récente : utilisation des LLMs comme sujets expérimentaux

Données de terrain

- Conditions de collecte souvent très spécifiques
- Dimension sociologique ou anthropologique déterminante
- Méthode partagée avec d'autres sciences sociales : possibilité de recherche croisée
- Les données collectées peuvent être introspectives, quantitatives (corpus) et même expérimentales

Plan

Exploitation

selon catégories : exploitations distinctes, méthodes distinctes :

Données introspectives souvent exploitées dans une perspective catégorique, pour répondre à la question “est-ce que telle constuction est possible ou pas ?”

Données de corpus souvent exploitées pour répondre à des questions quantitatives : “telle forme est-elle plus répandue que telle autre”, ou “telle construction est-elle (fréquemment) associée à telle propriété ?”.
L’exploitation des données de corpus repose souvent sur une phase d’annotation

Données expérimentales très fréquemment exploitées pour trancher entre deux théories concurrentes qui font des prédictions incompatibles
N.B. Dans ces deux cas, rôle crucial : significativité statistique

Données de terrain en plus des usages semblables aux autres types de données, les données de terrain ont fréquemment un rôle patrimonial ou de documentation.

Sorbonne Nouvelle

Références

Amsili, Pascal, Ellsiepen, Emilia, & Winterstein, Grégoire. 2016. Optionality in the use of 'too' : The role of reduction and similarity. *Revista da Abralin (Associação Brasileira de Linguística)*, **15**(1), 229–252.

Fillmore, Charles. 1992. "Corpus linguistics" or "Computer-aided armchair linguistics". *Pages 35–60 of :* Svartvik, Jan (ed), *Directions in corpus linguistics.* Trends in Linguistics. Studies and Monographs, vol. 65. Berlin and New York : Mouton de Gruyter. Proceedings of the Nobel Symposium 82 Stockholm, 4-8 August 1991.

Fort, Karën. 2011. *Corpus Linguistics : history*. Diapos présentées le 18 novembre 2011. Inist.

Green, Georgia M. 1968. On *too* and *either*, and not just *too* and *either*, either. *Pages 22–39 of : CLS (Chicago Linguistics Society)*, vol. 4.

Ioannidis, John P. A. 2005. Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8).

Kaplan, Jeff. 1984. Obligatory *too* in English. *Language*, **60**(3), 510–518.

McEnery, Tony, & Wilson, Andrew. 1996. *Corpus Linguistics*. Edinburgh University Press.

Poinsot, Denis. 2004. *Statistiques pour statophobes*. En ligne : `http://perso.univ-rennes1.fr/denis.poinsot`.

Tellier, Isabelle. 2015. *Introduction à la fouille de textes*. Université Paris 3 – Sorbonne Nouvelle. Polycopié, master pluriTAL.

Urfist Méditerrannée. 2019. *DoRANum-Enjeux et bénéfices : les principes FAIR*. DoRANum.

Sorbonne Nouvelle