

## Faktoren der Kognatenerkennung in der germanischen Interkomprehension

**Auteur :** Schwall, Danny

**Promoteur(s) :** Möller, Robert

**Faculté :** Faculté de Philosophie et Lettres

**Diplôme :** Master en langues et lettres modernes, orientation germaniques, à finalité approfondie

**Année académique :** 2019-2020

**URI/URL :** <http://hdl.handle.net/2268.2/10514>

---

### *Avertissement à l'attention des usagers :*

*Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.*

*Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.*

---



Faculté de Philosophie et Lettres

# Faktoren der Kognatenerkennung in der germanischen Interkomprehension

Masterarbeit von Danny Schwall

Promoteur: Robert Möller

Master en langues et lettres modernes :  
orientation germanique

Année académique 2019-2020



## **Danksagung**

Ich möchte mich ganz herzlich bei all denen bedanken, die mich beim Schreiben dieser Masterarbeit unterstützt haben.

Ein besonderes Dankeschön geht an Professor Robert Möller, der mir immer mit guten Ratschlägen und konstruktiver Kritik zur Seite gestanden hat.

Des weiteren möchte ich mich bei Daniel Lybeer bedanken. Seine informatische Expertise hat diese Arbeit erst in diesem Maße möglich gemacht.

Schließlich gilt mein Dank auch noch an meinen Eltern und meiner Freundin, die mich immer unterstützt und motiviert haben.

## Inhaltsverzeichnis

1. Einleitung.....	4
2. Theoretische Hintergründe.....	5
3. Bisherige Interkomprehensions-Forschung.....	14
4. Extralinguistische Faktoren der Kognaten-Erkennung.....	21
5. Das <i>Interactive Activation Model</i> (IAM).....	24
6. Kontext und Kotext.....	26
7. Ziel dieses Worterkennungstests.....	27
8. Testaufbau.....	27
9. Auswahl potentieller Faktoren.....	28
10. Verarbeitung der Testresultate.....	34
11. Nicht-systematische Faktoren der Kognaten-Erkennung.....	75
12. Diskussion und Fazit.....	83
13. Literaturverzeichnis.....	85
14. Anhang.....	87

## 1. Einleitung

Die Verständigung zwischen Menschen unterschiedlicher Sprachräume stellte schon immer ein ziemlich großes Problem dar. Diese Problematik schien sogar von solchem Interesse, dass in der Bibel die Existenz unterschiedlicher Sprachen und die damit verbundenen Kommunikationsschwierigkeiten in der Geschichte um den Turmbau zu Babel thematisiert wurden. Dieser Schöpfungsmythos der Sprachen zeigt, dass die Menschen schon damals wussten, wie wichtig Verständigung für erfolgreiche Kooperation ist. In vielen Fällen hat sich in Situationen, in denen sich Personen verschiedener Kulturen austauschen mussten, eine Handelssprache durchgesetzt, die als Basis der Kommunikation dienen sollte. Dies ist auch heute noch der Fall und die internationale politische und wirtschaftliche Kommunikation findet auf Englisch statt. Es gibt jedoch durchaus Fälle, in denen die Mitglieder einer Sprachengemeinschaft mehr oder weniger problemlos mit Fremdsprachigen kommunizieren können, ohne dass eine *Lingua franca* verwendet werden muss oder eine der beiden Personen die Sprache des jeweils anderen erlernt hat.

Ten Thije und Zeevaert gehen davon aus, dass Englisch als *Lingua Franca* in Europa teilweise durch rezeptive Mehrsprachigkeit ersetzt werden könnte (siehe Ten Thije & Zeevaert 2007). *Rezeptive Mehrsprachigkeit* (*receptive multilingualism*) oder *passive Mehrsprachigkeit* (siehe Arntz 1997) bezeichnet das Verstehen einer Sprache durch einen/eine Sprecher/in einer anderen Sprache, ohne dass diese Person unbedingt Vorkenntnisse dieser anderen Sprache besitzt (vgl. auch Riionheimo et al. 2017). Ein sehr schönes Beispiel dafür ist die Kommunikation zwischen Norwegern und Schweden, die sich ohne größere Probleme miteinander unterhalten können, wenn jede Person ihre eigene Muttersprache spricht. Ein weiteres Beispiel ist die Kommunikation zwischen Luxemburgisch-Sprachigen und Einwohnern der deutschsprachigen Region Belgiens, die Plattdeutsch sprechen. Solche intersprachliche Kommunikation ist jedoch an gewisse Faktoren geknüpft.

Die Forschung zur rezeptiven Mehrsprachigkeit hat sich mit verschiedenen Aspekten der Kommunikation zwischen Personen mit verschiedenen Muttersprachen beschäftigt. Dabei kann der Fokus entweder auf der geschriebenen oder der gesprochenen Sprache liegen. Außerdem kann man zwischen dem Verständnis ganzer Sätze oder sogar Texte und dem Verständnis einzelner Buchstaben unterscheiden. In dieser Arbeit wird vor allem die Erkennung einzelner Wörter in anderen Sprachen untersucht. Mehrere Autoren (siehe z.B. Heeringa et al. (2014), Möller & Zeevaert (2010), Marx (2012)) haben versucht herauszufinden, welche Aspekte die

richtige Identifikation von Wörtern ermöglichen. Diese Arbeit baut auf den bisherigen Recherchen von Robert Möller auf und verwendet Datensätze, die von diesem erstellt wurden.

Diese Arbeit versucht herauszufinden, welche Faktoren einen Einfluss auf die Erkennung von einzelnen Wörtern, die aus anderen Sprachen stammen, haben. Es wird geprüft, wann deutschsprachige Personen es leichter oder schwerer haben, Wörter anderer germanischer Sprachen zu erkennen. Die Wörter, die erkannt werden sollen sind Kognaten, also Wörter gleichen Ursprungs, die in verschiedenen Sprachen existieren. Zuerst wird jedoch auf die verschiedenen Konzepte eingegangen, die eine große Rolle in dieser Forschungsdisziplin spielen. Danach wird kurz auf den sprachgeschichtlichen Hintergrund des Deutschen und der anderen germanischen Sprachen im Allgemeinen eingegangen. Dies dient dazu, die Verwandtschaftsbeziehungen zwischen den germanischen Sprachen zu erläutern und auf die etwas besondere Position des Deutschen innerhalb dieser Sprachfamilie einzugehen. Dann wird auf die bisherige Forschung eingegangen, die sich mit diesem Thema beschäftigt hat. Dort wird dann auch auf die Erkennung ganzer Texte bei der rezeptiven Mehrsprachigkeit eingegangen, um zu zeigen, wodurch sich diese von der Einzelworterkennung unterscheidet und welche die Vorteile der beiden Methoden sind. Darauf folgt die Erklärung der verschiedenen Faktoren, die analysiert werden sollen und die Analyse selbst. In dieser werden zum einen die systematischen Faktoren untersucht, also jene, die einen stetigen Einfluss auf die Erkennung der Kognaten haben und zum anderen werden die Faktoren untersucht, die nur in manchen Situationen die Kognaten-Erkennung beeinflussen

## **2. Theoretische Hintergründe**

### **2.1 Begriffserklärungen**

Es existieren verschiedene Definitionen für das Prinzip der *Interkomprehension*. So erklärt Keipert (2009:234-235) es folgendermaßen:

„Mit dem Terminus *Interkomprehension* wird [...] das seit langem bekannte Phänomen bezeichnet, daß Sprecher verschiedener Sprachen sich wechselseitig verstehen können, wenn wie im Falle einer genetischen Verwandtschaft der beteiligten Idiome ähnliche morphosyntaktische Strukturen der Äußerungen, vor allem aber in größerer Zahl formal und semantisch wenigstens teilweise übereinstimmende Lexeme in den verwendeten Wortschätzen dem jeweiligen Hörer dabei helfen, sich in Kenntnis der Gesprächssituation und/oder des besprochenen Sachverhalts das jeweils vom Sprecher Gemeinte zumindest annähernd zu erschließen.“

Interkomprehension ist also rezeptive Mehrsprachigkeit zwischen Sprechenden von Sprachen, die Teil einer Sprachfamilie sind (vgl. auch Möller 2011). Die in dieser Arbeit untersuchte Interkomprehension basiert auf der Erkennung von Kognaten. Als Kognaten werden Wörter in verschiedenen Sprachen bezeichnet, die auf denselben Ursprung, d.h. auf ein gemeinsames Wort, zurückgehen. In dieser Arbeit wird aus Verständnisgründen als *Kognat\_f* immer das Wort bezeichnet, das sich in der Sprache befindet, von der die Testpersonen keine Kenntnisse haben bzw. das aus einer LX<sup>1</sup> der Testpersonen stammt. Das muttersprachliche Pendant dieses Wortes wird hier als *Kognat\_dt* bezeichnet. So wäre zum Beispiel für eine deutschsprachige Person das Wort *book* der *Kognat\_f* und *Buch* der *Kognat\_dt*. Anhand dieses Beispiels wird bereits ersichtlich, dass Kognaten nicht unbedingt in ihrer Orthographie und Phonologie übereinstimmen müssen. Dies ist auf den Ursprung der Kognaten zurückzuführen, die auf mehrere verschiedene Arten entstehen können. So können zum Beispiel zwei Sprachen dasselbe Wort aus einer dritten entlehnen. Heutzutage trifft dies vor allem auf Wortschatz aus dem Englischen zu. Es kann jedoch auch vorkommen, dass eine der beiden untersuchten Sprachen ein Wort aus der anderen entlehnt hat. Bei diesen beiden Vorgängen ist der ausdrucksseitige Unterschied zwischen den beiden Kognaten vor allem davon abhängig, wie weit die Entlehnung in der Vergangenheit zurückliegt. Bei rezenten Entlehnungen sind die Unterschiede für gewöhnlich noch recht gering. Im Laufe der Zeit können die beiden Kognaten sich jedoch sehr verändern.

Die letzte Art und Weise, auf die Kognaten entstehen können, ist über einen gemeinsamen Vorfahren der beiden Wörter, der aus einer gemeinsamen sprachlichen Vorstufe stammt. So stammen etwa dt. *Hund*, nl. *hond*, engl. *hound* und schwed. *hund* von einem germanischen Wort ab (*Hund*, DWDS). Die Unterschiede zwischen den Kognaten können jedoch weitaus größer sein als in diesem Fall. So sind zum Beispiel auch die Wörter dt. *Keks* und engl. *cakes* Kognaten. Das englische Wort wurde im 19. Jhd ins Deutsche entlehnt und hat sich sowohl ausdrucksseitig als auch semantisch verändert (*Keks*, DWDS). Während ein *cake* mit Kuchen übersetzt werden kann, ist ein *Keks* für gewöhnlich ein „trockenes Kleingebäck“ (*Keks*, DWDS). Die Ähnlichkeit zwischen Kognaten spielt natürlich eine große Rolle dabei, den einen Kognaten anhand des anderen zu finden. Natürlich ist die Ähnlichkeit kein binärer Wert und manche Kognatenpaare sehen sich ähnlicher als andere. Es existiert ein Messprinzip, das dazu dient, den Ähnlichkeitsgrad eines Kognatenpaares zu bestimmen: die Levenshtein-Distanz.

---

<sup>1</sup> Als LX bezeichnet man eine nach der L1 gelernte Sprache (siehe Dewaele 2017). In manchen Interkomprehensions-Tests wurden auch Sprachen abgefragt, die die Testpersonen zumindest ein wenig beherrschen oder sogar in der Schule gelernt haben (vgl. Gooskens & Swarte 2017).



Die Levenshtein-Distanz (LD) gibt an, wie ähnlich sich zwei Buchstabenfolgen (in diesem Fall Wörter) sind. Dazu wird berechnet, wie viele Bearbeitungsschritte (deshalb wird sie auch „Editier-Distanz“ genannt) mindestens benötigt werden, um ein Wort in das andere zu verwandeln. Dabei wird bei der klassischen Levenshtein-Distanz von drei möglichen Operationen ausgegangen: der Einfügung, der Tilgung und dem Austausch von Buchstaben. Dies soll an dieser Stelle anhand eines Beispiels erläutert werden. Die Levenshtein-Distanz zwischen den Wörtern *Pferd* und *Herde* beträgt 3, da man eine Einfügung, eine Löschung und einen Austausch vornehmen muss (siehe **Tab.1**).

**Tabelle 1:** Bearbeitungsschritte, um von dem Wort *Pferd* zu dem Wort *Herde* zu gelangen

1	2	3	4	5	6
<b>p</b>	<b>f</b>	<b>e</b>	<b>r</b>	<b>d</b>	
	<b>h</b>	<b>e</b>	<b>r</b>	<b>d</b>	<b>e</b>
1 Tilgung	1 Austausch				1 Einfügung

Je kleiner die Levenshtein-Distanz ist, desto ähnlicher sehen sich also die beiden Wörter, da man nicht so viele Bearbeitungsschritte durchführen muss. Auf den Vergleich zwischen deutschen Substantiven und jenen der anderen Sprachen muss dabei besonders Acht gegeben werden, da der Großbuchstabe die Levenshtein-Distanzen systematisch für diese Wortart erhöht. Großbuchstaben werden deshalb in diesem Fall bei der Berechnung der orthographischen Levenshtein-Distanzen nicht berücksichtigt (siehe auch Heeringa et al. 2014:378). Die Großschreibung muss nicht berücksichtigt werden, weil bei der Abfragung der Kognaten\_f immer jeweils angegeben wurde, ob es sich bei dem betreffenden Wort um ein Substantiv, ein Verb oder ein Adjektiv/Adverb handelt. So wussten die Testpersonen, ob sie einen Großbuchstaben am Wortanfang erwarten mussten oder nicht. Neben der klassischen Levenshtein-Distanz gibt es noch andere Varianten, die die Basismethode ein wenig verändern oder erweitern, um bestimmte Faktoren miteinzubeziehen. Auf diese Varianten wird genauer eingegangen, wenn die in dieser Arbeit untersuchten Faktoren der Kognaten-Erkennung vorgestellt werden.

Die Definition von Interkomprehension wird häufig noch weiter eingeschränkt und bezieht sich dann nur auf geschriebene Sprache (vgl. Möller & Zeevaert 2010). In Keiperts Definition liegt der Fokus auf der gesprochenen Sprache (siehe die Verwendung des Wortes „Hörer“ in Keiperts Definition). Dabei ist die Differenzierung zwischen geschriebener und mündlicher Kommunikation für intersprachliche Verständigung jedoch von großer Bedeutung. Bei

geschriebener Sprache ist die Verständigung zwischen zwei Sprachen einfacher, wenn beide über ein „konservatives orthographisches System“ (Möller & Zeevaert 2015:314) verfügen. Das bedeutet, dass sich die geschriebene Form der gesprochenen Sprache im Laufe der Zeit nicht zu stark verändert hat. Sprachwandel setzt sich im Regelfall zuerst in der gesprochenen Sprache durch und die geschriebene Sprache folgt. So kann sich die Lautform von Wörtern ändern, während der schriftliche Ausdruck zunächst unverändert bleibt. Dadurch ist der gemeinsame Ursprung von Wörtern in verschiedenen Sprachen leichter zu erkennen. Möller & Zeevaert (2015) weisen zudem darauf hin, dass das Erkennen von Kognaten in der geschriebenen Sprache im Gegensatz zur gesprochenen einfacher ist, da man mehrmals auf dasselbe Wort zurückkommen kann und Zeit hat, sich Gedanken darüber zu machen, welches Wort der L1 wohl der Kognat ist (vgl. auch Gooskens & Swarte 2017:142).

Wenn Sprechende einer germanischen Sprache A eine andere, ihnen unbekannte germanische Sprache B verstehen, dann ist es durchaus möglich, dass Sprechende dieser anderen Sprache B die Sprache A auch verstehen. Keipert (2007) bezeichnet dies in seiner Definition der Interkomprehension als Wechselseitigkeit. Es kann jedoch durchaus vorkommen, dass die Interkomprehension der einen Person bedeutend leichter fällt als der anderen. Ein Beispiel dafür ist die Kommunikation zwischen Dänisch- und Norwegischsprachigen. Dänischsprachige Menschen haben weniger Probleme (gesprochenes) Schwedisch zu verstehen als schwedischsprachige gesprochenes Dänisch (siehe Gooskens & Swarte 2017). Es lässt sich also eine eindeutige Asymmetrie in der Interkomprehension zwischen diesen beiden Sprachen feststellen.

Manche Sprachen sind also eine bessere Grundlage für die Interkomprehension als andere. Dies ist auf diverse Sprachwandelerscheinungen zurückzuführen. Durch Lautwandel können sich zwei vorher sehr ähnliche Sprachen im Laufe der Zeit erheblich unterscheiden. Dies führt dazu, dass gewisse Sprachen sich ähnlicher sehen als andere. Dabei sind Sprachen, die näher miteinander verwandt sind, sich für gewöhnlich ähnlicher. Wenn die Entstehung der Sprachen aus einer gemeinsamen Vorstufe weiter zurück liegt, dann hatten die Sprachen natürlich mehr Zeit, sich unterschiedlich zu entwickeln. Diese Entwicklung führte zu unterschiedlicher Phonologie, die schließlich in der Schriftsprache kodiert wurde. So sind sich Norwegisch und Schwedisch, die beide auf das Altnordische zurückgehen, ähnlicher als Norwegisch und Deutsch. Norwegisch und Deutsch, die beide aus dem Germanischen entstanden sind, sind sich wiederum ähnlicher als Norwegisch und Französisch etc. Die Entstehung von Sprachen geht teilweise mit großen systematischen Änderungen einher. So trennt die 1. Lautverschiebung die

germanischen Sprachen von den anderen indogermanischen Sprachen. Diese systematischen Änderungen gewisser Laute machen die Interkomprehension zwischen den germanischen und den anderen indoeuropäischen Sprachen schwieriger. Da das Hochdeutsche auch eine Lautverschiebung erfahren hat, ist es also wichtig, kurz auf die Stellung des Deutschen innerhalb der germanischen Sprachen einzugehen.

## 2.2 Die deutsche Sprache und die germanische Sprachfamilie

Innerhalb der germanischen Sprachen unterscheidet man zwischen drei Gruppen, von denen zwei noch heute existieren<sup>2</sup>: die westgermanischen Sprachen und die nordgermanischen Sprachen. Zu den westgermanischen Sprachen gehören z.B. Englisch, Deutsch, Niederländisch und Luxemburgisch. Sprachen wie z.B. Schwedisch, Norwegisch, Dänisch und Isländisch bilden die Gruppe der nordgermanischen Sprachen. Diese Sprachen weisen sowohl Kognaten auf, die auf Ursprungswörter in einer gemeinsamen sprachlichen Vorstufe zurückgehen, als auch solche, die auf Entlehnungen zurückgehen. Der Wortschatz, der auf die Verwandtschaft dieser Sprachen zurückzuführen ist, ist in der Regel älter, da die gemeinsamen sprachlichen Vorstufen dieser Sprachen bereits weiter in der Vergangenheit zurückliegen. Diese Kognaten können z.B. aus dem Indoeuropäischen (*Mutter* von ie. \**māter*-, DWDS) oder auch aus dem Germanischen (*Stein* von germ. \**staina*-, DWDS) stammen. Dabei ist der Ähnlichkeitsgrad zwischen den Kognaten aufgrund von Sprachwandelerscheinungen seit der Trennung der einzelnen Sprachen heutzutage unterschiedlich stark ausgeprägt:

- (1) dt. *Stein*
- (2) schwed. *sten*
- (3) nl. *steen*
- (4) norw. *stein*
- (5) dän. *sten*
- (6) isl. *steinn*
- (7) lux. *Steen*

Bei der Entstehung von Kognaten durch Entlehnung in der germanischen Sprachfamilie spielt das Niederdeutsche eine entscheidende Rolle. Durch die Ausbreitung und den zunehmenden

---

<sup>2</sup> Der dritte Zweig beinhaltet die ostgermanischen Sprachen, von denen heute jedoch keine mehr existiert. Die letzte ostgermanische Sprache, Krimgotisch, ist im 18. Jahrhundert ausgestorben (siehe Elspaß & Möller 2015).

Einfluss der Hanse gegen Ende des 13.Jahrhunderts breitete sich das Mittelniederdeutsche im Nord- und Ostseeraum aus. Dies lag daran, dass das Mittelniederdeutsche die Verwaltungssprache der Hanse war und auch die Niederlassungen in Hansestädten, die sich nicht im niederdeutschen Gebiet befanden (z.B. Stockholm), verwendeten diese Sprache (Stedje 2007:136). Dadurch wurde ein sehr großer Einfluss auf die dort gesprochenen Sprachen ausgeübt:

„Manche Nordisten sind der Ansicht, dass fast die Hälfte des gesamten schwed. Wortschatzes in der einen oder anderen Hinsicht niederdeutscher Herkunft sei (abgesehen von dem großen anglo-amerikanischen Einfluss der letzten Jahrzehnte).“

(Stedje 2007:136)

Dies führte also zur Entstehung vieler ähnlicher Kognaten in diesen Sprachen. Obwohl das Niederdeutsche nur einen sehr geringen Effekt auf die hochdeutsche Sprache hatte (siehe Stedje 2007:137), haben diese Sprachen doch sehr viele gemeinsame Kognaten. Dadurch gibt es natürlich auch viele Kognaten in jenen Sprachen, die durch das Mittelniederdeutsche beeinflusst wurden, die es auch im Hochdeutschen gibt.

Das Beispiel der englischen Sprache, die auch eine Zeit intensiver Entlehnungen erfahren hat, zeigt, dass sich Verwandtschaftsbeziehungen nicht unbedingt mit Kognatenähnlichkeit decken. Schepens et al. (2012) haben Kognaten der germanischen und romanischen Sprachfamilien auf Basis der orthographischen Ähnlichkeit miteinander verglichen und festgestellt, dass das Englische in dieser Hinsicht sehr hohe Ähnlichkeiten zum Französischen, Spanischen und Italienischen aufweist (161). Sie sagen dazu:

„This similarity of English to Romance languages derives to considerable extent from the consequences of the famous Norman-French victory at Hastings in 1066 by William the Conqueror. By losing this battle, Harold became the last English-speaking king for nearly 300 years, and in the following centuries many Romance words were introduced into English.“

Schepens et al. (2012:161)

Das Englische teilt die meisten der Kognaten mit dem Französischen und nicht mit irgendeiner anderen germanischen Sprache (Schepens et al. 2012:163). Dies passt sehr gut mit Schätzungen zusammen, dass „[m]ehr als die Hälfte der heutigen englischen Wörter [...] lateinisch-romanischen Ursprungs [sind]“ (Mader 2008:26).

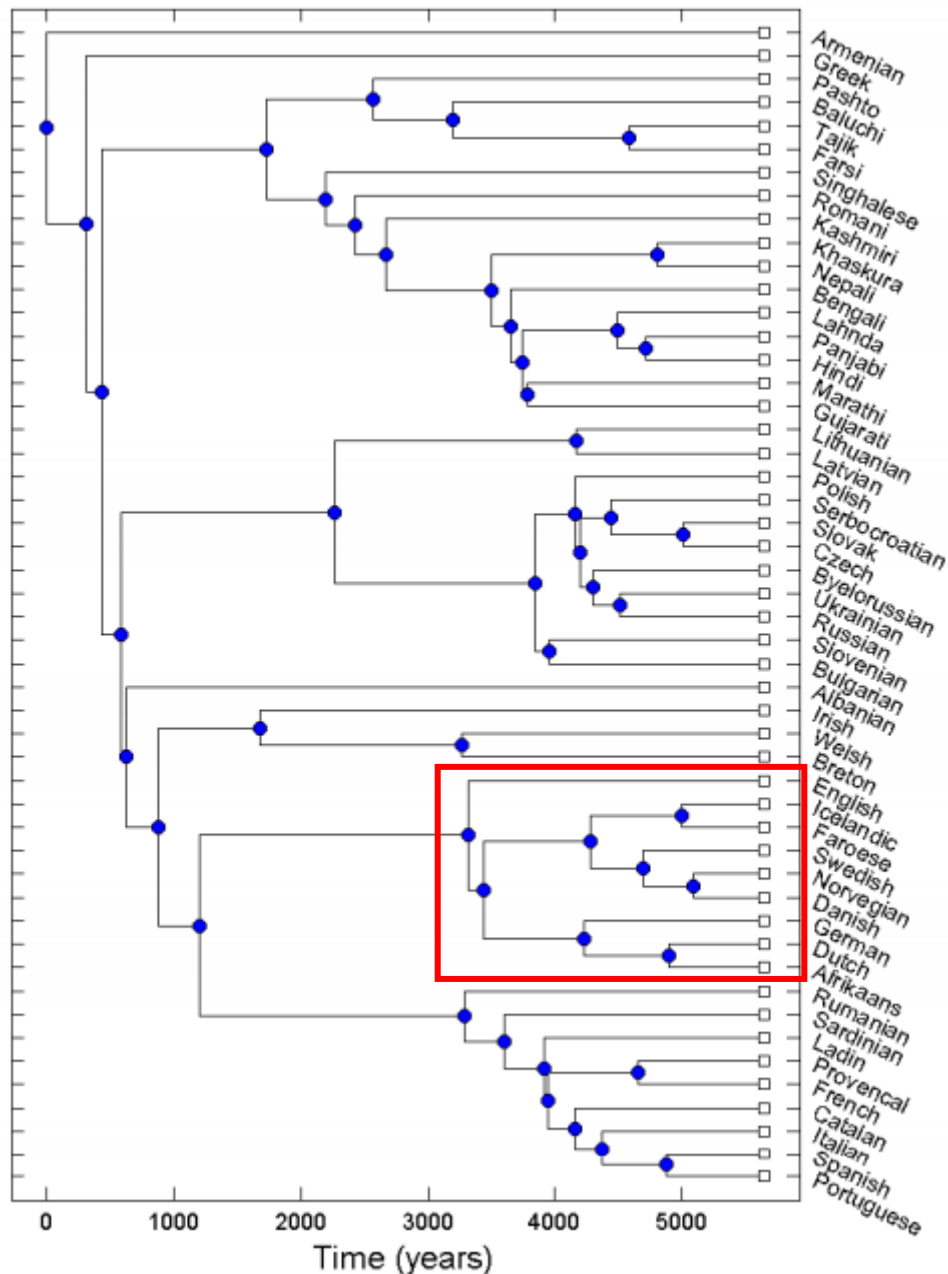
In Serva & Petroni (2008) wird dies sehr deutlich gezeigt: Sie haben in ihrer Studie die Ähnlichkeit der Kognaten mehrerer indoeuropäischer Sprachen untersucht. Dazu wurden die orthographischen Levenshtein-Distanzen der Kognatenpaare miteinander verglichen (siehe

Serva & Petroni 2008:1). Wenn die Verwandtschaftsbeziehungen von Sprachen indikativ für die Ähnlichkeit ihrer Kognaten sind, so ist zu erwarten, dass die Einteilung der germanischen Sprachen in nordgermanische und westgermanische in den Ergebnissen dieser Studie sichtbar ist. Dies ist auch für den Großteil, der in Serva & Petroni (2008) untersuchten germanischen Sprachen der Fall (siehe **Abb.1**). Man erkennt, dass Färöisch, Isländisch, Schwedisch, Norwegisch und Dänisch über einen Punkt verbunden sind. Die weiteren Einteilungen innerhalb dieser Gruppe stellen ebenfalls Verwandtschaftsbeziehungen dar. So bilden Schwedisch, Norwegisch und Dänisch die skandinavischen Sprachen und Norwegisch und Dänisch haben noch ähnlichere Kognaten, weil eine der norwegischen Schreibsprachen aus dem Dänischen entstanden ist. Deutsch und Niederländisch, beide westgermanische Sprachen, sind ebenfalls unter einem Knotenpunkt vereint (mit Afrikaans, das ja direkt vom Niederländischen abstammt). Die x-Achse (Time(years)) gibt die theoretisch errechnete Zeitspanne an, die zwischen dem Start der Trennung zweier Sprachen und der Gegenwart verstrichen ist. Diese Werte stimmen natürlich nicht zwangsläufig mit den richtigen Zeitabständen überein und sollen eher den Ähnlichkeitsgrad zwischen den Kognaten der verschiedenen Sprachen repräsentieren.

Der Unterschied zwischen den errechneten und den tatsächlichen Zeitspannen wird insbesondere für das Englische ersichtlich. Das Englische hat sich laut Abbildung 1 schon vor ungefähr 2700 Jahren von den anderen germanischen Sprachen getrennt. Dies ist natürlich falsch, da das Altenglische sich erst ab dem 7.Jh. n.Chr. entwickelt hat (siehe Möller & Elspaß 2015:55). Diese sehr großen errechneten Zeitspannen verdeutlichen jedoch, wie stark die englischen Kognaten aufgrund des französischen Einflusses von jenen der anderen germanischen Sprachen abweichen. Der Unterschied zwischen den englischen Kognaten und den Kognaten der westgermanischen Sprachen ist dabei genauso groß, wie jener zwischen den englischen und den nordgermanischen Kognaten. Die Annahme, dass die Kognatenähnlichkeiten den Verwandtschaftsgrad zwischen Sprachen widerspiegeln, wird dadurch jedoch nicht zwangsläufig widerlegt. Es wird nur offensichtlich, dass keine perfekte Entsprechung zwischen diesen beiden Größen besteht und die allgemeine Kognatenähnlichkeit zwischen zwei Sprachen durch Entlehnung verändert werden kann. Diese aus anderen germanischen oder nicht-germanischen Sprachen übernommenen Wörter waren aber seit ihrer Entlehnung (natürlich abhängig davon, wie lange diese her ist) Sprachwandelphänomenen unterworfen, wie es bereits bei den durch Verwandtschaftsbeziehungen entstandenen Kognaten der Fall war. Dadurch haben sich im Laufe der Jahrhunderte auch wieder mehr oder weniger

große Unterschiede zwischen diesen Kognaten gebildet und die Kognatenschaft ist deshalb heute in unterschiedlichem Maße sichtbar.

**Abbildung 1:** Klassifizierung indogermanischer Sprachen auf der Basis der durchschnittlichen Levenshtein-Distanzen zwischen Kognaten, Time (years) gibt an, wie viel Zeit verstrichen ist, seit sich zwei Äste des Dendrogramms geteilt haben (theoretisch berechneter Wert) (Serva & Petroni 2008:3, roter Rahmen zur Kennzeichnung der relevanten Daten hinzugefügt)



Wenn es um Sprachwandel und germanische Sprachen geht, dann rücken zwei Prozesse natürlich in den Vordergrund: die erste und die zweite Lautverschiebung. Die erste Lautverschiebung (1.LV) war ein systematischer Lautwandel, der die germanischen Sprachen

von den übrigen indoeuropäischen Sprachen trennte. Diese ist für die größeren Unterschiede zwischen den englischen (aus dem französischen entlehnten) Kognaten und den Kognaten der übrigen germanischen Sprachen verantwortlich.

Die zweite oder auch hochdeutsche Lautverschiebung (2.LV) trennte die hochdeutsche Sprache von den anderen germanischen Sprachen. Dies ist deshalb natürlich besonders relevant für die germanische Interkomprehension, wenn die L1 der Testpersonen Hochdeutsch ist. Die 2.LV ist auch die Basis für die Aufteilung in Hoch- und Niederdeutsch. Hochdeutsch ist die Sprache, die ursprünglich im Süden, also in den geographisch höher gelegenen Regionen des heutigen Deutschlands gesprochen wurde. Niederdeutsch hingegen wurde und wird noch heute eher im Norden gesprochen. Aufgrund dieser geographischen Verteilung der beiden Sprachen ist die niederdeutsche Sprache auch dem Niederländischen ähnlicher als das Hochdeutsche.

Die einzelnen Lautwandel, die unter dem Begriff 2. *Lautverschiebung* zusammengefasst werden sind zu zahlreich, um hier aufgezählt zu werden und dies ist auch nicht notwendig. Wichtig ist vor allem, dass diese systematische Veränderung von Lauten und deren Kodifizierung in der geschriebenen Sprache im Laufe der Jahrhunderte dazu geführt hat, dass die Ähnlichkeit zwischen hochdeutschen Kognaten und solchen aus anderen germanischen Sprachen sehr stark variieren kann.

Auf die durch die 2. Lautverschiebung zurückgehenden Unterschiede als mögliches Problem wurde auch bereits im Rahmen der Interkomprehensions-Forschung hingewiesen:

„Between Germanic cognates, differences can be rather wide (e.g., Dutch *doop* – German *Taufe*), and as a consequence of the Old High German sound shift especially a great many German words differ systematically and to a considerable degree from their cognates in the other Germanic languages.”

(Möller & Zeevaert 2015)

Anhand des Beispiels *doop* kann man die Nähe zwischen den germanischen Sprachen erkennen, die die 2. Lautverschiebung nicht mitgemacht haben (9 – 13):

(8) dt. *Taufe*

(9) nd. *dööp*

(10) nl. *doop*

(11) schwed. *dop*

(12) norw. *dåp*

(13) dän. *dåb*

Trotz dieser besonderen Situation des Deutschen innerhalb der germanischen Sprachfamilie ist es trotzdem interessant, die Interkomprehension anhand von Kognaten aus der Perspektive von Testpersonen mit Deutsch als L1 zu untersuchen, da Deutsch die germanische Sprache mit den meisten Muttersprachlern ist. Die bisher durchgeführte Forschung zum Thema Interkomprehension hat sich mit verschiedenen Sprachen und Sprachfamilien beschäftigt. Die Studien, auf die nun in dieser Arbeit eingegangen wird, beschränken sich jedoch zum Großteil auf die germanischen Sprachen.

### **3. Bisherige Interkomprehensions-Forschung**

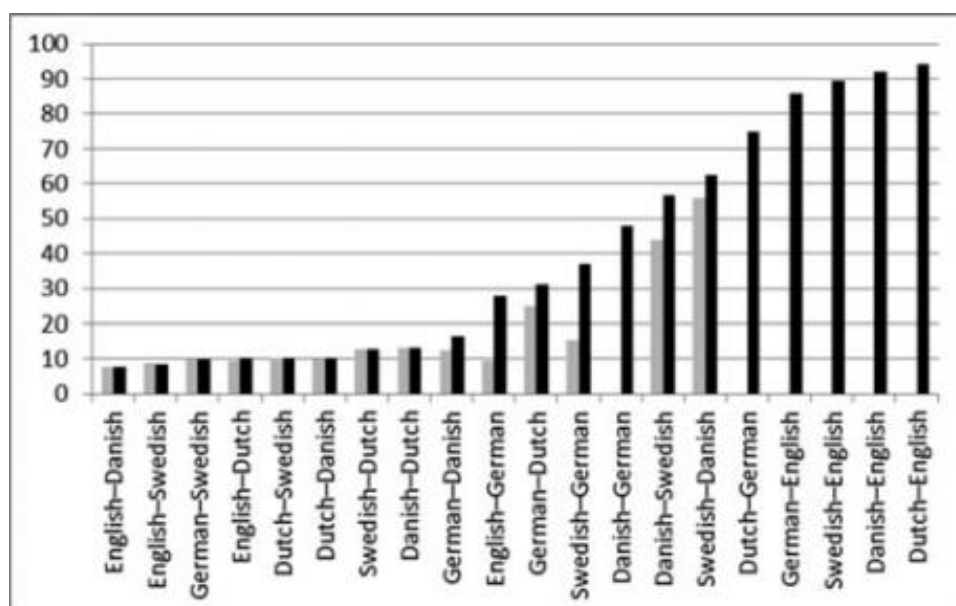
Für sehr lange Zeit lag (und teilweise ist dies auch heute noch der Fall) der Fokus der Forschung zum Thema Interkomprehension auf der didaktischen Komponente und „[t]heoretische Überlegungen [standen] deshalb eher im Hintergrund“ (Möller & Zeevaert 2010:218). Die Frage nach den linguistischen und nicht-linguistischen Faktoren, die die Interkomprehension ermöglichen oder erschweren, ist jedoch im Laufe der Jahre immer häufiger untersucht worden. Die zugrundegelegten Definitionen von Interkomprehension stimmen nicht in allen diesen Arbeiten überein. Manchmal wurde in Interkomprehensions-Studien nicht zwischen Personen mit Fremdsprachen-Kenntnissen der anderen untersuchten Sprachen und solchen ohne diese Kenntnisse unterschieden. Dies ist zum Beispiel der Fall in Gooskens & Swarte (2017). Dort wurden außerdem sowohl geschriebene als auch gesprochene Sprache untersucht, während in anderen Tests, wie z.B. denen von Möller & Zeevaert (2010), nur schriftliche Interkomprehension untersucht wurde. Da diese Arbeit nur auf schriftlich erhobenen Datensätzen aufbaut, wird hier auch nur die schriftliche Interkomprehension untersucht. Die hiernach folgende Zusammenfassung der Forschung zur germanischen Interkomprehension beschränkt sich demnach auch (fast) ausschließlich auf die Resultate, die auf der Erkennung geschriebener fremdsprachlicher Kognaten basieren. Dabei kann man zwischen zwei Arten von Studien unterscheiden. Zum einen gibt es diejenigen, die sich exklusiv mit der Erkennung von Kognaten beschäftigen, wie z.B. Möller & Zeevaert (2010) sowie die vorliegende Arbeit. Zum anderen gibt es diejenigen, die sich mit dem Erschließen von Texten in verwandten Sprachen beschäftigen. Hierbei spielt der Kontext eine wichtige Rolle bei der Erschließung. Wörter, die vielleicht normalerweise nicht erkannt werden könnten, werden anhand des Kontextes richtig identifiziert. Eine solche Studie wurde von Gooskens & Swarte (2017) durchgeführt. Auf diese wird an dieser Stelle genauer eingegangen, da sie zeigt, wie unterschiedlich die



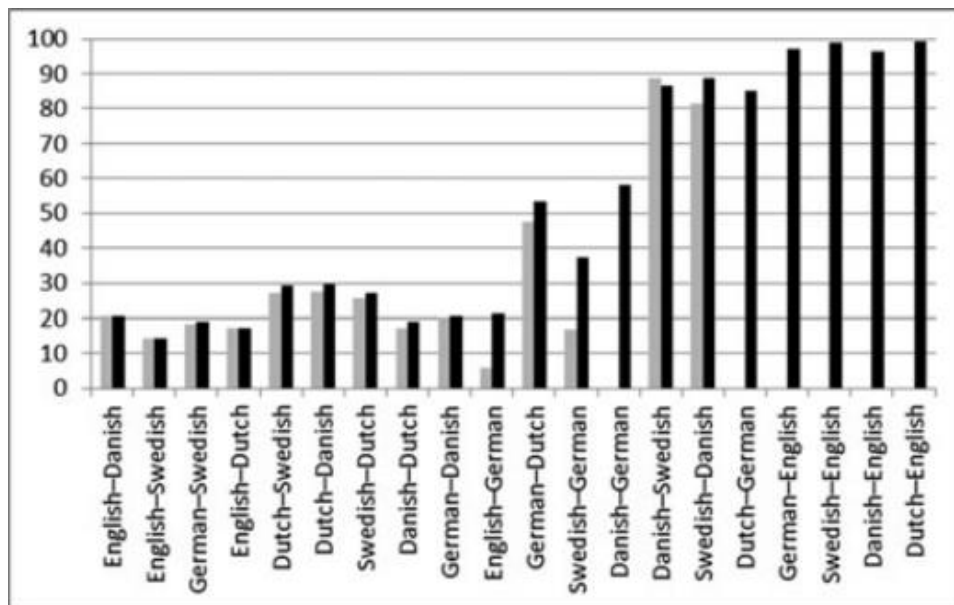
Untersuchungsmethoden für die Texterschließung im Vergleich zur Kognaten-Erkennung bei einzelnen Wörtern sind.

Gooskens & Swarte (2017) haben die Interkomprehension zwischen fünf germanischen Sprachen (Deutsch, Englisch, Schwedisch, Dänisch und Niederländisch) untersucht. Dazu wurden Sprechende der fünf Sprachen gebeten, einen Lückentext in den jeweils anderen Sprachen auszufüllen. Es wurde nicht darauf geachtet, dass diese Wörter Kognaten sind, da es hier nicht um Kognaten-Erkennung ging. In dieser Studie wurden auch Testpersonen akzeptiert, die Kenntnisse von einer der Sprachen (mit Ausnahme ihrer L1) hatten. Man kann jedoch nur dann von Interkomprehension (wie es oben definiert wurde) sprechen, wenn Personen versuchen, eine Sprache zu verstehen, die sie nicht beherrschen. M.E. kann man in diesem Fall jedoch bedingt von Interkomprehension sprechen, da einige der Testpersonen, die an dieser Studie teilgenommen haben, nur minimalen Kontakt zu einigen der anderen Sprachen haben (d.h. (fast) nie Kontakt zu den untersuchten Sprachen haben). Bei der Auswertung in Gooskens & Swarte (2017) wurden die Ergebnisse dieser Personen nicht nur mit den anderen zusammen, sondern auch separat behandelt. Da sich die vorliegende Arbeit nur mit wahrer Interkomprehension beschäftigt, werden an dieser Stelle vor allem die Ergebnisse der Personen, die keine häufigen Kontakte (bzw. keine Kenntnisse) der jeweiligen anderen Sprachen haben, besprochen. Gooskens & Swarte (2017) haben diesen Test sowohl in mündlicher (**Abb.2**) als auch in schriftlicher (**Abb.3**) Form durchgeführt.

**Abbildung 2:** Prozentuale Ergebnisse des mündlichen Lückentext-Tests von Gooskens & Swarte (2017). Bei jedem der Sprachenpaare steht die L1 der Testpersonen an erster Stelle und die Testsprache an zweiter. Die hellgrauen Balken stellen die Ergebnisse der Personen dar, die keine Kenntnisse der jeweiligen Testsprachen haben (Gooskens & Swarte 2017:130).



**Abbildung 3:** Prozentuale Ergebnisse des geschriebenen Lückentext-Tests von Gooskens & Swarte (2017). Bei jedem der Sprachenpaare steht die L1 der Testpersonen an erster Stelle und die Testsprache an zweiter. Die hellgrauen Balken stellen die Ergebnisse der Personen dar, die keine Kenntnisse der jeweiligen Testsprachen haben (Gooskens & Swarte 2017:130).



Ein Vergleich der Ergebnisse der mündlichen und schriftlichen Tests lässt erkennen, dass die Probanden und Probandinnen viel weniger Schwierigkeiten dabei haben, die Testsprachen zu verstehen, wenn sie einen Text selber lesen anstatt die Sprache zu hören. Gooskens & Swarte (2017) erklären, wie bereits oben besprochen, dass es beim Lesen wohl einfacher ist, auf das Gesagte zurückzukommen:

„It is easier to grasp the concrete written form of the language than the volatile spoken form. In addition, the written form of a closely related language is likely to be easier to match with the corresponding form in the native language because written language reflects an earlier form of the language where the two languages had diverged less from the original common form than in the spoken form.”

(Gooskens & Swarte 2017:142)

Aufgrund der konservativeren Natur der geschriebenen Sprache ähneln sich die schriftlichen Ausdrücke der beiden Sprachen der jeweiligen Sprachenkombinationen mehr als die Lautformen.

Anhand der Ergebnisse wird ersichtlich, dass es sechs Sprachenkombinationen gibt, in denen alle Testpersonen mehr oder weniger gute Kenntnisse der Testsprache hatten (Sprachenkombinationen ohne graue Balken in **Abb.2** und **Abb.3**). Vier von diesen sind jene Kombinationen, in denen Englisch die Testsprache war. Dies geht natürlich auf die Tatsache zurück, dass Englisch die am häufigsten in der EU gelernte Schulsprache ist: 2017 lernten

94,7% der SchülerInnen Englisch in der Sekundarstufe II (Eurostat). So ist es natürlich unwahrscheinlich, Testpersonen zu finden, die nur minimale Englischkenntnisse haben. Dies ist insbesondere in der Studie von Gooskens & Swarte (2017) der Fall, da nur Universitäts-Studierende für die Studie ausgewählt wurden und diese häufig regelmäßigen Kontakt zum Englischen haben. Die anderen beiden Sprachenkombinationen, für die es keine Testpersonen mit nur minimalen Kenntnissen der Testsprache gab, waren Niederländisch – Deutsch und Dänisch – Deutsch. Das bedeutet, dass alle dänisch- und niederländischsprachigen Testpersonen ein gewisses Maß an Kontakt zur deutschen Sprache hatten. Die Testpersonen mussten angeben, ob sie regelmäßig mit Medien (Fernsehen, DVDs, etc.) in einer der anderen Sprachen in Kontakt kommen, oder nicht (Gooskens & Swarte 2017:129). Aufgrund der Nähe von Dänemark und den Niederlanden zu Deutschland ist es wohl wahrscheinlich, dass Personen aus diesen Ländern z.B. deutsches Fernsehen empfangen. Außerdem ist Deutsch ein Pflichtfach in den Niederlanden (Gooskens et al. 2015:256). Um diesen Test mit Niederländisch- oder Dänischsprachigen durchzuführen, die keine Kenntnisse des Deutschen haben, müssten relativ junge Kinder, die noch keinen Deutschunterricht hatten, als Testpersonen ausgewählt werden. Dies wurde in Gooskens et al. (2015) für Niederlandischsprachige getan und auf diese Studie wird weiter unten genauer eingegangen.

Wenn man die Sprachenkombinationen mit Deutsch für den geschriebenen Test (**Abb.3**) heranzieht und nur die Ergebnisse der Testpersonen mit minimalen Kenntnissen der Testsprachen (graue Balken) auswählt, dann sieht man, dass Deutsch als L1 der Testpersonen die besten Ergebnisse liefert (**Tab.2**).

**Tabelle 2:** Prozentualer Anteil richtiger Antworten der Personen, die nur minimale Kenntnisse für die Testsprache haben, für die Sprachenkombinationen, die Deutsch enthalten (graue Balken in **Abb.3**; Gooskens & Swarte 2017).

L1	Testsprache	Richtige Antworten
Deutsch	Niederländisch	~47%
Deutsch	Dänisch	~20%
Deutsch	Schwedisch	~18%
Schwedisch	Deutsch	~16%
Englisch	Deutsch	~5%

Anhand dieses direkten Vergleiches sieht man, dass Deutschsprachige bei der Interkomprehension am häufigsten den niederländischen Text richtig ausfüllen. Dies liegt vermutlich daran, dass beide westgermanische Sprachen sind. Die nächste Sprache, Dänisch, wird nicht einmal halb so häufig richtig ausgefüllt. Dies deckt sich mit der Feststellung von

Gooskens & Swarte (2017), dass die niedrigsten gemittelten Resultate für Kombinationen bestehen, in denen eine Sprache zu den westgermanischen und die andere zu den nordgermanischen Sprachen gehört. Deutsch ist jedoch besonders für Englischsprachige sehr schwierig zu verstehen, wenn diese keine Vorkenntnisse haben. Dies geht vermutlich wieder auf die großen Wortschatzunterschiede zwischen dem Englischen und den anderen germanischen Sprachen zurück. Dies wird dadurch bestätigt, dass das Englische in den drei Sprachenkombinationen, die am schlechtesten bei dem schriftlichen Test abgeschnitten haben (bei Personen ohne sprachliche Kenntnisse), immer die L1 der Testpersonen war. Die Interkomprehension zwischen Schwedisch und Deutsch hingegen ist mehr oder weniger symmetrisch.

Gooskens & Swarte (2017) haben anhand dieser Ergebnisse untersucht, welche Rolle verschiedene Faktoren bei der Texterschließung spielen. Zum einen wurden linguistische Distanzen, wie z.B. die phonologische oder syntaktische Distanzen, untersucht. Phonologische Distanzen werden mithilfe der Levenshtein-Distanz berechnet. Es werden nur anstelle von Graphemen Phoneme verwendet. Auf dieses Distanzprinzip wird weiter unten genauer eingegangen, wenn die Faktoren, die in der vorliegenden Arbeit untersucht werden, vorgestellt werden. Die syntaktische Distanz ist spezifisch für Texterschließungsaufgaben, da sie die Position von Wörtern innerhalb der zu erschließenden Sätze untersucht. Dazu werden die Sätze in der Sprache der Testperson mit jenen in der Testsprache verglichen. Dabei wird zum einen gemessen, um wie viele Stellen ein gegebenes Wort im Satz verrückt werden muss, wenn man den Satz in die andere Sprache übersetzt. Dies wird *movement measure* genannt. Zum anderen wird gemessen, ob und wie viele Wörter bei der Übersetzung hinzukommen oder wegfallen. Dies ist das Äquivalent der Insertion (bei der Levenshtein-Distanz), aber auf der Satz- anstelle der Wortebene und wird *indel measure* genannt. Außerdem wird noch eine Trigramm-Berechnung durchgeführt. Ein Trigramm ist eine Einheit aus drei Elementen (Buchstaben, Wörter, etc., hier: Wörter) und dient dazu, Nachbarschaftsbeziehungen zu untersuchen. Dies soll anhand des Beispiels (14) erklärt werden. a. bis d. stellen die drei Trigramme dar, aus denen der Satz „Ich habe großen Hunger“ besteht. Das \$ steht für den Satzanfang und das # das Satzende (siehe auch Gooskens & Swarte 2017:135).

(14) Ich habe großen Hunger.

- |    |    |          |      |          |              |
|----|----|----------|------|----------|--------------|
| a. | \$ | Pronomen | Verb |          |              |
| b. |    | Pronomen | Verb | Adjektiv |              |
| c. |    |          | Verb | Adjektiv | Substantiv   |
| d. |    |          |      | Adjektiv | Substantiv # |

Abhängig davon, wie häufig ein gewisses Trigramm in einer Sprache vorkommt, ist es einfacher für eine Person diese Struktur in einem Satz zu erkennen, der in einer der Testperson unbekannten Sprache geschrieben ist. Das Trigramm /Satzbeginn/-/Pronomen/-/Verb/ ist im Deutschen sehr oft vertreten und wird so vermutlich im Deutschen einfacher erkannt als in einer SOV-Sprache. Gooskens & Swarte (2017) haben herausgefunden, dass in den Fällen, in denen echte Interkomprehension vorliegt, d.h. in denen die Testpersonen nur wenig oder überhaupt keinen Kontakt zur Testsprache hatten, diese syntaktischen Distanzen mit der Texterschließung korrelieren (insbesondere die Trigramm-Frequenzen)(Gooskens & Swarte 2017:140). Dies bedeutet, dass diese Faktoren in der Tat eine Rolle bei der Texterschließung spielen und, dass die Interkomprehension durch Texterschließung von Faktoren abhängig ist, die in der vorliegenden Arbeit nicht auftreten, da der Fokus hier auf der Einzelworterkennung liegt.

Wie zuvor bereits kurz erwähnt wurde, haben Gooskens et al. (2015) die Interkomprehension zwischen deutsch- und niederländischsprachigen Kindern untersucht, die keine Kenntnisse der anderen Sprache hatten und relativ weit von der Deutsch-Niederländischen-Grenze leben (Gooskens et al. 2017:259). Beide Gruppen mussten versuchen, Kognaten aus der jeweils anderen Sprache in die eigene zu übersetzen. Diese Kognaten wurden den Testpersonen vorgelesen und war so ein Test, der auf der mündlichen Worterkennung beruhte. Dabei haben Gooskens et. al. (2015) herausgefunden, dass die deutschsprachigen Kinder mehr Probleme dabei hatten, die niederländischsprachigen zu verstehen, als umgekehrt. Gooskens et al. (2015) haben anhand der phonologischen Distanz, also anhand der Ähnlichkeit der Aussprache der beiden Kognaten versucht, diese Unterschiede in der Worterkennung zu erklären. Aber die Resultate zeigten, dass die phonologischen Distanzen nicht ausreichen, um die Worterkennungsrate der Testpersonen vollständig zu erklären (Gooskens et al. 2015:278). Gooskens et al. (2015) gehen davon aus, dass bei jedem der einzelnen Testitems zusätzliche besondere Bedingungen vorliegen, die diese verschiedenen Wörter entweder besonders einfach oder besonders schwierig für die Sprechenden der jeweils anderen Sprache erkennbar machen. So kann es sein, dass ein Kognat\_f, dessen Kognat\_dt mehrere ähnlich aussehende Nachbarwörter hat, öfters falsch erkannt wird, da eines dieser Nachbarwörter anstelle des

Kognaten\_dt ausgesucht wird (Gooskens et al. 2015:277). Das zeigt, wie komplex bereits die Erkennung von einzelnen Kognaten in anderen Sprachen sein kann.

In Möller & Zeevaert (2010), (2015) und Möller (2011) standen die einzelnen Kognaten und deren Erkennung im Mittelpunkt. Es wurde überprüft, welche Intuitionen deutschsprachige Personen für Lautkorrespondenzen haben, die zwischen dem Deutschen (L1 der Testpersonen) und den anderen germanischen Sprachen existieren. Dabei wurde nicht zwischen den wirklichen Kognaten\_dt und den anderen Antworten der Testpersonen unterschieden. Es wurde nur gemessen, wie plausibel die einzelnen Antworten waren. Als „plausibel“ wurden dabei jene Antworten angesehen, die nach geläufigen Lautkorrespondenzmustern anhand des Kognaten\_f gebildet werden können. Die Kognatenpaare wurden für diesen Test so ausgewählt, dass die Kognaten\_f „jeweils zu verschiedenen deutschen Substantiven dieselbe ungewichtete Levenshtein-Distanz (Editier-Distanz) aufweisen“ (Möller und Zeevaert 2010:222, die Liste der verwendeten Wörter ist in Möller (2011) zu finden). So konnte man dann überprüfen, welche Wörter, die als Kognat\_dt in Frage kommen, ausgewählt wurden und warum.

Es wurden zwei Arten von Tests durchgeführt: eine freie Abfrage und ein Multiple-Choice-Test (siehe Möller 2011). Bei der freien Abfrage mussten die teilnehmenden Personen (75 Studierende der Universität Bonn und der Universität Münster, siehe Möller 2011) für eine Liste von Kognaten\_f die Kognaten\_dt angeben. Der Multiple-Choice-Test (mit anderen Testpersonen) verwendete die vorher ausgesuchten Wörter und die Antworten der freien Abgabe als mögliche Kognaten\_dt für die verschiedenen Kognaten\_f (siehe Möller 2011). In manchen Abfragen wurden die richtigen Antworten nicht in die Liste der möglichen Items aufgenommen. Interessanterweise wurden in diesen Fällen von den Testpersonen häufiger Lösungen ausgewählt, die aufgrund ihrer fehlenden Lautkorrespondenzen bei Möller (2011) als „unwahrscheinlich“ eingestuft wurden (siehe Möller 2011:93). Demnach sind bei nicht-offensichtlichen Lösungen auch diejenigen akzeptabel, die ansonsten nicht in Betracht gezogen werden. Möller (2011) vergleicht diese Ergebnisse mit Fällen, in denen der Kontext die Wahl der Kognat\_dt beeinflusst:

„[D]ie Wahrscheinlichkeit eines Kognatenverhältnisses [wird] weniger kritisch beurteilt [...], wenn die Suche dadurch eingeschränkt wird, dass der semantische Kontext eine bestimmte Lösung nahezu legen scheint.“

(Möller 2011:94)

Da die richtige (bzw. beste, da die Testpersonen ja nicht wissen, welcher der richtige Kognat\_dt ist) Option nicht angeboten wurde, wurde jedoch nicht unbedingt die zweitbeste ausgewählt.

Die Testpersonen sind also relativ flexibel bei ihrer Wahl des vermuteten Kognaten\_dt und diese Wahl kann durch eine Reihe weiterer Faktoren, die nicht auf der phonologischen oder graphemischen Ähnlichkeit basieren, beeinflusst werden. Die vollständigen Ergebnisse werden an dieser Stelle nicht angegeben, aber viele der am häufigsten richtig erkannten Kognaten\_dt wiesen ähnliche Lautkorrespondenzen auf, die von den Testpersonen präferiert wurden, wie z.B. die Anfügung eines Schwas am Wortende (Möller 2011:90). Diese Lautkorrespondenzen wurden schließlich benutzt, um eine phonologische Levenshtein-Distanz zu entwickeln, die die Intuitionen der Testpersonen in Betracht zieht und so den plausibleren Lautkorrespondenzen niedrigere phonologische Distanzen verleiht. Auf diese phonologische Levenshtein-Distanz wird weiter unten genauer eingegangen, wenn die Faktoren vorgestellt werden, die als Basis für die Analyse in dieser Arbeit dienen. Zuvor wird jedoch noch auf extralinguistische Faktoren genauer eingegangen, die bei allen Arten von Interkomprehension eine Rolle spielen können.

#### **4. Extralinguistische Faktoren der Kognaten-Erkennung**

Neben den linguistischen Faktoren, die bereits besprochen wurden, gibt es andere, die nicht von den Eigenschaften der untersuchten Sprachen („langue“) an sich abhängig sind. Diese nicht-linguistischen Faktoren können aber durchaus etwas mit „langage“, also mit der Sprachfähigkeit an sich, zu tun haben. Diese extralinguistischen Faktoren sind eher personenspezifische Charakteristika, die die Erkennung von Kognaten mitbeeinflussen. So werden diese in Vanhoven & Berthele (2017) auch „participant-related variables“ (38) genannt.

Gooskens & Swarte (2017) haben neben dem zuvor bereits besprochenen Faktor der LX als Schulsprache zwei weitere nicht-linguistische Faktoren untersucht: Kontakt zur untersuchten Sprache und die Einstellung der Testpersonen gegenüber dieser Sprache. Interessanterweise scheint die Einstellung der Testpersonen gegenüber der untersuchten Sprachen keine Rolle für deren Verständlichkeit zu spielen. Dies wird durch die Vorgehensweise in dieser Studie erklärt: „It is possible that this is caused by the experimental settings where participants may be less strongly influenced by their attitudes towards the language and its speakers than in real life“ (Gooskens & Swarte 2017:142). Durch die künstlichen Interkomprehensionsbedingungen während des Tests ist die Validität dieser Resultate also fraglich. Die Testpersonen interagieren nicht mit dieser ihnen fremden Sprache in einer Alltagssituation und sie wissen, dass man von ihnen erwartet, dass sie sich zumindest ein wenig anstrengen, um die Texte zu verstehen. Inwiefern dieser Faktor sich also wirklich auf die Interkomprehension auswirkt, ist schwer zu

sagen. Um dies herauszufinden, müsste man also die Interaktionen zwischen Personen in realen Situationen beobachten.

Die Intensität des Kontaktes zur untersuchten Sprache ist nach Gooskens & Swarte (2017) entscheidend für das Verständnis. Wenn eine Testpersonen nie (oder nur selten) Kontakt zu einer gewissen Sprache hat, dann ist diese Sprache laut Gooskens & Swarte (2017) nur über die linguistische Distanz zwischen den Sprachen zu erschließen. Meines Erachtens wird dabei jedoch die Verbindung zwischen der untersuchten Sprache und einer Art „Vermittlersprache“ übersehen. Eine LX, die diese Testperson bereits (zumindest teilweise) beherrscht, könnte Ähnlichkeiten zur untersuchten Sprache aufweisen, die das Erschließen der letzteren ermöglicht. Als Beispiel könnte man sich eine deutschsprachige Person vorstellen, die an der Grenze zu Dänemark lebt und so zumindest Basiskenntnisse des Dänischen hat. Dadurch ist es möglich, dass diese Person auch norwegische Texte teilweise verstehen könnte. Dies sind jedoch Grenzfälle und der Kontakt zur Zielsprache ist vermutlich insbesondere für die mündliche Interkomprehension wichtig. Man kann wohl davon ausgehen, dass Menschen, die häufig eine gewisse Sprache hören, die sie nicht beherrschen, sich an deren phonologisches Profil, d.h. deren Lautinventar und wiederkehrende Lautkombinationen gewöhnen.

Vanhove & Berthele (2017) haben zwei weitere nicht-linguistische Faktoren untersucht. Zum einen ging es um die Spracherfahrung und zum anderen um die fluide Intelligenz der einzelnen Testteilnehmenden. Die Spracherfahrung wurde anhand eines Wortschatztests bestimmt, bei dem die teilnehmenden Personen zwischen existierenden deutschen Wörtern und Nicht-Wörtern unterscheiden mussten (siehe Vanhove & Berthele 2017). Die Spracherfahrung wurde also anhand des Wortschatzreichtums der einzelnen Personen bestimmt. Als fluide Intelligenz bezeichnet man die Fähigkeit sich an bestimmte Situationen flexibel anzupassen. Vanhove & Berthele (2017) haben herausgefunden, dass sowohl höhere Niveaus fluider Intelligenz als auch bessere Spracherfahrung die Erkennung von Kognaten erleichtern. Ein größerer Wortschatz ist bei der Erkennung von gewissen Kognaten förderlich, deren Kognaten\_dt im Deutschen vielleicht nicht sehr gebräuchlich sind. Man kann ein gegebenes Kognat natürlich nur dann bestimmen, wenn man mit dessen L1-Pendant zumindest ein wenig vertraut ist. Es ist aber natürlich wieder eine andere Frage, ob man dann auch an dieses Wort denkt, wenn man den Kognaten liest bzw. hört.

Die fluide Intelligenz hilft dabei, die Verbindung zwischen den Kognaten\_f und deren Kognaten\_dt zu knüpfen:



„Participants with high fluid intelligence levels are arguably more adept at coping with obscured cognate relationships as they can treat the stimuli’s actual forms more flexibly by abstracting away from formal discrepancies and instead establishing similarities.”

(Vanhove & Berthele 2017:37)

Die Testteilnehmenden mit hoher fluider Intelligenz können also leichter feststellen, welche Teile eines Kognaten vermutlich nicht so im Kognaten\_dt vorhanden sind und so abwägen, wie der Kognat\_f wohl im Deutschen aussehen würde.

Die Ergebnisse dieser Studie zeigen jedoch auch, dass diese erleichterte Erkennung von Kognaten\_f bei höheren Niveaus fluider Intelligenz bei niedrigen linguistischen Distanzen zwischen den Wörtern eine negative Korrelation aufzeigt (siehe Vanhove & Berthele 2017:37). Das bedeutet, dass bei Wörtern, die sich ähnlicher sehen, häufiger die falsche Antwort gegeben wird, als bei Wörtern, die sich weniger ähnlich sehen. In der Studie wird davon ausgegangen, dass dies auf eine Überkompensation durch die fluide Intelligenz zurückzuführen ist:

“When cognate relationships are obscured, this [die Flexibilität bei höherer fluider Intelligenz] is usually advantageous, but when they are not, this flexibility may backfire. [...] When decoding a stimulus like h rd ‘hard’, not much flexibility is required; in fact, decoders who are too flexible can sometimes come up with incorrect translations such as Herz ‘heart’ or Herd ‘stove’.

(Vanhove & Berthele 2017:37)

Die Teilnehmenden überschätzen also in diesem Fall die Schwierigkeit der Aufgabe und suchen so nach weniger geeigneten Lösungen, weil sie davon ausgehen, dass die Antwort ansonsten wohl zu einfach zu finden wäre. Auch wenn dieser negative Effekt nur relativ klein ist (siehe Vanhove und Berthele 2017:37), so muss man doch davon ausgehen, dass aufgrund dieser Überkompensation vermeintlich leicht zu erkennende Kognaten\_f falsch identifiziert werden.

Die hier besprochenen Ergebnisse beziehen sich nur auf die geschriebenen Tests und die mündliche Kognaten-Erkennung zeigt teilweise widersprüchliche Ergebnisse. Die mündlichen Tests haben gezeigt, dass ein größerer Wortschatz negative Auswirkungen auf die Erkennung von Kognaten hat. Das Niveau fluider Intelligenz scheint hingegen überhaupt keine klare Wirkung auf die Interkomprehension zu haben.

Da die für diese Arbeit erhobenen Resultate aber nur die einzelnen Testwörter in Betracht ziehen, d.h. „item-related“ (siehe Vanhove & Berthele 2017) sind, werden die nicht-linguistischen Faktoren hier nicht in Betracht gezogen. Es steht jedoch fest, dass die für diese Arbeit untersuchten Faktoren linguistischer Natur nicht die einzigen Einflüsse auf die Kognaten-Erkennung darstellen. Diese Faktoren sollten jedoch keine größeren Auswirkungen auf die Ergebnisse des analysierten Tests haben. Für die einzelnen Kognatenpaare sollten die

Resultate der Testpersonen, die es besonders leicht oder besonders schwierig haben, sich gegenseitig aufheben. Wenn nun alle Teilnehmenden besondere Schwierigkeiten haben oder überdurchschnittlich (im Vergleich zum Durchschnitt der Gesamtheit aller Deutschsprachigen) gut abschneiden, dann kann man davon ausgehen, dass die Testpersonengruppe an sich nicht repräsentativ ist. Auf die Testteilnehmenden wird weiter unten genauer eingegangen.

Obwohl Faktoren wie fluide Intelligenz und Wortschatz von einer Testperson zur nächsten variieren, gibt es dennoch einen Faktor, den alle Testteilnehmenden gemeinsam haben: die Buchstaben- und Worterkennung im mentalen Lexikon. Im nächsten Kapitel wird erklärt, wie dieses Erkennungssystem funktioniert und wie dies sich auf die Erkennung von Kognaten\_f auswirken könnte.

## **5. Das *Interactive Activation Model* (IAM)**

Um zu begreifen, warum eine Testperson ein Wort anstelle eines anderen als möglichen Kognaten\_dt für einen Kognaten\_f auswählt, ist es wichtig zu verstehen, wie diese Wörter im mentalen Lexikon organisiert sind und wie diese Wörter abgerufen werden. Das *Interactive Activation Model* oder IAM wurde in McClelland & Rumelhart (1981) präsentiert. Es war eines der ersten und einflussreichsten Modelle, das versucht hat, die kognitiven Vorgänge bei der Wortverarbeitung zu erklären und es ist auch heute noch aktuell (vgl. z.B. McClelland et al. (2014)). In diesem Modell wird davon ausgegangen, dass für jedes Wort im mentalen Lexikon ein Wortknoten (eine sogenannte *node*) vorhanden ist (McClelland & Rumelhart 1981). Diese *node* enthält nicht nur die graphische sondern auch die phonologische und semantische Seite dieses Wortes und dieser wird aktiviert, wenn die betreffende Person das dazugehörige Wort liest oder hört (siehe McClelland & Rumelhart 1981:379). Diese Aktivierung findet schneller statt, wenn das betreffende Wort häufig vorkommt, da die nötige Stimulation des Wortknotens geringer ist (siehe McClelland & Rumelhart 1981:379). Da der Wortknoten häufig verwendet wird, ist er sensibler für die Erkennung der Phonologie, Graphie oder die Semantik des Wortes (McClelland & Rumelhart 1981:379).

Diese Aktivierung kommt nur unter bestimmten Bedingungen vor. Ein Mindestmaß an Charakteristika des betreffenden Wortes muss präsent sein, damit dessen Wortknoten aktiviert wird. Dabei werden zuerst die Buchstaben verarbeitet und darauf folgt dann das gesamte Wort (siehe McClelland & Rumelhart 1981). Hier stellt sich dann die Frage nach dem Mindestmaß

an Ähnlichkeit, das notwendig ist, um die Aktivierung des Wortknotens noch zu gewährleisten. Dasselbe gilt natürlich für die Erkennung gesprochener Wörter. Dass dieses System relativ flexibel ist, kann man daran erkennen, dass die Aussprache von Wörtern (viel mehr als dies bei geschriebener Sprache der Fall ist), sehr personenspezifisch ist. Ob es nun dialektale Unterschiede (/ɪç/ vs. /ɪʃ/) oder Eigenheiten der Aussprache sind, die auf den Sprecher unabhängig von seiner Sprachgemeinschaft zurückgehen. Trotz all diesen Variationen gibt es (meistens) keine größeren Probleme der Verständigung. Im Prinzip lassen sich fast alle linguistischen Faktoren für die Erkennung von Kognaten auf dieses Modell zurückführen: All jene Faktoren, die die Aktivierung des Wortknotens der deutschen Kognat\_dt ermöglichen bzw. die Anzahl der in Frage kommenden Wortknoten verringern, sollten also eine korrekte Erkennung des Kognaten begünstigen.

Das IAM kann ohne weiteres auf die in den bisherigen Studien beschriebenen Faktoren angewendet werden. Die in Vanhove & Berthele (2017) beschriebene Korrelation zwischen Spracherfahrung und erfolgreicher Kognaten-Erkennung ist z.B. nur möglich, wenn ein betreffender Knoten für das gesuchte Wort im mentalen Lexikon besteht. Ansonsten findet keine Aktivierung und somit keine Erkennung des Kognaten statt. Das IA-Modell ist von besonderem Interesse für die Kognaten-Erkennung durch Testpersonen, die passive Kenntnisse niederdeutscher Dialekte haben. Bei Dialektsprechern ist es leicht nachzuvollziehen, warum gewisse Kognaten anderer germanischer Sprachen leicht erkannt werden, da sie regelmäßig mit vergleichbaren Wörtern in Kontakt kommen und diese selbst verwenden. Bei passiver Kenntnis ist dies aber nicht der Fall. Hier kann es sogar vorkommen, dass sich die Testpersonen nicht bewusst sind, dass sie über solche Kenntnisse verfügen. Wenn sie nun ein gewisses Kognat sehen bzw. hören, dann ist es möglich, dass ihr mentales Lexikon einen Wortknoten aktiviert, der vielleicht einmal angelegt, aber noch nie oder schon lange nicht mehr aktiviert wurde. Inwiefern solche passiven Kenntnisse jedoch eine konkrete Rolle spielen, ist nur sehr schwierig festzustellen. Während in den Fällen, in denen sich die Teilnehmenden der Verbindung zwischen den Kognaten und deutschen Dialekten bewusst sind, eine Verbindung leicht nachzuvollziehen ist (siehe die Lautprotokolle in Möller & Zeevaert 2010), wird dies bei unterbewussten Kenntnissen schwieriger. Dies könnte die Basis einer Art „Sprachintuition“ bilden, deren Ursprung nicht nachzuvollziehen ist.

Die Levenshtein-Distanz ist ebenfalls mit dem IA-Modell kompatibel. Da die Aktivierung von Wortknoten auf die Aktivierung von Buchstabenknoten folgt (siehe McClelland & Rumelhart 1981), ist es natürlich wichtig, zu wissen, wie groß die Distanz zwischen dem Kognaten\_f und

dem gesuchten Kognaten\_dt ist. Je größer diese Distanz ist, desto weniger stimmen die Buchstabenfolgen der beiden Wörter überein. Dies führt dann dazu, dass der falsche Kognat\_dt auf der Basis der Buchstaben, die nicht mit denen im Kognaten\_dt übereinstimmen, im mentalen Lexikon aktiviert wird. Man könnte sich ein Wort wie *rande* (konstruiert) vorstellen, dessen Kognat\_dt zum Beispiel *rannte* ist. Die Levenshtein-Distanz beträgt also 2 (Tilgung eines <n> und Austausch des <d> durch ein <t>). Es gibt aber mehrere Wörter, die diesem Kognaten ähnlicher sehen, wie z.B. *Runde*, *Rinde* oder *Rand* (jeweils Levenshtein-Distanz 1). Wenn nun eines dieser Wörter aktiviert wird, dann haben das <d> und das zusätzliche <n> das mentale Lexikon dazu gebracht, den falschen Wortknoten zu aktivieren. Diese Aktivierung fand dann aufgrund der Annahme statt, dass das Wort im mentalen Lexikon mit den meisten Buchstabenübereinstimmungen der richtige Kognat\_dt ist.

## 6. Kontext und Kotext

Richtige Interkomprehension findet natürlich nur selten so statt, wie es in den oben beschriebenen Studien untersucht wird. Die Texterkennungsstudien wie jene von Gooskens & Swarte (2017) sind dabei durchaus repräsentativer für wirkliche Alltagsinterkomprehension als die Kognaten-Erkennungstests, da die Testpersonen Kotext zur Verfügung haben, der es ihnen erlaubt, ihnen unbekannte Wörter trotzdem zu erschließen. Dies ist natürlich ein wichtiger Faktor bei der realen Interkomprehension. Es kann jedoch vorkommen, dass der Kotext, der ein gewisses Wort umgibt, nicht ausreicht, um dieses zweifelsfrei zu erkennen. Hier kommt dann der Kontext ins Spiel. Wie bei der Kognaten-Erkennung in diesen Studien kann es auch im alltäglichen Leben vorkommen, dass ein Wort alleine steht. Ein Beispiel dafür sind Straßenschilder. Eine deutschsprachige Person hat möglicherweise Probleme damit, das Wort *flygplats* auf einem Schild zu verstehen. Wenn dieses Schild nun aber in der Nähe eines Flughafens steht oder sogar ein Piktogramm eines Flugzeugs darauf abgebildet ist, dann ist es wohl wahrscheinlich, dass die betreffende Person versteht, dass es sich bei einem *flygplats* um einen Flughafen handelt. All diese Hinweise auf die Bedeutung dieses Wortes sind bei der Erkennung isolierter Kognaten nicht gegeben. Dadurch ist es aber auch möglich, die Faktoren zu bestimmen, die die Lautform und den schriftlichen Ausdruck leichter erkennbar machen, ohne dass der Ko- und Kontext die Ergebnisse beeinflussen. Da der Ko- und Kontext alleine keine Interkomprehension möglich machen, sondern diese maßgeblich von der Ähnlichkeit der

Wörter der Sprachen im Allgemeinen und deren geteilten Kognaten insbesondere abhängt, ist es wichtig, die Kognaten-Erkennung isoliert zu untersuchen.

## **7. Ziel des Worterkennungstests**

Die oben besprochenen Studien, die sich mit rezeptiver Mehrsprachigkeit beschäftigen, haben für die linguistischen Faktoren immer auf die Ähnlichkeit zwischen den untersuchten Sprachen aufgebaut. Dabei lag der Fokus entweder auf der Ähnlichkeit der Sprachen im Allgemeinen (dabei wurden dann Texterkennungstests verwendet) oder auf der Ähnlichkeit spezifischer Wörter unterschiedlicher Sprachen. Wie bereits erwähnt, werden in dieser Studie die Kognaten verschiedener germanischer Sprachen untersucht und es wird überprüft, welche Faktoren dafür sorgen, dass Kognaten besonders einfach oder schwer zu identifizieren sind. Im folgenden Teil werden nun diese Faktoren näher beschrieben, deren Einfluss auf die Erkennbarkeit von Kognaten überprüft werden soll.

## **8. Testaufbau**

Der Kognaten-Erkennungstest wurde schriftlich in Form einer freien Abfrage online durchgeführt ([www.Kognatentest.de](http://www.Kognatentest.de)) und die Testpersonen wurden gebeten, für einen gegebenen Kognaten\_f den gesuchten Kognaten\_dt anzugeben. Insgesamt wurden 312 Testitems verwendet, die Kognaten aus verschiedenen germanischen Sprachen waren (Dänisch, Niederländisch, Isländisch, Schwedisch, Norwegisch, Luxemburgisch). Dabei wurden manche Wörter verwendet, die in allen skandinavischen Sprachen identisch sind (mit *skand.* gekennzeichnet). 180 der Testitems sind Substantive, 88 Verben und 44 Adjektive/Adverbien.

Einige dieser „Kognaten“ wurden nach geläufigen Lautkorrespondenzmustern aus deutschen Wörtern konstruiert und andere wurden so verändert, dass sie etwas leichter für eine deutschsprachige Person zu erkennen sind. Darauf wurden die Testpersonen zu Beginn des Tests auch hingewiesen, damit diejenigen mit Vorkenntnissen sich nicht dadurch verwirren ließen. Die Testpersonen wurden ebenfalls darauf hingewiesen, dass die Buchstaben (S), (V) und (A), die mit jedem Kognaten\_f angezeigt wurden, die Wortart des Kognaten\_f bezeichneten (Subjekt, Verb oder Adjektiv/Adverb).

Die Testpersonen hatten eine unbegrenzte Anzahl von Versuchen zur Verfügung, um den richtigen Kognaten\_dt zu finden. Wenn sie diesen nicht finden konnten, dann hatten sie die Möglichkeit, den Kognaten\_f zu überspringen und zum nächsten überzugehen. Nach dem Absenden einer Antwort konnten die Testpersonen ebenfalls einen Kommentar abgeben, in dem sie zum Beispiel erklären konnten, woran sie den angezeigten Kognaten\_f erkannt haben. Sie wurden außerdem gebeten, den Kommentar „bekannt“ in das Kommentarfeld zu schreiben, wenn sie den Kognaten\_f bereits kannten.

Am Ende des Tests wurden die Testpersonen noch gebeten, einige Informationen anzugeben. Diese betrafen z.B. Dialektkenntnisse und die Frage, ob die Testpersonen bereits Kurse besucht haben, die sich mit Sprachgeschichte, Phonetik/Phonologie, etc. beschäftigt haben. Dieser Fragebogen ist im Anhang dieser Arbeit aufgeführt. 91,2% der Testpersonen waren Deutschsprachige und von den 249 Personen, die an diesem Test teilgenommen haben, waren 230 Schüler oder Studierende – davon 197 Germanistikstudierende. Auf die verschiedenen Kurse, die die Testpersonen besucht haben, und mögliche Dialektkenntnisse wird in dieser Arbeit nicht genauer eingegangen, da sich Möller & Zeevaert (2010) und Möller (2011, 2015) bereits mit diesen Aspekten auseinandergesetzt haben. An dieser Stelle soll nur darauf hingewiesen werden, dass sowohl vorherige Erfahrungen mit Sprachwandelerscheinungen im Rahmen von Universitätskursen als auch (von den Testpersonen selbst angegebene) Dialektkenntnisse keinen bemerkbaren Einfluss zu haben scheinen (vgl z.B. Möller 2011:87).

## **9. Auswahl potentieller Faktoren**

### **9.1 Die Levenshtein-Distanz und deren Varianten**

Die Levenshtein-Distanz wurde auch in dieser Arbeit in Betracht gezogen, da sie eine sehr einfache Methode darstellt, um die Ähnlichkeit zwischen zwei Wörtern zu bestimmen. In dieser Arbeit wird jedoch von verschiedenen Varianten der Levenshtein-Distanz Gebrauch gemacht. Es wurde zum Beispiel ebenfalls eine gewichtete Levenshtein-Distanz verwendet. Diese zeichnet sich dadurch aus, dass nicht alle Operationen die Distanz um 1 erhöhen. So können zum Beispiel Austausch-Operationen mit höheren „Kosten“ verbunden werden als Tilgungs-Operationen, wenn man diese als besonders unwahrscheinlich ansieht. Demnach würde dann die Levenshtein-Distanz zwischen den Wörtern *Hosen* und *Hasen* (Austausch des <o> durch ein <a>) höher sein als jene zwischen *Hosen* und *Hose* (Tilgung des <n> am Wortende). Eine

weitere Art, eine gewichtete Levenshtein-Distanz zu kreieren, ist, nicht die Art der Operation in Betracht zu ziehen, sondern die Natur der Elemente, die von diesen Operationen betroffen sind. So kann man die Tilgung eines <e> am Ende eines Wortes als weniger ausschlaggebend ansehen als die Tilgung eines <t>. Diese Methode ermöglicht eine differenziertere Berechnung der Ähnlichkeit zwischen zwei Wörtern.

Die gewichtete Levenshtein-Distanz, die in dieser Arbeit verwendet wird, basiert vornehmlich auf phonologischen Kriterien und wurde von Robert Möller konzipiert (vorgestellt auf dem *International Symposium on Bilingualism* 7, 2009). Diese phonologische Levenshtein-Distanz baut auf den Intuitionen von Testpersonen zu den Lautentsprechungen zwischen dem Deutschen und den anderen germanischen Sprachen auf. Das bedeutet, dass eine Operation mit einem niedrigeren Preis versehen ist, wenn diese eine Änderung herbeiführt, die mit den von den Testpersonen präferierten Lautentsprechungen übereinstimmt. Da diese Intuitionen durchaus mit den tatsächlichen Lautkorrespondenzen zwischen germanischen Sprachen übereinstimmen (Möller 2011:97), kann man sagen, dass die niedrigen gewichteten phonologischen Levenshtein-Distanzen vor allem bei plausiblen Antworten der Kognaten-Erkennung auftreten.

Dabei kann es vorkommen, dass phonologische Distanzen von 0 zwischen dem Kognaten\_dt und dem Kognaten\_f existieren. Dies bedeutet jedoch nicht zwangsläufig, dass diese Wörter eine identische Aussprache haben. So wird zum Beispiel der Austausch eines Konsonanten durch einen Vokal als unwahrscheinlicher angesehen, als der Austausch eines Konsonanten durch einen anderen Konsonanten und deshalb mit einem höheren Wert versehen. Eine vollständige Liste der für die verschiedenen Operationen und Buchstabenarten verwendeten Werte sind im Anhang zu finden. Da häufig im Deutschen keine direkten Graphem-Phonem-Korrespondenzen bestehen, werden sich die orthographischen und phonologischen Distanzen unterscheiden. Dies erlaubt es, zu erkennen, ob die Phonologie, unabhängig von der Orthographie, eine Rolle bei der Erkennung von Kognaten spielt und wie groß deren Einfluss auf die Resultate ist.

Neben der Verwendung der absoluten Levenshtein-Distanz kann man auch die relative (bzw. normalisierte) Levenshtein-Distanz berücksichtigen. Dabei wird die Länge der Wörter, deren Distanz berechnet wird, berücksichtigt und ein Vergleich der Levenshtein-Distanzen von Wortpaaren mit Wörtern unterschiedlicher Länge wird erleichtert. Die relativen Distanzen sind auch bessere Messwerte für die Ähnlichkeit zwischen einem Kognaten\_f und einem Kognaten\_dt. Wenn zwei Operationen an einem Wort durchgeführt werden, das nur fünf

Buchstaben enthält, dann sieht das resultierende Wort natürlich sehr verschieden aus. In einem Wort mit zehn Buchstaben sind die Unterschiede schon nicht mehr so deutlich erkennbar, da sich nur ein relativ kleiner Teil des Wortes verändert hat (siehe auch Serva & Petroni 2008).

Es gibt verschiedene Methoden, um diese normalisierte Levenshtein-Distanz zu berechnen. Serva & Petroni (2008) haben für ihre Berechnungen die Levenshtein-Distanz durch die Anzahl Buchstaben im längeren der beiden Wörter geteilt. In dieser Arbeit wurde die Distanz durch die gemittelte Länge der beiden Wörter geteilt, um die relative Levenshtein-Distanz zu ermitteln. Für das Beispiel aus **Tab.1**, bei dem die Levenshtein-Distanz 3 war, findet man so eine relative Levenshtein-Distanz von 0,6. Genauso wie bei der gewöhnlichen Editier-Distanz sind kleinere Werte indikativ für eine größere Ähnlichkeit zwischen den beiden Wörtern und die Werte der relativen Levenshtein-Distanz liegen immer zwischen 0 und 1.

Für die in dieser Arbeit verwendete gewichtete Levenshtein-Distanz wurden die relativen Distanzen etwas anders berechnet. Da diese Editier-Distanzen nicht auf der Orthographie, sondern auf der Phonologie basieren, werden die relativen Distanzen nicht anhand der Anzahl Buchstaben berechnet. Stattdessen werden die Levenshtein-Distanzen durch die gemittelte Anzahl der Phoneme geteilt. Dies kann man anhand des Wortpaares *Schirm* und *Stirn* demonstrieren. Wenn man von der Orthographie ausgeht, dann findet man eine Levenshtein-Distanz von 3 (**Tab.2**) und daraus folgt dann eine relative Levenshtein-Distanz von  $3/5,5 = 0,54$ .

**Tabelle 2:** Bearbeitungsschritte, um von dem Wort *Schirm* zu dem Wort *Stirn* zu gelangen

1	2	3	4	5	6
s	c	h	i	r	m
s		t	i	r	n
	1 Tilgung	1 Austausch			1 Austausch

Wenn man *Schirm* und *Stirn* jedoch in Lautschrift transkribiert (jeweils ['ʃrɐm] und ['ʃtɪrən]), dann erhält man andere Bearbeitungsschritte und eine andere Anzahl an Segmenten (**Tab.3**). Bei einer gemittelten Segmentanzahl von 4,5 und 2 Operationen ergibt sich so eine relative gewichtete Levenshtein-Distanz von 0,44.



Tabelle 3: Bearbeitungsschritte, um von der IPA-Transkription des Wortes *Schirm* zu der IPA-Transkription des Wortes *Stirn* zu gelangen

1	2	3	4	5
ʃ		ɪ	e	m
ʃ	t	ɪ	e	n
	1 Tilgung			1 Austausch

Eine erweiterte Variante der Levenshtein-Distanz ist die Damerau-Levenshtein-Distanz (DLD). Diese fügt den drei Basisoperationen eine vierte hinzu: die Transposition. Bei der Transposition tauschen zwei nebeneinander liegende Buchstaben innerhalb eines Wortes die Position. Dies verringert die Distanz zwischen gewissen Wörtern. Die Levenshtein-Distanz zwischen *rief* und *reif* beträgt 2 (z.B. Tilgung des <e> hinter <i> und Einfügung des <e> vor <i>). Die Damerau-Levenshtein-Distanz beträgt aber nur 1, da man einfach eine Transposition des <e> und des <i> durchführen kann.

Die Präsenz des für die Bildung der richtigen Kognaten\_dt notwendigen Buchstaben im Kognat ist sehr wahrscheinlich förderlich für dessen Erkennung. Studien haben außerdem festgestellt, dass bei gewissen Sprachen die Buchstabenposition weniger relevant ist als deren Identität (siehe Lerner et al. 2014). Dies betrifft jene Sprachen, in denen sich die Wörter vor allem durch die Art der Buchstaben, die in ihnen enthalten sind, unterscheiden. Lerner et al. (2014) haben dies für Englisch, Spanisch und Französisch untersucht. Man kann davon ausgehen, dass das Deutsche auch eher zu einer dieser Sprachen gehört. Die Alternative dazu sind Sprachen, in denen sich die Wörter vor allem durch die Position von Buchstaben unterscheiden, d.h. in denen viele Anagramme existieren, wie z.B. in den semitischen Sprachen wie Arabisch oder Hebräisch (siehe Lerner et al. 2014). Falls die orthographische Damerau-Levenshtein-Distanz ein guter Prädiktor für die Interkomprehension ist, dann ist dies sicherlich nicht auf alle Sprachfamilien übertragbar<sup>3</sup>. In diesen Fällen wäre es auch unmöglich zu entscheiden, wie ähnlich sich zwei Wörter sehen und wie sehr diese Ähnlichkeit die Wahl der richtigen Kognaten\_dt beeinflusst. Der nächste Faktor hilft dabei, diese Ähnlichkeiten zwischen orthographischen Nachbarn mithilfe der Levenshtein-Distanz zu ermitteln.

<sup>3</sup> Dies gilt natürlich auch für sämtliche Sprachen, die gewisse andere Schriftsysteme verwenden, wie z.B. eine Morphem-Schrift (z.B. Mandarin). Dies trifft jedoch nur auf die orthographischen Distanzen zu. So haben Tang & van Heuven (2015) die Levenshtein-Distanz auf phonetische Transkriptionen angewandt, um die Interkomprehension zwischen chinesischen Dialekten zu untersuchen.

## **9.2 Die *Orthographic Levenshtein distance 20* (OLD20)**

Die *orthographic Levenshtein distance 20* (OLD20) wurde von Balota & Yap (2008) entworfen, um Ähnlichkeitsverhältnisse zwischen einem Wort und seinen orthographischen Nachbarn zu quantifizieren und wurde ursprünglich für die visuelle Worterkennung entwickelt. Sie lässt die Ähnlichkeit zwischen dem Kognaten\_f und dem Kognaten\_dt außer Acht und dient dazu, zu bestimmen, wie ähnlich sich ein Wort und seine 20 nächsten Nachbarn sind. Dazu werden auf Basis der (in diesem Fall einfachen) Levenshtein-Distanz die 20 Wörter herausgesucht, die der deutschen Kognaten\_dt am ähnlichsten sehen, d.h. die kleinsten Levenshtein-Distanzen aufweisen. Diese werden gemittelt und so erhält man die OLD20. Wörter, die einen kleinen OLD20-Wert haben, besitzen also sehr viele Nachbarn, die ihnen sehr ähnlich sehen. Es ist davon auszugehen, dass der Kognat\_dt eines Kognaten\_f seltener richtig erkannt wird, wenn dieser Kognat\_dt einen niedrigen OLD20-Wert hat. Das liegt daran, dass eine höhere Wahrscheinlichkeit besteht, dass man anstelle des richtigen Kognaten\_dt einen der orthographischen Nachbarn auswählt. Demnach müsste also eine positive Korrelation zwischen der OLD20 und der Anzahl richtiger Antworten bestehen.

## **9.3 Die Länge der Kognaten**

Die Wortlänge sowohl der Kognaten\_f als auch der Kognaten\_dt spielen aus zwei Gründen eine Rolle. Zum einen beeinflussen sie natürlich direkt die relative Levenshtein-Distanz zwischen einem Kognaten\_f und dessen Kognaten\_dt, da die Editier-Distanz ja durch die gemittelte Wortlänge der beiden Wörter geteilt wird. Zum anderen kann man außerdem anhand des Verhältnisses zwischen der Länge des Kognaten\_dt und der Länge des Kognaten\_f erkennen, ob ein Kognat\_f einfacher zu erkennen ist, wenn dieser länger oder kürzer als der Kognat\_dt ist. Dies zeigt, ob die Testpersonen es einfacher haben, wenn ein Buchstabe dem Kognaten\_f hinzugefügt oder weggenommen werden muss, um den Kognaten\_dt zu bilden. Für einen Kognaten\_f mit fünf Buchstaben, dessen Kognat\_dt vier enthält, beträgt das Längenverhältnis 0,8. Ein Kognat\_f mit sechs Buchstaben und dessen Kognat\_dt mit fünf führen zu einem Längenverhältnis von 0,833. Je größer diese Zahl, desto größer ist der Kognat\_dt im Vergleich zum Kognaten\_f. Wenn dieser Wert 1 beträgt, dann haben der Kognat\_f und der Kognat\_dt dieselbe Länge.

## 9.4 Die Wortfrequenz

Die Frequenz, mit der die Kognaten\_dt eines Kognaten\_f in einer Sprache vorkommen, könnte sich ebenfalls auf die Interkomprehension auswirken. Wie schon bei den nicht-linguistischen Faktoren erklärt, sind die Sprachenkenntnisse der verschiedenen Testpersonen wichtig für die Erkennung bestimmter Wörter. Wenn die Kognaten\_dt der abgefragten Kognaten selten vorkommen, d.h. niedrig-frequent sind, dann ist die Wahrscheinlichkeit größer, dass diese den Testpersonen nicht bekannt sind. Dies ist jedoch sicherlich nur in äußerst seltenen Wörtern der Fall. Man kann wohl davon ausgehen, dass die für diese Arbeit analysierten Wörter den meisten Testpersonen bekannt sein sollten. Die Frequenzen spielen jedoch eine Rolle bei dem Versuch des mentalen Lexikons die Kognaten\_dt der Kognaten\_f zu finden. McClelland & Rumelhart (1981) erklären, dass die Aktivierung der verschiedenen Wortknoten nur dann stattfinden kann, wenn ein gewisser Grenzwert, der von Wort zu Wort unterschiedlich ist, überschritten wurde. Die Frequenz der Wörter ist dabei entscheidend, um zu bestimmen, wie einfach es ist, einen spezifischen Wortknoten zu aktivieren:

„This resting level may differ from node to node and [is] determined by frequency of activation of the node over the long term. Thus, for example, the nodes for high-frequency words have resting levels higher than those for low-frequency words.”

(McClelland & Rumelhart 1981:379)

Je höher die Frequenz eines Wortes ist, desto höher ist also der Ruhepegel des Wortknotens. Da höhere Ruhepegel näher am Grenzwert für die Aktivierung dieses Wortknotens sind, werden diese für höherfrequente Wörter schneller aktiviert. Selbst wenn Wortknoten für alle Kognaten\_dt der abgefragten Kognaten\_f bestehen, ist es möglich, dass diese nicht aktiviert werden. Trotz einer hohen Übereinstimmung zwischen den Buchstaben des Kognaten\_f und der Kognaten\_dt könnte möglicherweise das falsche Wort als Kognat\_dt ausgesucht werden, wenn dieses eine höhere Gebrauchsfrequenz aufweist. Die Auswahl des höherfrequenten Wortes kann sich auf zwei verschiedenen Ebenen abspielen. Die erste Ebene betrifft eine unterbewusste Entscheidung zu Gunsten des Wortes, mit dem die Testperson vertrauter ist, weil sie diesem bereits häufiger begegnet ist. Auf der zweiten Ebene entscheidet sich die Testperson bewusst dafür, die niedrigfrequenter Option zu verwerfen, da diese als ein unwahrscheinlicherer Kognat\_dt angesehen wird. Wenn auch nur eine dieser Vermutungen zutrifft, dann müsste eine positive Korrelation zwischen der Frequenz der Kognaten\_dt und der Anzahl richtiger Antworten bestehen.

## **9.5 Die Wortart**

Der nächste Faktor, der untersucht wird, ist die Wortart der Kognaten\_f. Wenn die Wortart keinerlei Effekt auf die Erkennbarkeit der Kognaten hat, dann müsste man davon ausgehen, dass für alle Wortarten im gleichen Maße richtige Antworten gefunden werden. Aufgrund der Tatsache, dass Verben immer die Endung *-en* haben, sind in allen Kognaten\_f, die Verben sind, immer bereits zwei Buchstaben bzw. Phoneme bekannt. Sofern die Verben also nicht länger sind als die Kognaten, die zu den anderen Wortarten gehören, müssten diese am einfachsten zu erkennen sein. Deshalb wird nicht nur die Erkennungsfrequenz der Kognaten\_f nach Wortarten sortiert berechnet, sondern es wurde auch der mögliche Einfluss des Längenverhältnisses auf die Kognatenpaare überprüft.

## **9.6 Identische Anfangsbuchstaben**

Die Frage, wie viele Buchstaben im Kognaten\_f und in den Kognaten\_dt übereinstimmen müssen, damit der Kognat erkannt wird, kann durch eine andere Frage erweitert werden: Ist die Position der übereinstimmenden Buchstaben relevant? Möller & Zeevaert (2015) haben anhand der Kommentare der Testteilnehmenden herausgefunden, dass Buchstabenunterschiede in der Mitte von Wörtern „logischer“ sind als am Anfang (siehe Möller & Zeevaert 2015:340). Daran erkennt man, dass Testpersonen wohl eher dazu tendieren sollten, die Buchstaben im Inneren eines Wortes an den Anfangsbuchstaben anzupassen, als den Anfangsbuchstaben an die inneren anzupassen. Demnach müssten Kognaten dt, die denselben Anfangsbuchstaben wie ihre entsprechenden Kognaten f besitzen, öfters erkannt werden, als jene, bei denen sich der Anfangsbuchstabe unterscheidet.

# **10. Verarbeitung der Testresultate**

## **10.1 Aufbau der Datensätze**

Die für diese Analyse verwendeten Daten wurden zwischen 2011 und 2014 erhoben und enthalten die Antworten von 249 Personen. Während der Studie wurden 24111 Antworten abgegeben, deren Rohdaten aufbereitet und in einzelne Datensätze umgewandelt wurden. Dies wurde mit drei für diese Studie geschriebenen Powershell-Programmen durchgeführt. Jeder der so erhaltenen Datensätze ist nach dem gleichen Muster aufgebaut:

*Kognat\_f ; Antwort ; Versuche ; nextTrue ; Zeit ; Kommentar ; userID ; file ; Wortart ; Kognat\_dt ;  
Resultat*

- *Kognat\_f* und *Antwort* geben den abgefragten Kognaten\_f und die von der Testperson abgegebene Antwort an.
- *Versuche* gibt an, bei welchem Versuch die Testperson für den gegebenen Kognaten ist.
- *nextTrue* ist ein Binärwert, der angibt, ob eine Person einen Kognaten übersprungen hat oder nicht.
- In *Zeit* steht die Zeit in Sekunden, die zwischen dem Erscheinen des Kognaten und dem Absenden der Antwort durch die Testperson verstrichen ist.
- Die Testteilnehmenden hatten die Möglichkeit, Kommentare zu ihren Antworten zu geben (auch wenn die meisten Testpersonen dies nicht getan haben) und diese sind unter *Kommentare* eingetragen.
- Die darauffolgende *userID* wurde anhand der verschiedenen personenspezifischen Daten (wie z.B. das Datum der Abfrage oder auch die Ursprungsdatei, in der die Antworten dieser Person zu finden sind) gebildet. Die Erschaffung dieser *userID* war notwendig, um die verschiedenen Daten der einzelnen Testpersonen voneinander zu trennen. Dies liegt daran, dass die abgefragten Kognaten\_f und die Antworten der Testpersonen alle automatisch zusammengefügt wurden und das Anhängen dieser *userID* ermöglichte es, nachzuvollziehen, welche Antwort von welcher Person abgegeben wurde.
- *file* bezeichnet die Ursprungsdatei, aus der der Datensatz stammt.
- *Wortart* gibt an, ob es sich bei dem Kognaten\_f (und dessen Kognaten\_dt) um ein Substantiv (S), ein Verb (V) oder ein Adjektiv/Adverb (A) handelt.
- *Kognat\_dt* gibt schließlich den richtigen Kognaten\_dt zu dem Kognaten\_f an und
- *Resultat* (ein weiterer binärer Wert) zeigt an, ob die Testperson den Kognaten in dieser Abfrage richtig erkannt hat.

Hier nun ein konkretes Beispiel für einen Datensatz:

**kreft;kraft;1;0;10;;8df1575d822984001e85343e3c93f63f;andere\_1.txt;(S);Kraft;1**

Aufgrund der sehr großen Datenmenge können jedoch nicht alle einzelnen Datensätze im Anhang aufgeführt werden. Für die 312 Kognaten (siehe Anhang) wurden insgesamt 4126 verschiedene Antworten angegeben. Diese werden jedoch alle (mit Ausnahme einiger Non-

Sense-Antworten wie z.B. einzelne Buchstaben) für die Analyse verwendet. Nun folgt die Vorgehensweise der Analyse, die an diesen Datensätzen durchgeführt wurde.

## **10.2 Vorgehensweise**

Im Prinzip kann man die Analyse eines Kognatenerkennungstests auf zwei verschiedene Arten durchführen. Zum einen kann man überprüfen, wie häufig die Kognaten\_dt für die jeweiligen Kognaten\_f angegeben wurden. Dabei wird nur überprüft, ob bzw. wie oft die richtige Antwort angegeben wurde. Alternativ kann man auch mit den einzelnen abgegebenen Antworten arbeiten.

Diese Datenanalyse ist deshalb in zwei Durchgänge eingeteilt. Im ersten Durchgang wird überprüft, welche der untersuchten Variablen einen Einfluss auf die Trefferquote hat, d.h. welche der den Kognaten\_f und deren Kognaten\_dt inhärenten Charakteristika einen Einfluss auf die richtige Erkennung der Kognaten haben. Bei dieser Methode werden die abgegebenen Antworten, die nicht dem Kognaten\_dt entsprechen, außen vor gelassen und es wird nur untersucht, wie häufig der Kognat\_dt angegeben wurde. Dadurch wird ein binäres Bewertungssystem geschaffen, das in Bezug auf die Interkomprehension keine Nuancen zulässt: Entweder ist ein Wort richtig oder es ist falsch. Diese binäre Sichtweise ist insofern hilfreich, als sie eine mehr oder weniger problemlose und nicht zu komplexe Analyse zulässt. Man kann sagen: Dieser Faktor führt zur erleichterten Erkennung des Kognaten und dieser nicht.

Im zweiten Durchgang geht es um die Analyse der Abfrageergebnisse, d.h. die von den Testpersonen abgegebenen Antworten werden untersucht, indem diese mit den richtigen Kognaten\_dt der abgefragten Kognaten und den Kognaten\_f selbst verglichen werden. Anhand von Eigenschaften der Kognatenpaare, wie z.B. Levenshtein-Distanzen, kann man so Rückschlüsse darauf ziehen, wie ähnlich die abgegebenen Antworten den Kognaten\_f und den Kognaten\_dt sehen. Dieses Verfahren erlaubt es, zu sehen, wie nah die Testpersonen an die richtige Antwort gekommen sind. Die Frage ist so nicht mehr, ob die Antworten richtig oder falsch sind, sondern ob die Antworten plausibel sind. Haben die Testpersonen tatsächlich Wörter vorgeschlagen, die dem Kognaten ähnlich sehen oder nicht.

Auf diese beiden Analysedurchgänge folgt ein Kapitel, dass sich mit nicht-systematischen Faktoren beschäftigt. Diese Faktoren spielen nicht immer eine Rolle, können aber, abhängig vom einzelnen Kognaten\_f und von den Testpersonen, die Erkennung des Kognaten\_f entweder

erschweren oder erleichtern. Dafür werden dann einzelne abgegebene Antworten der Testpersonen herangezogen, an denen diese nicht-systematischen Faktoren veranschaulicht werden können. Der nun folgende erste Analysedurchgang geht aber erst einmal auf die systematischen Faktoren ein, die die Erkennung von Kognaten\_f möglicherweise beeinflussen.

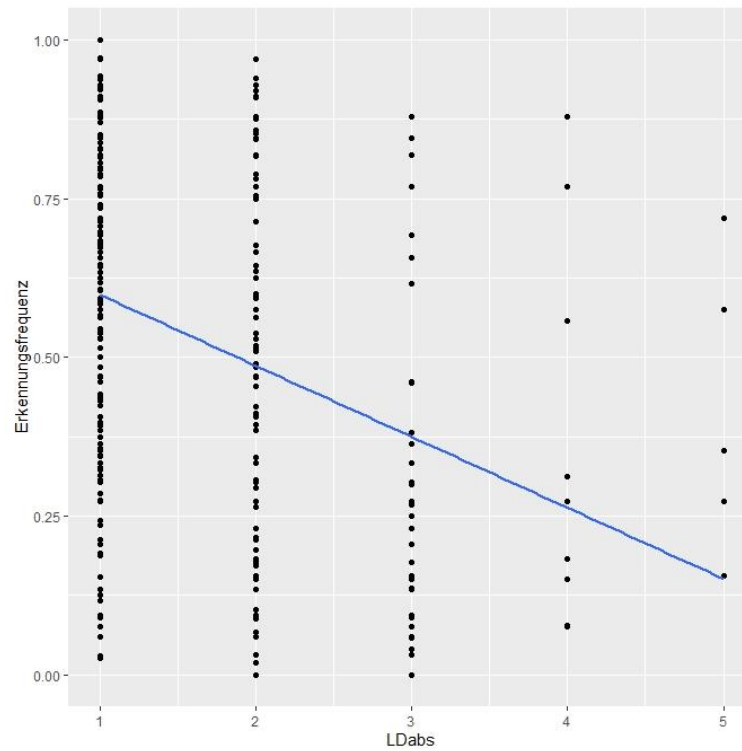
### 10.3 Analysedurchgang 1

Für diese Analyse wurden zwei Werte verwendet, anhand derer der Einfluss der verschiedenen Variablen berechnet wurde. Zum einen wurde die prozentuelle Anzahl an Testteilnehmenden, die einen bestimmten Kognaten richtig erkannt haben, herangezogen und der Einfluss der verschiedenen Faktoren auf den Erkennungserfolg untersucht. Zum anderen wurde dasselbe Verfahren auf die Erkennungserfolg für den ersten Abfrageversuch angewandt. Das bedeutet, dass ermittelt wurde, wie häufig ein gegebener Kognat\_f beim ersten Versuch richtig erkannt wurde. Dies ist interessant, da es einen Einblick in die Durchsichtigkeit der Kognatenverbindungen erlaubt. Zwei Kognaten\_f könnten dieselbe globale Trefferquote besitzen, aber sehr unterschiedliche Trefferquoten für den ersten Versuch aufweisen. Wenn die Korrelation zwischen einer Variablen und dem Erkennungserfolg geringer ist als für die gesamten Versuche, dann kann man davon ausgehen, dass diese Variable bei der Kognaten-Erkennung nicht sofort eine Rolle spielt und eine Antwort vielleicht aufgrund anderer Kriterien ausgesucht wird. Die Interpretation der Analyseergebnisse der verschiedenen Variablen wird am Ende dieses Analysedurchgangs vorgenommen.

#### 10.3.1 Absolute und relative Levenshtein-, Damerau-Levenstein- und gewichtete Levenshtein-Distanzen

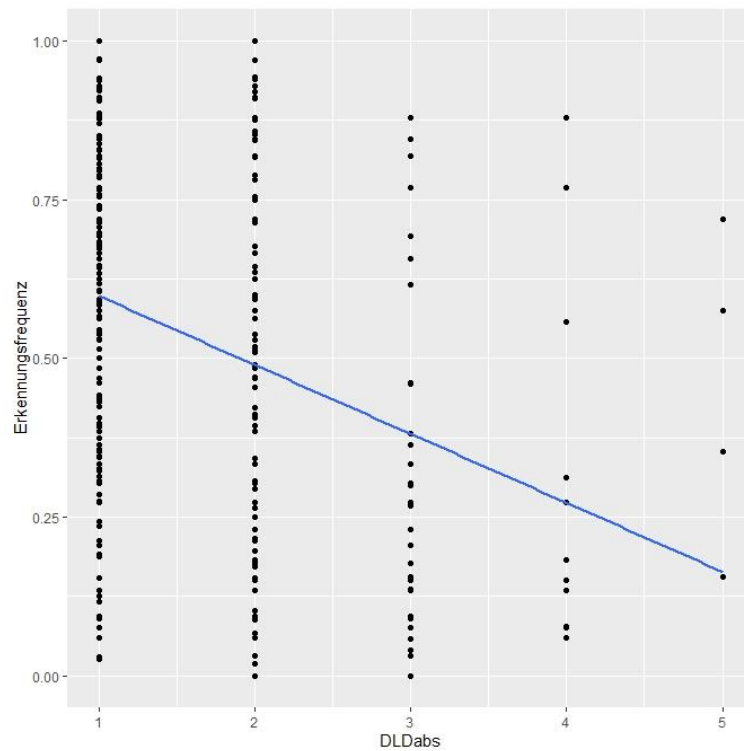
Die ersten Variablen, die an dieser Stelle überprüft werden sollen, sind die verschiedenen Arten von Levenshtein-Distanzen, d.h. die Editier-Distanzen, die dem Ähnlichkeitsgrad zwischen dem Kognaten\_f und dem Kognaten\_dt einen numerischen Wert zuschreiben. Dabei werden zuerst die absoluten und dann die relativen Distanzen, d.h. diejenigen, die von der Länge der Kognaten abhängig sind, miteinander verglichen. Die **Grafiken 1 – 3** zeigen die Distanzen (**Graf.1:** Levenshtein, **Graf.2:** Damerau-Levenshtein, **Graf.3:** gewichtete (phonologische) Levenshtein) im Verhältnis zum Anteil richtig erkannter Kognaten.

**Grafik 1:** Absolute Levenshtein-Distanz (LDabs) zwischen dem Kognaten\_f und dem dazugehörigen Kognaten\_dt im Verhältnis zur Erkennungsfrequenz; jeder der Punkte stellt ein Kognatenpaar dar.

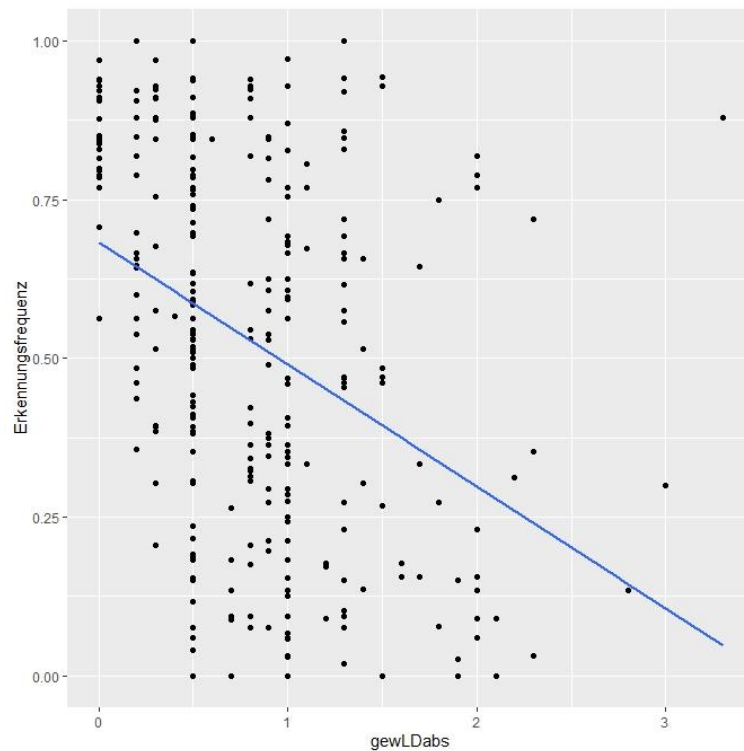


**Grafik 2:** Absolute Damerau-Levenshtein-Distanz (DLDabs) zwischen dem Kognaten\_f und dem dazugehörigen Kognaten\_dt im Verhältnis zur Erkennungsfrequenz; jeder der Punkte stellt ein Kognatenpaar dar.





**Grafik 3:** Absolute gewichtete phonologische Levenshtein-Distanz (gewLDabs) zwischen dem Kognaten\_f und dem dazugehörigen Kognaten\_dt im Verhältnis zur Erkennungsfrequenz; jeder der Punkte stellt ein Kognatenpaar dar.



Die blaue Linie stellt die Regressionsgerade dar und für die drei Distanzmessungen erkennt man einen eindeutigen negativen Trend. Anhand der Grafik der gewichteten Levenshtein-Distanz erkennt man eine gleichmäßigere Verteilung der Werte, da mehr Nuancen bei der

Berechnung dieser Distanz möglich sind. Dies ist darauf zurückzuführen, dass nicht alle Editierschritte die gewichtete Levenshtein-Distanz um eins erhöhen, sondern für manche Operationen kleinere Werte verwendet werden und so insgesamt mehr Datenpunkte entstehen.

Anhand der berechneten  $p$ -Werte (**Tab.4**) erkennt man, dass die Korrelationen zwischen den Distanzen und der Anzahl richtig erkannter Kognaten bei einem Signifikanzniveau von 0,001 statistisch signifikant sind. Dies bedeutet, dass die Werte mit einer sehr hohen Wahrscheinlichkeit (in diesem Fall 99,9%) nicht auf einen Zufall zurückzuführen sind.

Die Korrelationskoeffizienten (**Tab.4**) zeigen an, wie nahe die einzelnen Datenpunkte an der Regressionsgeraden anliegen. Je näher die Datenpunkte an der Regressionsgeraden liegen, desto besser ist die prädiktive Funktion der untersuchten Variable. Dieser Wert liegt immer irgendwo zwischen -1 und +1, wobei diese Grenzwerte eine maximale Korrelation beschreiben. Je näher dieser Wert an 0 liegt, desto geringer ist die Korrelation. Wenn die Korrelation einen positiven Wert hat, dann bedeutet dies, dass die Werte auf der Y-Achse ansteigen, wenn jene auf der X-Achse auch ansteigen. In diesem Fall sind die Korrelationskoeffizienten jedoch negativ, d.h., dass die Werte auf der Y-Achse sinken, wenn jene auf der X-Achse größer werden. Bei steigender orthographischer und phonologischer Distanz sinkt also die Erkennungsfrequenz.

**Tabelle 4:** Korrelationen zwischen den verschiedenen absoluten Levenshtein-Distanzen und dem Anteil richtig erkannter Kognaten und die dazugehörigen  $p$ -Werte

	Korrelationskoeffizienten	$p$ -Wert
Absolute LD	-0,351	1,67E-10
Absolute DLD	-0,343	4,98E-10
Absolute gewichtete LD	-0,374	8,26E-12

Eine Korrelation von 0,374 lässt auf einen etwas stärkeren Einfluss der absoluten gewichteten LD auf die Frequenz, mit der Kognaten erkannt werden, schließen. Für die phonologische Levenshtein-Distanz stimmen also die Punkte am ehesten mit der Regressionsgeraden überein und diese beschreibt besser die Abnahme der Erkennungsfrequenz bei zunehmender phonologischer Distanz. Die Werte liegen aber alle relativ nahe aneinander und es kann durchaus von einer recht guten Korrelation zwischen den Distanzen und der Erkennungsfrequenz ausgegangen werden.

An den Regressionskoeffizienten (**Tab.5**), die die Steigung der Regressionsgeraden beschreiben, kann man ablesen, wie stark eine Zunahme der absoluten gewichteten

Levenshtein-Distanz die Erkennung der Kognaten beeinträchtigt. Negative Regressionskoeffizienten beschreiben eine fallende Regressionsgerade, positive eine steigende und wenn der Regressionskoeffizient den Wert 0 hat, dann verläuft die Regressionsgerade parallel zur X-Achse.

**Tabelle 5:** Regressionskoeffizienten zwischen den verschiedenen absoluten Levenshtein-Distanzen und dem Anteil richtig erkannter Kognaten

	Regressionskoeffizienten
Absolute LD	-0,112
Absolute DLD	-0,109
Absolute gewichtete LD	-0,192

Der Regressionskoeffizient liegt hier für die einfache Levenshtein-Distanz bei -0,112, für die Damerau-Levenshtein-Distanz bei -0,109 und für die phonologische, also die gewichtete Levenshtein-Distanz bei -0,192. Das bedeutet, dass bei zunehmender Distanz in allen Fällen die Erkennungsfrequenz sinkt. Der sehr negative Regressionskoeffizient für die phonologische Distanz zeigt jedoch, dass die Kognaten\_dt, deren Aussprache sich sehr stark von jener der Kognaten\_f unterscheidet, sehr schnell nicht mehr erkannt werden. Pro Zunahme der absoluten gewichteten Levenshtein-Distanz um 1 sinkt die durchschnittliche Frequenz, mit der Kognaten erkannt werden, um 19,2%. Für die beiden orthographischen Distanzen (LD und DLD) ist dieser Wert nur halb so groß. Das bedeutet, dass bei einer identischen Zunahme der Distanz, die Erkennungsfrequenz nur ungefähr halb so schnell sinkt.

Anhand der Schnittpunkte der verschiedenen Regressionsgeraden mit der Y-Achse kann man erkennen, welche der Distanztypen für die jeweiligen Minimalwerte die höchste gemittelte Erkennungsfrequenz aufweist (**Tab.6**).

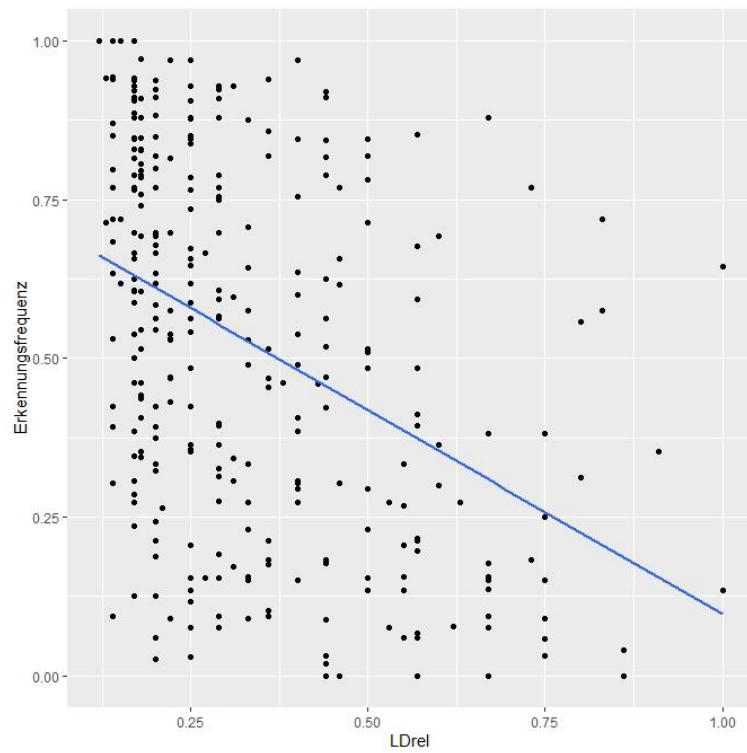
**Tabelle 6:** Schnittpunkte zwischen den Regressionsgeraden und den Y-Achsen für die absoluten Levenshtein-Distanzen (LD), Damerau-Levenshtein-Distanzen (DLD) und phonologischen Levenshtein-Distanzen (gewichtete LD)

	Schnittpunkt mit Y-Achse
Absolute LD	0,598
Absolute DLD	0,598
Absolute gewichtete LD	0,682

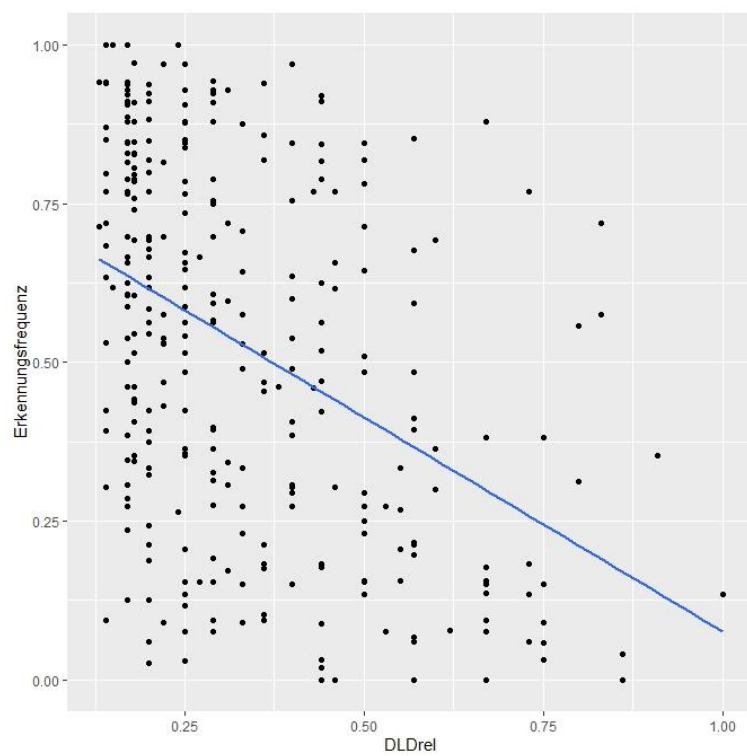
Die kürzesten absoluten orthographischen Distanzen, die einen Kognaten\_f vom entsprechenden Kognaten\_dt trennen, führen zu einer durchschnittlichen Erkennungsfrequenz von 59,8%. Die kleinste phonologische Levenshtein-Distanz macht es den Testpersonen möglich, den richtigen Kognaten\_dt in durchschnittlich 68,2% der Fälle zu finden. Dies ist ein Unterschied von fast 10%. Demnach werden die Kognatenpaare mit den phonologisch ähnlichsten Wörtern öfters richtig erkannt, als jene mit den orthographisch ähnlichsten. Dies sind jedoch nur die Resultate für die absoluten Distanzen, d.h., dass die Länge der Kognaten nicht in Betracht gezogen wurde.

Dieselben Berechnungen wurden hiernach ebenfalls für die relativen Distanzen durchgeführt. Diese repräsentieren die Anzahl Editierschritte dividiert durch die gemittelte Länge der beiden Kognaten des Kognatenpaares. In den **Grafiken 4, 5 und 6** sind die relativen Werte der Levenshtein-Distanz, Damerau-Levenshtein-Distanz und phonologischen, d.h. der gewichteten Levenshtein-Distanz dargestellt.

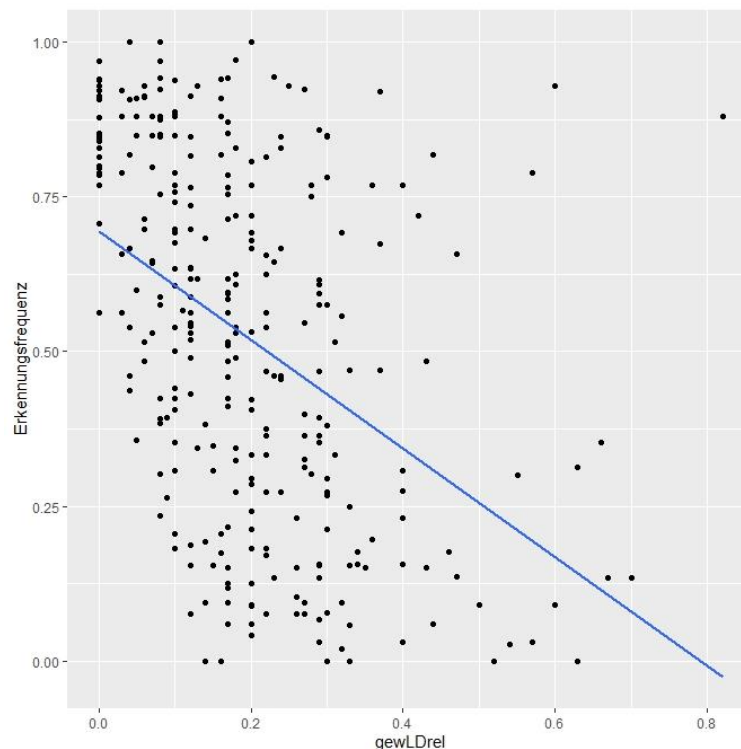
**Grafik 4:** Relative Levenshtein-Distanz (LDrel) zwischen dem Kognaten\_f und dem dazugehörigen Kognaten\_dt im Verhältnis zur Erkennungsfrequenz; jeder der Punkte stellt ein Kognatenpaar dar.



**Grafik 5:** Relative Damerau-Levenshtein-Distanz (DLDrel) zwischen dem Kognaten\_f und dem dazugehörigen Kognaten\_dt im Verhältnis zur Erkennungsfrequenz; jeder der Punkte stellt ein Kognatenpaar dar.



**Grafik 6:** Relative gewichtete (phonologische) Levenshtein-Distanz (gewLDrel) zwischen dem Kognaten\_f und dem dazugehörigen Kognaten\_dt im Verhältnis zur Erkennungsfrequenz; jeder der Punkte stellt ein Kognatenpaar dar.



Man erkennt wiederum einen eindeutig negativen Trend für die drei Regressionsgeraden. Alle Korrelationen sind für ein Signifikanzniveau von 0,001 statistisch signifikant (siehe  $p$ -Werte in **Tab.7**).

**Tabelle 7:** Korrelationen zwischen den verschiedenen relativen Levenshtein-Distanzen und dem Anteil richtig erkannter Kognaten und die dazugehörigen  $p$ -Werte

	Korrelationskoeff.	$p$ -Wert
Relative LD	-0,411	4,01E-14
Relative DLD	- 0,421	7,71E-15
Relative gewichtete LD	- 0,428	2,33E-15

Die Korrelationskoeffizienten sind alle bei den relativen Distanzen höher als bei den absoluten und liegen wieder sehr nahe beieinander. Die einzelnen Datenpunkte folgen also besser der Regressionsgeraden und die relativen Distanzen sind bessere Prädiktoren für die Erkennungsfrequenz der Kognatenpaare. Die höchste Korrelation ist bei der relativen phonologischen Distanz zu finden. Diese ist auch diejenige, die den mit Abstand höchsten Regressionskoeffizienten aufweist (**Tab.8**).

**Tabelle 8:** Regressionskoeffizienten der relativen Distanzen (**Graf.4**, **Graf.5** und **Graf.6**) und deren Differenz zu den Regressionskoeffizienten der absoluten Distanzen

	Regressionskoeffizient	Differenz Regress. Abs.LD
Relative LD	-0,642	+[-0,530]
Relative DLD	-0,675	+[-0,566]
Relative gewichtete LD	-0,877	+[-0,684]

Die Regressionskoeffizienten sind für die drei Distanzarten sehr stark angestiegen. Das „+“ vor den negativen Zahlen bei den einzelnen Differenzen bedeutet, dass die negative Regression größer geworden ist. Für die relative phonologisch gewichtete Levenshtein-Distanz ist sie von -0,192 auf -0,877 gestiegen. Das bedeutet, dass eine Erhöhung der relativen gewichteten LD um 0,5 die Erkennungsfrequenz um 43,85% senkt. Eine Erhöhung der relativen orthographischen Distanzen, also der Levenshtein- und Damerau-Levenshtein-Distanzen, senkt die Erkennungsfrequenzen um jeweils 32,1% und 33,75%. Hohe phonologische Unterschiede zwischen einem Kognaten\_f und einem Kognaten\_dt führen also schneller zur Nicht-Erkennung des Kognaten\_f als hohe orthographische Unterschiede. Diese Werte sind jedoch alle sehr hoch und weisen darauf hin, dass sowohl phonologische als auch orthographische Ähnlichkeit entscheidend für die Erkennungsfrequenz sind.

Bei den kleinstmöglichen Distanzen für die beiden orthographischen und die phonologische Distanz ist die Erkennungsfrequenz wieder bei der phonologischen am höchsten (**Tab.9**).

**Tabelle 9:** Schnittpunkte zwischen der Regressionsgeraden und der Y-Achse

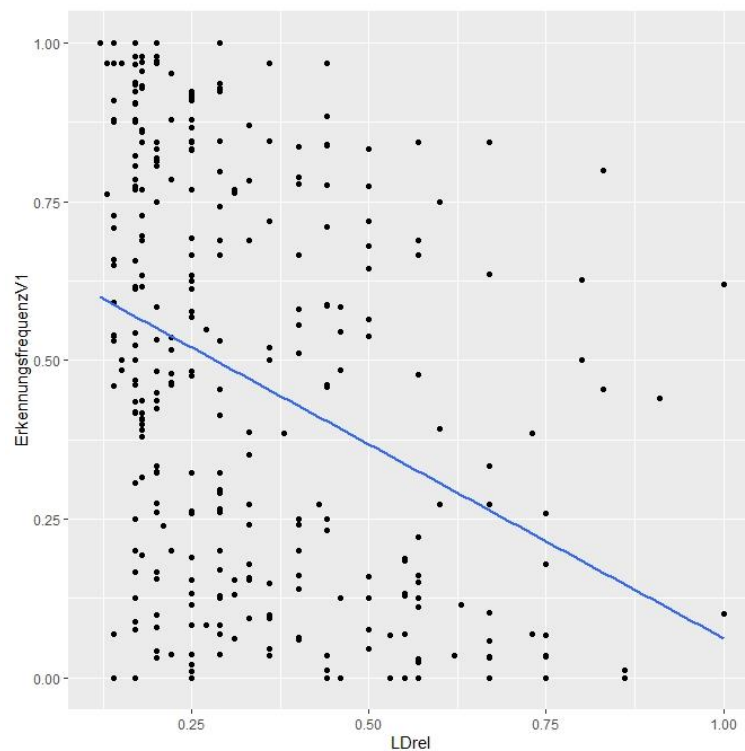
	Schnittpunkt mit Y-Achse
Relative LD	0,662
Relative DLD	0,663
Relative gewichtete LD	0,693

Die relative gewichtete Levenshtein-Distanz hat für ihren Minimalwert mit 69,3% die höchste durchschnittliche Erkennungsfrequenz. Die Kognatenpaare, in denen sich der Kognat\_f und der Kognat\_dt durch die kürzeste phonologische Levenshtein-Distanz unterscheiden, wurden durchschnittlich in 3,1% (bzw. 3%) mehr Fällen richtig erraten als jene Kognatenpaare, die die kleinste relative Levenshtein-Distanz (bzw. Damerau-Levenshtein-Distanz) aufweisen.

Insgesamt hat die phonologisch gewichtete Levenshtein-Distanz die besten prädiktiven Eigenschaften, wenn es um die Bestimmung der Erkennungsfrequenz anhand der Unterschiede zwischen dem Kognat\_f und dem Kognat\_dt geht.

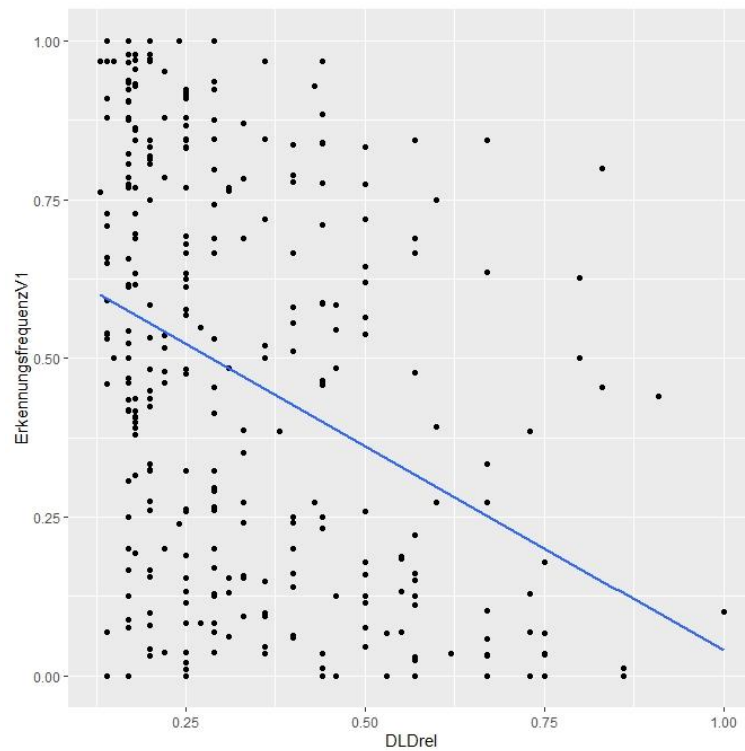
Dieselben Analysemethoden werden nun verwendet, um die Korrelationen zwischen diesen Distanzen und dem Anteil beim ersten Abfrageversuch richtig erkannter Kognaten zu bestimmen (**Graf.7**, **Graf.8** und **Graf.9**). Dabei wird jedoch nur mit den relativen Distanzen gearbeitet, da ja nur bei diesen alle Distanztypen statistisch signifikant waren. Die Streuung der Punkte in diesen Grafiken hat im Vergleich zu den Gesamt-Trefferquoten zugenommen. Dies ist ebenfalls an den niedrigeren Korrelationskoeffizienten erkennbar (**Tab.9**).

**Grafik 7:** Relative Levenshtein-Distanz (LDrel) zwischen dem Kognaten\_f und dem dazugehörigen Kognaten\_dt im Verhältnis zur Erkennungsfrequenz beim ersten Abfrageversuch (ErkennungsfrequenzV1); jeder der Punkte stellt ein Kognatenpaar dar.

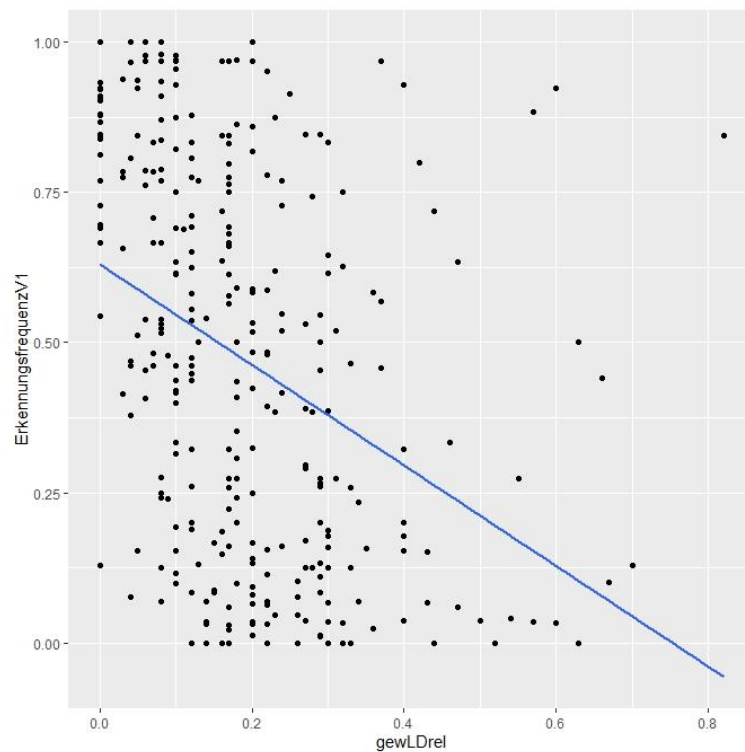




**Grafik 8:** Relative Damerau-Levenshtein-Distanz (DLDrel) zwischen dem Kognaten\_f und dem dazugehörigen Kognaten\_dt im Verhältnis zur Erkennungsfrequenz beim ersten Abfrageversuch (ErkennungsfrequenzV1); jeder der Punkte stellt ein Kognatenpaar dar.



**Grafik 9:** Relative gewichtete (phonologische) Levenshtein-Distanz (gewLDrel) zwischen dem Kognaten\_f und dem dazugehörigen Kognaten\_dt im Verhältnis zur Erkennungsfrequenz beim ersten Abfrageversuch (ErkennungsfrequenzV1); jeder der Punkte stellt ein Kognatenpaar dar.



Die Korrelationen zwischen den verschiedenen relativen Distanzen und den Erstversuch-Frequenzen richtig erkannter Kognaten sind nur ein wenig niedriger als bei den Gesamt-Trefferquoten (siehe **Tab.10**). Die Korrelationen sind statistisch signifikant (Signifikanzniveau: 0,001).

**Tabelle 10:** Korrelationen zwischen den verschiedenen relativen Levenshtein-Distanzen und der Erkennungsfrequenz beim ersten Versuch, die dazugehörigen  $p$ -Werte und die Differenz zu den Korrelationskoeffizienten für die Gesamt-Trefferquoten.

	Korrelationskoeff.	$p$ -Wert	Diff. Korrelation Gesamt-Trefferquoten
Relative LD	-0,353	1,32E-10	-[-0,058]
Relative DLD	-0,364	3,21E-11	-[-0,057]
Relative gewichtete LD	-0,370	1,46E-11	-[-0,058]

Die Tatsache, dass die Korrelationen etwas niedriger sind, zeigt, dass der Einfluss der relativen Levenshtein-, Damerau-Levenshtein- und gewichtete Levenshtein-Distanzen auf die Trefferquoten durch andere Faktoren beim ersten Erkennungsversuch beeinflusst werden, die sich bei den folgenden Versuchen nicht mehr so stark bemerkbar machen.

Ein Vergleich der Regressionskoeffizienten lässt erkennen, dass die Regression für alle drei Distanztypen gleichermaßen abgenommen hat (**Tab.11**).

**Tabelle 11:** Regressionskoeffizienten der verschiedenen relativen Levenshtein-Distanzen im Verhältnis zu den Erstversuch-Frequenzen und die Differenz zu den Regressionskoeffizienten für die Gesamt-Trefferquoten

	Regressionskoeff.	Diff. Regression Gesamt-Trefferquoten
Relative LD	-0,611	-[-0,031]
Relative DLD	-0,645	-[-0,030]
Relative gewichtete LD	-0,837	-[-0,040]

Dies zeigt, dass obschon, wie anhand der Korrelationskoeffizienten ersichtlich ist, der Einfluss der Distanzen auf die Erkennbarkeit der Kognaten\_f beim ersten Versuch niedriger ist, die Steigung doch mehr oder weniger gleich geblieben ist. Das bedeutet, dass die Einfachheit, mit der die Kognaten\_dt gefunden werden, bei steigenden orthographischen (LD und DLD) und phonologischen (gewichtete LD) Distanzen relativ schnell abnimmt.

Die Regressionsgerade schneidet die Y-Achse für die LD, DLD und gewichtete LD an einem niedrigeren Wert im Vergleich zu den jeweiligen Grafiken, die die relativen Distanzen mit den Gesamt-Trefferquoten vergleichen. Das bedeutet, dass die niedrigsten Distanzen beim ersten Versuch seltener die richtige Erkennung der Kognaten ermöglichen (**Tab.12**).

**Tabelle 12:** Schnittpunkte zwischen den Regressionsgeraden und der Y-Achse

	Schnittpunkt mit Y-Achse
Relative LD	0,599
Relative DLD	0,601
Relative gewichtete LD	0,630

Die Erkennungsfrequenzen der relativen Levenshtein-Distanz, der relativen Damerau-Levenshtein-Distanz und der relativen phonologischen Levenshtein-Distanz sind beim ersten Erkennungsversuch für die Minimaldistanzen um ungefähr 6% gesunken. Keine von diesen Distanztypen liefert bessere Resultate für die niedrigsten Distanzen beim ersten Versuch. Sie haben alle in gleichem Maße für den ersten Versuch abgenommen und daran wird ersichtlich, dass die relative Levenshtein-, Damerau-Levenshtein und phonologische (gewichtete) Levenshtein-Distanz bei späteren Versuchen eine größere Rolle spielen.

#### **Zwischenfazit: Levenshtein-, Damerau-Levenshtein- und gewichtete Levenshtein-Distanz**

Wenn es um die Identifikation der Kognaten\_dt anhand der Kognaten\_f geht, so erkennt man, dass die relativen Distanzen einen viel größeren Einfluss auf die Erkennungsfrequenz haben als die absoluten. Dies zeigt, dass die Anzahl Operationen pro Wort einer gewissen Länge äußerst entscheidend. Die durchweg schlechteren Korrelations- und Regressionswerte für den ersten Erkennungsversuch weisen darauf hin, dass andere Faktoren die Wahl der Antworten so stark beeinflussen, dass die Effekte der orthographischen und phonologischen Distanz beim ersten Kontakt mit dem Kognaten\_f manchmal eine weniger entscheidende Rolle spielen.

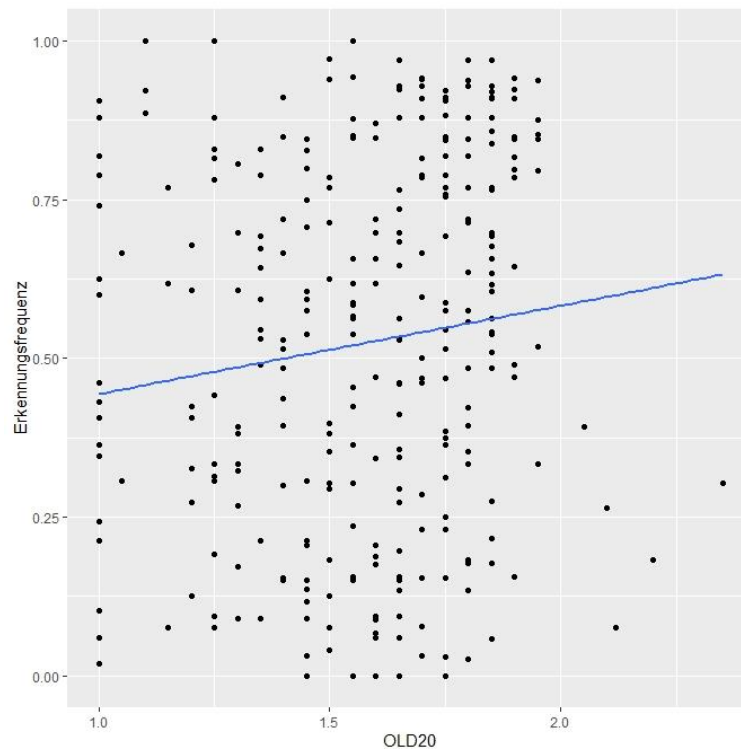
Die gewichtete phonologische Distanz zeigt den stärksten Einfluss auf die Frequenz, mit der Kognaten\_dt gefunden werden. Das weist darauf hin, dass die Testpersonen sich wohl eher an der (vermuteten) Phonologie orientieren sollten, um den Kognaten\_f richtig zu erkennen. Die Minimalwerte der drei Distanzen führen auch durchschnittlich am häufigsten bei der phonologischen Distanz zur richtigen Erkennung des Kognaten\_f. Die Damerau-Levenshtein-Distanz zeigt für alle Messwerte ähnliche oder sogar bessere Resultate als die klassische Levenshtein-Distanz. Die Präsenz der für die Bildung des Kognaten\_dt notwendigen Buchstaben im Kognaten\_f ist also durchaus förderlich für dessen Erkennung.

### 10.3.2 Die *Orthographic Levenshtein Distance 20* (OLD20)

Die nächste Variable, die nun untersucht wird, ist die *Orthographic Levenshtein Distance* (OLD20), also die gemittelte Levenshtein-Distanz der 20 nächsten Nachbarn der Kognaten\_dt. Hier wird auch wieder zuerst der Einfluss der OLD20 auf die Gesamt-Trefferquoten und danach auf die Erstversuch-Frequenz untersucht. Die OLD20-Werte, die hier verwendet werden, wurden mithilfe des Programmes *R* und dessen Erweiterung *visual word recognition* (*vwr*) berechnet. Diese Erweiterung verwendet Wörter aus der *CELEX Lexical Database* (Baayen et al. 1995).

In **Graf.10** kann man eine sehr große Streuung der einzelnen Punkte erkennen. Die Berechnung des *p*-Wertes für die Erkennung der Kognaten im Allgemeinen ergibt einen Wert von 0,03 und diese Resultate sind somit statistisch signifikant für ein Signifikanzniveau von 0,05. Der Korrelationskoeffizient (0,124) lässt jedoch erkennen, dass die OLD20 insgesamt nur einen sehr geringen Einfluss auf die richtige Identifizierung des Kognaten\_f hat. Dies ist auch graphisch an der sehr hohen Streuung der einzelnen Datenpunkte erkennbar. Man sieht jedoch sehr deutlich, dass die *Orthographic Levenshtein Distance 20* eine steigende Regression (Regressionskoeffizient: 0,139) aufweist. Das bedeutet, dass, obwohl die orthographischen Nachbarn des Kognaten\_dt keinen großen Einfluss auf die richtige Erkennung des Kognaten\_f haben, so steigt die Erkennungsfrequenz doch an. Je ähnlicher sich der Kognat\_dt und seine 20 nächsten Nachbarn also sehen, desto schwieriger wird es für die Testpersonen, diesen zu finden. Dies ist durchaus nachvollziehbar. Wenn die Testpersonen mehr mögliche Kognaten\_dt haben, weil sich viele Wörter sehr ähnlich sehen, dann ist es wahrscheinlicher, dass eines dieser anderen Wörter ausgewählt wird.

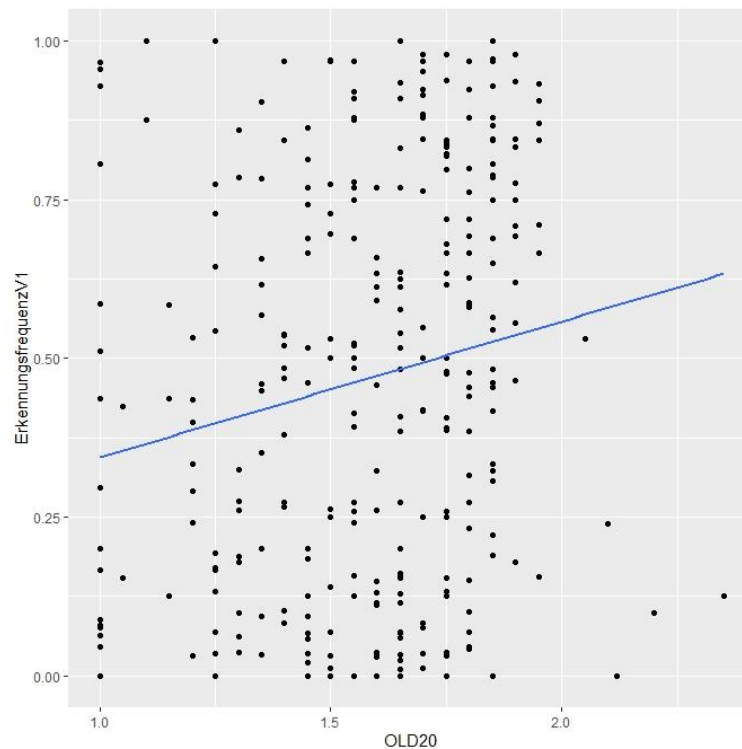
**Grafik 10:** OLD20 der Kognaten\_dt im Verhältnis zum Anteil richtig erkannter Kognaten\_f (Erkennungsfrequenz), jeder der Punkte stellt ein Kognatenpaar dar.



Der Einfluss der OLD20 auf die Erkennungsfrequenz ist beim ersten Versuch (**Graf.11**) schon ein wenig größer. Der Korrelationskoeffizient liegt hier bei 0,173 ( $p$ -Wert: 0,002; statistisch signifikant für ein Signifikanzniveau von 0,01).

Der Regressionskoeffizient hat ebenfalls zugenommen und liegt nun bei 0,214. Die Nähe zu den orthographischen Nachbarn ist also einflussreicher beim ersten als bei allen Versuchen zusammen genommen. Daraus kann man schließen, dass besonders beim ersten Erkennungsversuch die Kognaten\_dt, deren Nachbarn ihnen sehr ähnlich sehen, häufig nicht richtig erkannt werden. Die niedrigere Korrelation für alle Versuche im Vergleich zu jener des ersten ist vermutlich darauf zurückzuführen, dass diese falschen Wörter öfters beim ersten Versuch ausgesucht werden. Nachdem die Testpersonen dann erfahren haben, dass das abgegebene Wort nicht der Kognat\_dt war, können sie ein anderes Wort probieren, dass dann alleine durch den Ausschluss eines sehr ähnlichen Wortes wahrscheinlicher richtig erkannt wird. Dadurch spielt die Nähe eines Wortes zu seinen Nachbarn natürlich besonders am Anfang eine Rolle.

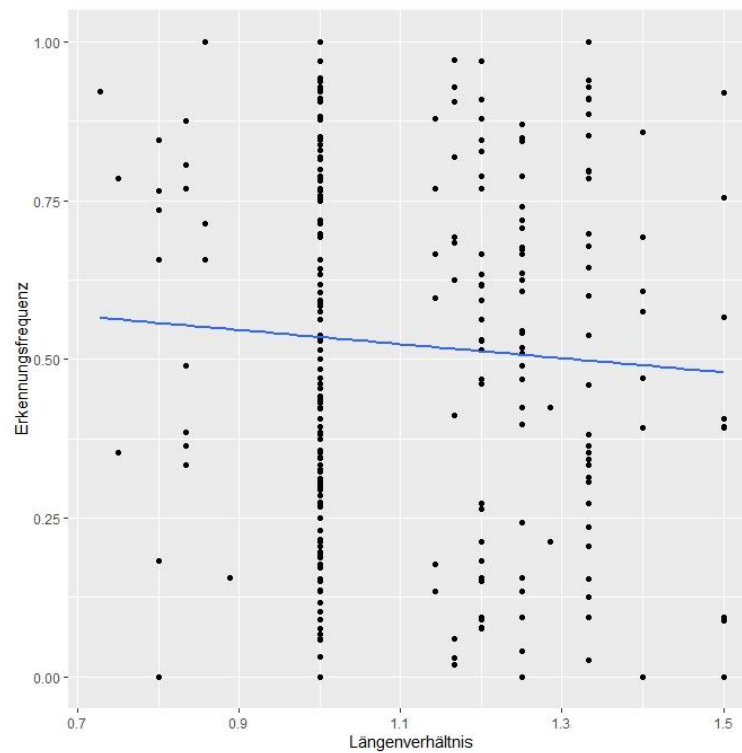
**Grafik 11:** OLD20 der Kognaten\_dt im Verhältnis zum Anteil beim ersten Versuch richtig erkannter Kognaten\_f (Erkennungsfrequenz), jeder der Punkte stellt ein Kognatenpaar dar.



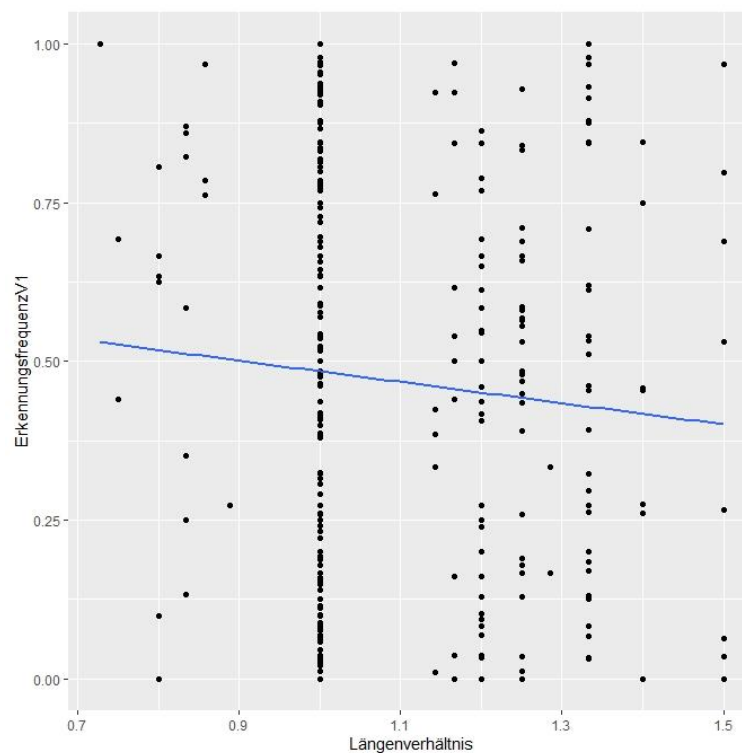
### 10.3.3 Wortlänge

Um diese Variable zu untersuchen, wurden die Längenverhältnisse der Kognatenpaare herangezogen. Dazu wurde die Länge des Kognaten\_dt durch die des Kognaten\_f geteilt. Aus den in **Graf.12** repräsentierten Werten wurden die Ausreißer entfernt. Der  $p$ -Wert für die Korrelation zwischen dem Längenverhältnis und der Erkennungsfrequenz beträgt 0,26 und ist damit nicht statistisch signifikant. Dasselbe gilt auch für die Korrelation zwischen dem Längenverhältnis und der Trefferquote beim ersten Versuch (**Graf.13**). Der Korrelationskoeffizient ist sehr niedrig (0,087) und die Ergebnisse sind statistisch nicht signifikant ( $p$ -Wert: 0,130).

**Grafik 12:** Längenverhältnis der Kognatenpaare im Verhältnis zum Anteil richtig erkannter Kognaten\_dt (Erkennungsfrequenz) jeder der Punkte stellt ein Kognatenpaar dar



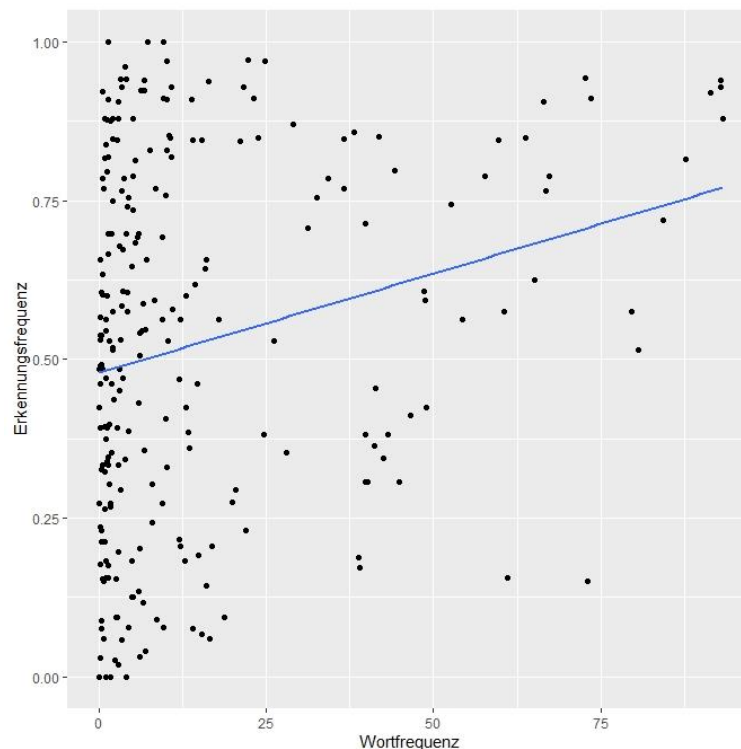
**Grafik 13:** Längenverhältnis der Kognatenpaare im Verhältnis zum Anteil beim ersten Versuch richtig erkannter Kognaten\_dt (ErkennungsfrequenzV1)



### 10.3.4 Die Wortfrequenz

Die Frequenzen für die Kognaten\_dt wurden aus dem online Wortschatzkorpus der Universität Leipzig bezogen. In diesem Korpus sind 720 421 868 Tokens, also einzelne Wörter, enthalten. Die Frequenzen wurden in Treffern pro 1 000 000 Wörter gerechnet. Wenn alle Frequenzwerte zur Berechnung der Korrelation verwendet werden, sind die Daten statistisch nicht signifikant. Nachdem die Ausreißer jedoch entfernt wurden, sieht man einen positiven Trend (**Graf.14**). Mit einem  $p$ -Wert von 0,0001 sind die Ergebnisse für einen Signifikanzwert von 0,001 statistisch signifikant. Die entfernten Ausreißer waren Wörter, die viel häufiger als die anderen vorkamen. Daraus kann man schließen, dass die Wortfrequenz nur einen Einfluss bei Kognaten dt hat, die nicht zu häufig vorkommen.

**Grafik 14:** Frequenz der Kognaten\_dt pro 1 000 000 Wörter im Verhältnis zum Anteil richtig erkannter Kognaten\_dt (Erkennungsfrequenz)



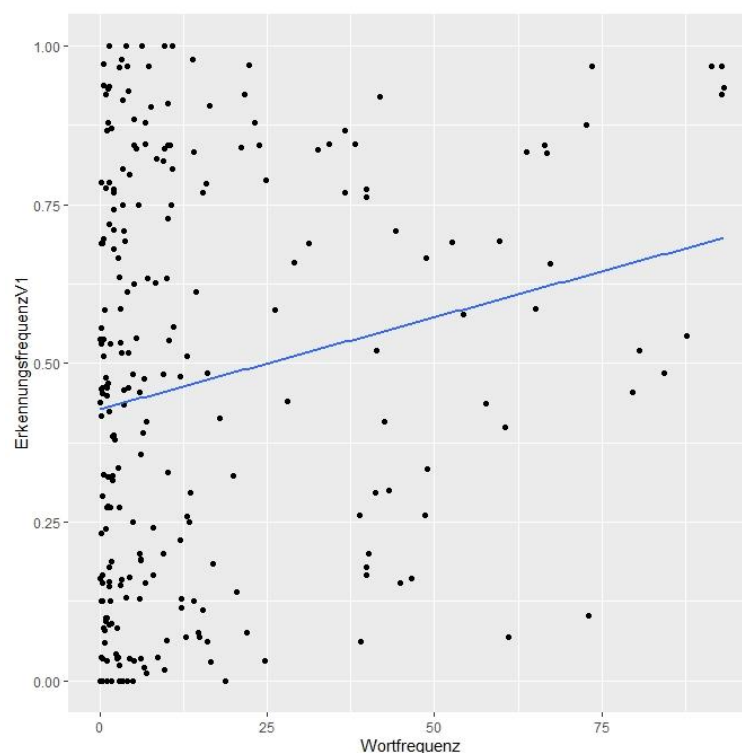
Der Korrelationskoeffizient liegt bei 0,238, was auf einen relativ kleinen Einfluss der Wortfrequenz auf die Erkennung des Kognaten\_f hindeutet. Der Regressionskoeffizient beträgt 0,003, d.h., dass die Erkennungsfrequenz bei einem Anstieg der Wortfrequenz um 1 um 0,3%



zunimmt. Der Einfluss der Wortfrequenz ist also nur sehr klein und beschränkt sich auf die niedrigfrequenten Kognaten dt.

Die gleichen Berechnungen wurden für den Frequenzeinfluss beim ersten Erkennungsversuch durchgeführt (**Graf.15**). Dort mussten auch wieder die höchsten Wortfrequenzen herausgenommen werden. Danach war diese Korrelation für ein Signifikanzniveau von 0,01 (bei einem  $p$ -Wert von 0,0015) statistisch signifikant. Beim ersten Versuch zeigt sich keine bessere Verteilung der Werte entlang der Regressionsgeraden. Dies wird auch am Korrelationskoeffizienten ersichtlich, der im Vergleich zu den Gesamtabfragen abgenommen hat: 0,198. Der Regressionskoeffizient hingegen ist gleich geblieben.

**Grafik 15:** Frequenz der Kognaten\_dt pro 1 000 000 Wörter im Verhältnis zum Anteil beim ersten Versuch richtig erkannter Kognaten\_dt (Erkennungsfrequenz)

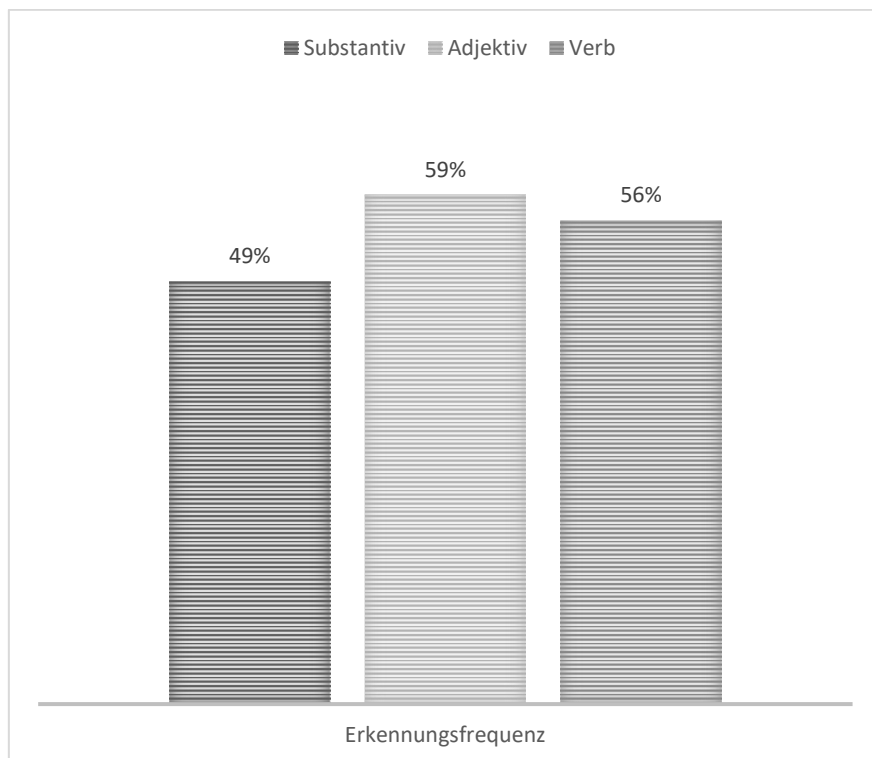


Insgesamt hat die Wortfrequenz, mit der ein Kognat\_dt vorkommt, nur einen sehr kleinen Effekt auf die Erkennung der Kognaten. Hinzu kommt, dass dieser Effekt nur bei seltenen Kognaten\_dt auftritt und so nur bei diesen einen Einfluss auf die Erkennung des Kognaten\_f hat. Ab einer gewissen Frequenz spielt ebendiese also keine Rolle mehr bei der Kognaten-Erkennung.

### 10.3.5 Die Wortart

Die Berechnung der gemittelten Frequenzen für Substantive, Verben und Adjektive haben ungleiche Erkennungsfrequenzen der Kognaten\_dt ergeben (**Graf.14**). Die Adjektive sind am häufigsten richtig erkannt worden und weisen eine Differenz von 10% zu den Substantiven auf, die am seltensten richtig erkannt werden. Die Verben liegen mit 56% nur knapp hinter den Adjektiven.

**Grafik 16:** Durchschnittliche Erkennungsfrequenzen der Kognatenpaare nach Wortarten



Um zu prüfen, woran diese Unterschiede liegen könnten, wurde die Korrelation zwischen den Erkennungsfrequenzen der einzelnen Wortarten und der gemittelten Länge der Kognatenpaare für die drei Wortarten berechnet. Es ergibt sich ein sehr niedriger Korrelationskoeffizient (0,066) und dieser ist überhaupt nicht statistisch signifikant ( $p$ -Wert: 0,958). Es wurde ebenfalls überprüft, ob die Wortfrequenz der verschiedenen Kognatenpaare, die nach Wortarten sortiert waren, mit den Erkennungsfrequenzen korrelieren. Es ergibt sich ein sehr hoher Korrelationskoeffizient (0,983), der jedoch statistisch nicht signifikant ist ( $p$ -Wert: 0,117). Weder die Frequenz noch die gemittelte Länge der Kognatenpaare haben also einen statistisch signifikanten Einfluss auf die Erkennungsfrequenz. Die deutlichen Unterschiede zwischen den Wortarten werden also nicht durch diese beiden Faktoren erklärt.

### 10.3.6 Die Präsenz identischer Anfangsbuchstaben

Die Analyse der Erkennungsfrequenzen für Kognaten\_dt, bei denen die Anfangsbuchstaben dieselben wie im Kognat f sind, zeigt, dass es keine großen Unterschiede zwischen diesen beiden gibt. Die gemittelte Erkennungsfrequenz für die Kognatenpaare, in denen der Anfangsbuchstabe in beiden Wörtern nicht identisch ist, liegt bei 51,95%. Dieser Wert liegt nicht einmal einen Prozentpunkt unterhalb der Erkennungsfrequenz für die Kognatenpaare mit denselben Anfangsbuchstaben (52,58%). Die Differenz zwischen diesen Werten ist beim ersten Versuch zwar noch etwas größer, aber dort ist es das Kognatenpaar mit unterschiedlichen Anfangsbuchstaben, das häufiger richtig erkannt wird:

- Identischer Anfangsbuchstabe: 46,63%
- Unterschiedlicher Anfangsbuchstabe: 48,42%

Es lassen sich also keine eindeutigen Schlüsse aus diesen Ergebnissen ziehen. Außerdem hat die Berechnung der Korrelation zwischen der Präsenz unterschiedlicher (bzw. identischer) Anfangsbuchstaben und der Erkennungsfrequenz einen sehr niedrigen Korrelationskoeffizienten (0,008) und keine statistisch signifikanten Ergebnisse ergeben ( $p$ -Wert: 0,893). Wenn man anstelle der orthographischen Anfangsbuchstaben die Anfangsphoneme vergleicht, so zeigt sich noch immer eine sehr niedrige Korrelation (0,068), die ebenfalls nicht statistisch signifikant ist ( $p$ -Wert: 0,233). Es gibt also keinen Hinweis darauf, dass die Anfangsbuchstaben oder die Anfangsphoneme eine Rolle bei der Erkennung der Kognaten\_f spielen.

## 10.4 Interpretation des Analysedurchgangs 1

Der Analysedurchgang 1 hat gezeigt, wie unterschiedlich – wenn überhaupt – sich die verschiedenen Faktoren auf die Erkennung von Kognaten\_f auswirken. Hier wird die Analyse dieser verschiedenen Faktoren in derselben Reihenfolge ausgewertet, in der sie zuvor analysiert wurden.

#### 10.4.1 Die absolute und relative Levenshtein-, Damerau-Levenshtein- und gewichtete Levenshtein-Distanzen

Die absoluten Distanzen spielen bei den Levenshtein-Distanzen und deren Varianten die kleinste Rolle. Die Anzahl der Operationen ist alleine nicht so ausschlaggebend wie die Anzahl der Operationen, die in einem Kognaten\_f gegebener Länge notwendig sind, um diesen in den entsprechenden Kognaten\_dt zu verwandeln und so sind die relativen Distanzen bedeutend einflussreicher. Die relativen Distanzen sind ein besserer Indikator für die Ähnlichkeit zwischen dem Kognaten\_f und dem Kognaten\_dt. Von den drei Distanzen, die untersucht wurden, ist die phonologische Levenshtein-Distanz immer diejenige, die die höchste Korrelation aufweist und damit den größten Einfluss auf die Erkennungsfrequenz hat. Außerdem sinken bei zunehmender phonologischer Levenshtein-Distanz die Erkennungsfrequenzen der Kognatenpaare viel schneller, als es bei den beiden orthographischen Distanzen der Fall ist. Daran erkennt man, dass die Testpersonen orthographische Unterschiede leichter überbrücken können als phonologische. In Kombination mit den durchschnittlich höheren Erkennungsfrequenzen für die phonologische Distanz kann man davon ausgehen, dass die Erkennung der Kognaten durch ähnliche Phonologie am stärksten vereinfacht wird. Die Testpersonen stützen sich auf die deutsche Aussprache der Buchstaben(-kombinationen), die in den Kognaten\_f vorkommen. Dadurch ist es natürlich auch wichtig, dass sich der schriftliche Ausdruck nicht zu sehr von dem des Kognaten\_dt unterscheiden darf. Von den orthographischen Distanzen zeigt die Damerau-Levenshtein-Distanz den größeren Einfluss auf die Erkennung der Kognaten\_f. Die relative Damerau-Levenshtein-Distanz beeinflusst die Erkennungsfrequenz sogar fast genauso sehr wie die phonologische LD. Obwohl die Resultate der einfachen Levenshtein- und der Damerau-Levenshtein-Distanz häufig fast identisch sind, erweisen sich die Unterschiede dennoch als interessant. Die Damerau-Levenshtein-Distanz, die ja die Transposition von nebeneinander liegenden Buchstaben zulässt, zeigt, dass die Präsenz der Buchstaben des Kognaten\_dt im Kognaten\_f bei der Erkennung eine Rolle spielt.

Die Tatsache, dass die Erkennungsfrequenzen der ersten Versuche weniger mit den verschiedenen relativen Distanzen korrelieren als die Gesamt-Trefferquoten, ist äußerst interessant. Dies könnte auf verschiedene Ursachen zurückzuführen sein. Zum einen könnte dies zeigen, dass die Personen bei den ersten Versuchen seltener auf die Distanzen, sei es nun orthographische oder phonologische, geachtet haben, als es im Durchschnitt für alle Versuche der Fall gewesen ist. Daraus könnte man schließen, dass, zumindest in einem Teil der Abfragen, zuerst ein Wort auf der Basis anderer Kriterien ausgewählt wurde und erst danach, also

nachdem die Antwort als falsch abgelehnt wurde, die Distanzen zwischen Kognat\_f und Kognat\_dt in Betracht gezogen wurden. Zum anderen könnte dies aber auch einfach daran liegen, dass es ein anderes Wort gibt, das dem Kognaten\_f genauso ähnlich sieht wie das Kognat\_dt. Dieser zweite Fall kann natürlich nur überprüft werden, wenn man sich die einzelnen Antworten genauer ansieht. Das wird im zweiten Analysedurchgang getan.

#### 10.4.2 Die *Orthographic Levenshtein Distance 20* (OLD20)

Die *Orthographic Levenshtein Distance 20* (OLD20) zeigt zwar nur einen sehr geringen Einfluss auf die Erkennung der Kognaten\_f, aber der Unterschied zwischen dem Einfluss auf die Trefferquote des ersten Versuches und der Gesamt-Trefferquote ist sehr interessant. Wie bereits oben erklärt, werden die Wörter, die ihren orthographischen Nachbarn sehr ähnlich sehen, besonders beim ersten Versuch falsch erkannt. Wenn viele Buchstaben von einem Wort zum nächsten übereinstimmen, dann ist es sehr wahrscheinlich, dass nicht der wahre Kognat\_dt ausgewählt wird. Obwohl die OLD20 beim ersten Versuch nur eine relativ kleine Rolle spielt, muss man davon ausgehen, dass sie in Kombination mit anderen Faktoren die Auswahl der richtigen Antwort in einem Kognaten-Erkennungstest beeinflusst. Dies wird auch durch Kommentare von Testpersonen bestätigt. Das folgende Zitat stammt von einer Person, die den Kognaten\_f *batter* richtig erkannt hat (Kognat\_f *bitter*): „gibt nicht viele Adjektive im Deutschen, die dem ähneln“. Hier haben die Wortart und die OLD20 zusammen zum Erkennen des richtigen Kognaten\_dt geführt. Ob dies aber in der Tat nur an diesen beiden Faktoren gelegen hat oder nicht, ist nicht mit Sicherheit zu sagen, da vielleicht auch unterbewusst andere Faktoren eine Rolle gespielt haben.

#### 10.4.3 Die Wortlänge

Da die Analyse des Einflusses der Wortlänge auf die Erkennungsfrequenz der Kognaten\_f keine statistisch signifikanten Resultate geliefert hat, kann man nicht sagen, ob die Länge des Wortes an sich wirklich eine Rolle spielt. Die Länge sowohl des Kognaten\_f als auch des Kognaten\_dt wird natürlich direkt in den Berechnungen der relativen Levenshtein-, Damerau-Levenshtein- und phonologischen Levenshtein-Distanzen miteinbezogen. Anhand der hier präsentierten Ergebnisse lassen sich jedoch keine Schlüsse darüber ziehen, ob eher das Hinzufügen von Buchstaben an oder das Wegnehmen von Buchstaben vom Kognaten\_f die Erkennung erleichtern.

#### 10.4.4 Die Wortfrequenz

Die Tatsache, dass nur bei den kleineren Wortfrequenzen überhaupt statistisch signifikante Korrelationen auftreten, deutet darauf hin, dass ab einer gewissen Frequenz ein Kognat\_dt wohl als „frequent“ angesehen werden kann. Ab diesem Punkt scheint es nicht mehr wichtig zu sein, wie groß die Frequenz ist. Demnach sind wohl auch nur die Kognaten\_f, deren Kognaten\_dt nur sehr selten vorkommen, besonders schwierig zu erkennen. Diese Ergebnisse decken sich durchaus mit dem *Interactive Activation Model* (IAM). Dieses geht ja davon aus, dass die verschiedenen Wortknoten, die die graphemische, phonologische und semantische Seite des Wortes enthalten, unterschiedliche Aktivationslevel haben, die unter anderem von der Frequenz des Wortes abhängig sind (McClelland & Rumelhart 1981:...). Die äußerst seltenen Kognaten\_dt werden nur sehr schwer (wenn überhaupt) erkannt, da der Eintrag für dieses Wort im mentalen Lexikon nicht häufig abgerufen wird.

#### 10.4.5 Die Wortart

Obwohl die Wortarten teilweise sehr große Unterschiede in der Erkennungsfrequenz aufweisen, (Differenz von 10% zwischen den Substantiven und den Adjektiven), haben die Korrelationsberechnungen für einen möglichen Einfluss der Frequenzen und der Längenverhältnisse der Kognatenpaare keine statistisch signifikanten Resultate geliefert. Es ist möglich, dass die Substantive eine relativ niedrige Erkennungsfrequenz haben, weil sie nicht als Substantive erkannt worden sind. Bei jeder Abfrage stand zwar daneben, um welche Wortart es sich bei dem Kognaten\_f (und so auch Kognaten\_dt) handelte, aber es ist durchaus möglich, dass die Testpersonen aufgrund des fehlenden Großbuchstabens am Wortanfang den Kognaten\_f nicht als ein Substantiv erkannt haben. Das dies durchaus vorkommt, ist am folgenden Kommentar ersichtlich: „Das (S) in Klammern hinter dak hat mich auf die Idee gebracht. Allerdings weiß ich nicht, weshalb das S umklammert ist.“ Es wurde der Kognat\_f *dak* abgefragt und die Testperson hat *Dachs* als Antwort angegeben. Diese Antwort war zwar auch ein Substantiv, aber die Testperson wusste offensichtlich nicht, dass das „S“ bedeutet, dass es sich bei dem Kognaten\_f um ein Substantiv handelt.

#### 10.4.6 Die Präsenz identischer Anfangsbuchstaben

Die Anfangsbuchstaben haben keinen systematischen Einfluss auf die Erkennungsfrequenz der Kognaten\_f. Es ist jedoch auch bei diesem Faktor sehr wahrscheinlich so, dass dieser von den Kognatenpaaren abhängig ist. Dies kann anhand von zwei Kognatenpaaren aus diesem Test illustriert werden. Zum einen gab es in diesem Test den Kognaten\_f *twilling*, der immer richtig als *Zwilling* erkannt wurde. Diese Erkennungsfrequenz ist trotz unterschiedlicher Anfangsbuchstaben vermutlich deshalb so hoch, weil es keine anderen Wörter im Deutschen gibt, bei denen nach einem Anfangsbuchstaben die Buchstabenfolge *willing* steht. Die sehr geringe phonologische Distanz zwischen /t/ und /ts/ ist vermutlich ebenfalls eine Hilfe bei der Erkennung. Der Kognat\_f *ár*, dessen Kognat\_dt *Jahr* ist, hatte nur eine Erkennungsfrequenz von 13,40%. In diesem Fall spielt die Länge des Wortes an sich vermutlich eine Rolle, da nicht sehr viele orthographische oder phonologische Elemente zur richtigen Identifikation des Kognaten\_f verwendet werden können. Die Tatsache, dass in diesem Fall der erste Buchstabe nicht ersetzt, sondern fallen gelassen wurde, ist vermutlich auch ausschlaggebend. Bei *twilling* hatten die Testpersonen einen Anhaltspunkt, der bei diesem Kognaten\_f fehlt.

### 10.5 Analysedurchgang 2

An dieser Stelle werden die von den Testpersonen abgegebenen Antworten untersucht. Dabei spielt es keine Rolle, ob die Antworten den Kognaten\_dt entsprechen oder nicht. Dazu wird von einigen der im ersten Analysedurchgang bereits verwendeten Faktoren Gebrauch gemacht. Der erste Teil ist eine Gesamtanalyse, die sich mit allen abgegebenen Antworten beschäftigt. Anhand der absoluten Levenshtein-Distanzen wird ermittelt, wie viele Editierschritte die Testpersonen meistens verwenden, um die ihrer Meinung nach dem Kognaten\_dt entsprechende Antwort zu finden.

Der zweite Teil betrachtet die häufigsten Antworten, die nicht die Kognaten\_dt sind, genauer. Da diese häufig von sehr vielen Personen ausgewählt wurden, ist es interessant zu sehen, warum genau diese Wörter für die Testteilnehmenden als plausible Kognaten\_dt angesehen werden. Da die phonologische Distanz auf von Testpersonen präferierten Lautentsprechungen zwischen dem Deutschen und anderen germanischen Sprachen basiert, kann man mit der Hilfe der gewichteten Levenshtein-Distanz überprüfen, ob die Antworten der Testpersonen plausibel sind.

### 10.5.1 Gesamtanalyse der abgegebenen Antworten

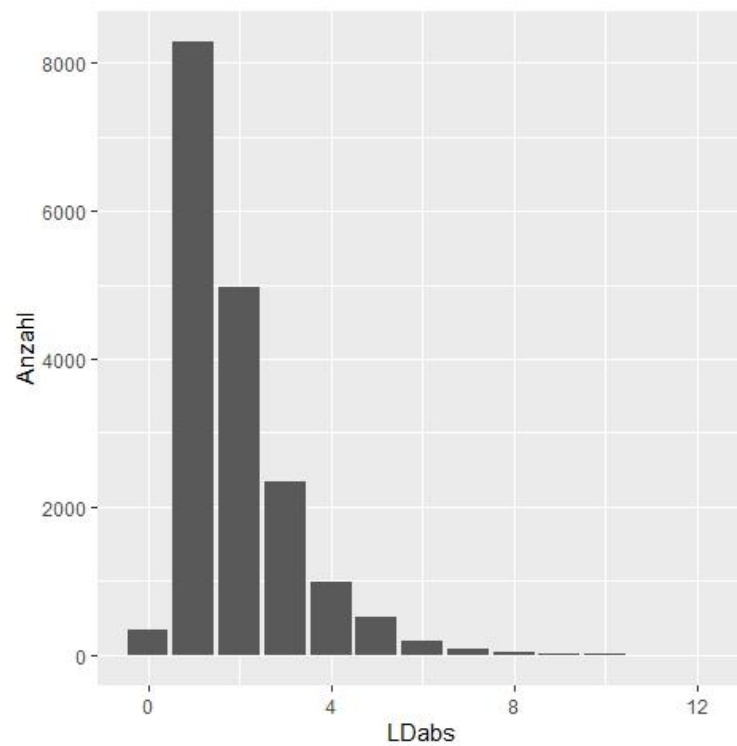
Die absoluten Distanzen werden hier zuerst analysiert, da anhand dieser Werte bestimmt werden kann, wie viele einzelne Operationen die Testpersonen anwenden, um vom angegebenen Kognaten\_f zu ihrer abgegebenen Lösung zu gelangen. So wird deutlich, in welchem Maße die Testpersonen bereit sind, einen gegebenen Kognaten\_f zu verändern, um den möglichen Kognaten\_dt zu bilden.

Die absoluten Levenshtein und Damerau-Levenshtein-Distanzen zeigen sehr deutlich, dass die Testpersonen dazu tendieren, die kleinstmögliche Anzahl an Operationen (unabhängig von der Länge des Kognaten\_f) am Kognaten\_f durchzuführen um den potentiellen Kognaten\_dt zu bilden (**Graf.17** und **Graf.18**). Wie man erkennen kann, bestehen hier keine größeren Unterschiede zwischen diesen beiden Distanztypen. Bei der absoluten Damerau-Levenshtein-Distanz sind jedoch mehr Antworten zu finden, die sich durch eine Distanz von 1 von den Kognaten\_f unterscheiden. Diese zusätzlichen Antworten sind die Fälle, in denen durch eine Transposition, also durch den Platztausch von zwei nebeneinander liegenden Buchstaben, ein Editierschritt der einfachen Levenshtein-Distanz wegfällt.

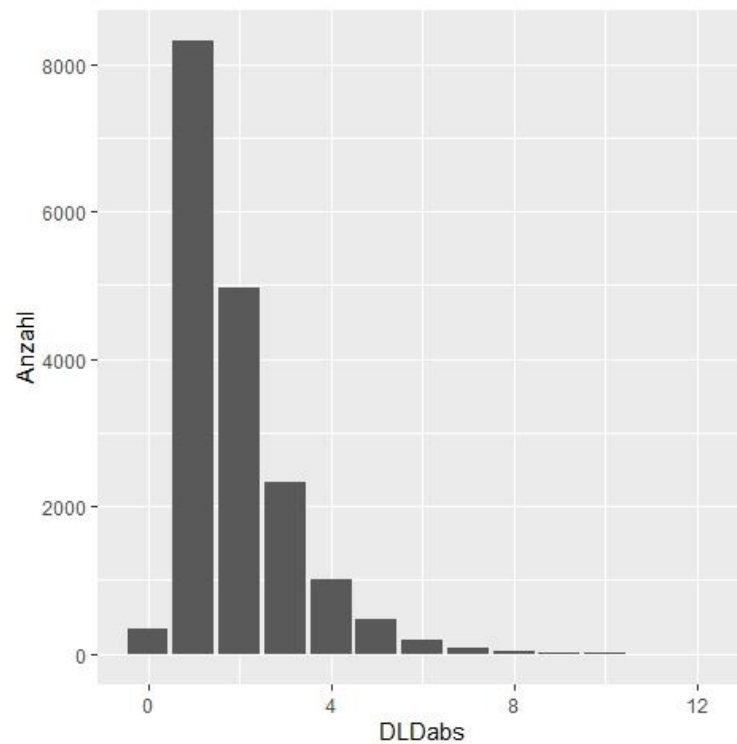
Interessanterweise wurden relativ häufig Wörter mit orthographischen Distanzen von 0 abgegeben. Das bedeutet, dass die Testpersonen einfach nur den Kognaten\_f als potentiellen Kognaten\_dt angegeben haben. Die meisten Antworten der Testteilnehmenden unterscheiden sich durch eine orthographische Distanz von 1 von den Kognaten\_dt und mit steigender absoluter orthographischer Distanz sinkt die Anzahl der Antworten exponentiell. Sowohl für die absolute LD als auch die absolute DLD sind etwas mehr als 46% der Antworten eine Operation von den jeweiligen Kognaten\_f entfernt. Die zweithäufigsten Antworten sind jene mit einer Levenshtein-Distanz von 2 mit fast 28% der Gesamtantworten. Dies deutet darauf hin, dass die meisten Personen in der Tat versuchen, ihre Antwort anhand des gegebenen Kognaten\_f durch das Verändern einzelner Buchstaben zu bilden.



**Grafik 17:** Anzahl der Antworten (Anzahl), die sich in einem gegebenen absoluten LD-Wert (LDabs) vom Kognat  $f$  unterscheiden



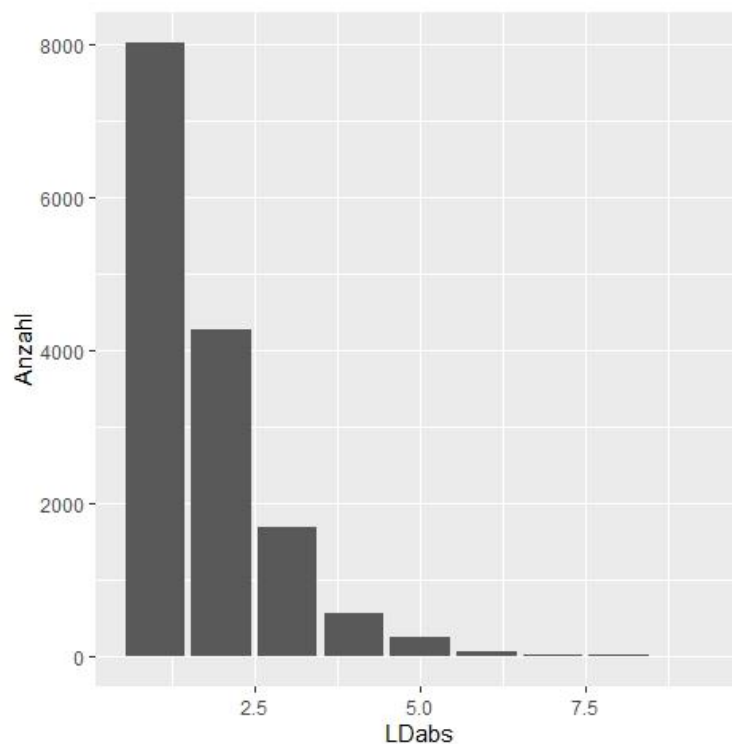
**Grafik 18:** Anzahl der Antworten (Anzahl), die sich in einem gegebenen absoluten DLD-Wert (DLDabs) vom Kognat  $f$  unterscheiden



Der Korrelationskoeffizient zwischen den absoluten Levenshtein-Distanzen der Kognaten\_f und Kognaten\_dt und den absoluten LD der Kognaten\_f und abgegebenen Antworten beträgt 0,522 ( $p$ -Wert: 3,34E-23, signifikant für ein Signifikanzniveau von 0,001). Das bedeutet, dass bei steigender Distanz zwischen dem Kognaten\_f und dem Kognaten\_dt auch die Distanz zwischen dem Kognaten\_f und den Antworten der Testpersonen zunimmt. Etwas weniger als die Hälfte der gesamten Antworten (46,6%) unterscheiden sich durch eine Levenshtein-Distanz von 1 von den Kognaten\_f.

Viele der Antworten sind Wörter, die nur ein einziges mal angegeben wurden. Bei solchen Einzelantworten ist die Wahrscheinlichkeit natürlich größer, dass die Auswahl eines bestimmten Wortes nicht wirklich auf einem System basiert, sondern auf eine individuelle Entscheidung der betreffenden Testpersonen zurückzuführen ist. So ist es gut möglich, dass ein Wort aufgrund einer bestimmten Assoziation ausgewählt wurde, die nur diese eine Testperson gebildet hat. Für **Grafik 19** wurden deshalb alle Antworten entfernt, die nur ein einziges Mal angegeben wurden. Außerdem wurden die Antworten entfernt, die sich nicht vom Kognaten\_f unterscheiden. Diese Analyse wird nur mit der einfachen absoluten LD durchgeführt, da sich die DLD nur unwesentlich von dieser unterscheidet.

**Grafik 19:** Anzahl der Antworten (Anzahl), die sich in einem gegebenen absoluten LD-Wert (LDabs) vom Kognaten\_f unterscheiden, mit Ausnahme der den Kognaten\_f identischen und der nur einmal angegebenen Antworten.



Es ist erkennbar, dass die Anzahl der Antworten, die sich durch eine Operation von den Kognaten *f* unterscheiden, im Vergleich zu den anderen Werten nur ein wenig abgenommen hat. Ein Vergleich der einzelnen Werte (**Tab.13**) zeigt, dass nur 254 von den ursprünglich 8283 Antworten für eine LD von 1 einmalige Antworten waren. Das bedeutet, dass die Wörter, die nur einmal als Antwort bei der Kognaten-Erkennung abgegeben wurden, sich in 96,03% der Fälle nicht durch die kürzeste orthographische Distanz von den Kognaten *f* unterscheiden. Daran erkennt man, dass die einmaligen Antworten, deren Verbindungen zum Kognaten *f* also nur schwierig nachzuvollziehen sind, sich durch hohe absolute Levenshtein-Distanzen von diesem unterscheiden.

53,76% der einmaligen Antworten sind Wörter, die sich durch zwei oder drei Editierdistanzen von den Kognaten *f* unterscheiden. Insgesamt sind 89,89% der einmaligen Antworten mehr als eine Editierdistanz vom abgefragten Kognaten *f* entfernt.

**Tabelle 13:** Vergleich der Anzahl Treffer pro absoluter Levenshtein-Distanz zwischen **Graf.17** und **Graf.19**, die Differenz zwischen diesen Werten; außerdem wird der prozentuale Anteil der jeweiligen Differenzen an der gesamten Anzahl einmaliger Antworten (%Anteil) angegeben

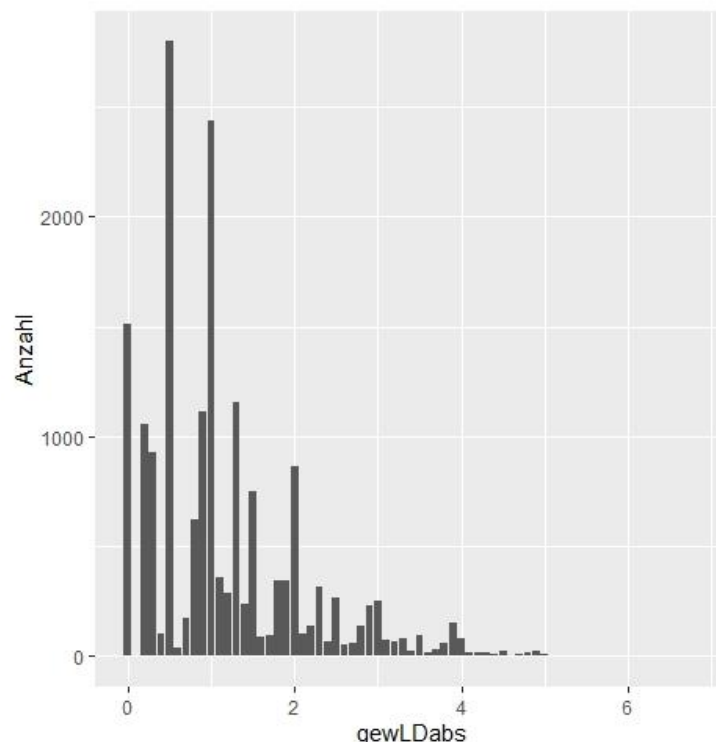
LDabs	Gesamtanzahl ( <b>Graf.17</b> )	Anzahl ohne LD 0 und einm. Treffer ( <b>Graf.19</b> )	Differenz	%Anteil
0	346	--	--	--
1	8283	8029	-254 (-3,07%)	10,11%
2	4964	4274	-690 (-13,90%)	27,46%
3	2348	1687	-661 (-28,15%)	26,30%
4	991	566	-425 (-42,89%)	16,91%
5	507	250	-257 (-50,69%)	10,23%
6	186	56	-130 (-69,89%)	5,17%
7	73	22	-51 (-69,86%)	2,03%
8	33	12	-21 (-63,64%)	0,84%
9	19	4	-15 (-78,95%)	0,60%
10	7	0	-7 (-100%)	0,28%
11	1	0	-1 (-100%)	0,04%
12	1	0	-1 (-100%)	0,04%

Der prozentuale Anteil der Gesamtheit der einmaligen Antworten (%Anteil) sinkt stetig von Distanz 2 zu Distanz 12. Die einmaligen Werte bilden jedoch einen immer größeren Teil der Antworten mit zunehmender absoluter Levenshtein-Distanz. So sind die 254 einmaligen Antworten, die eine Editier-Distanz von 1 haben, nur 3,07% der gesamten Antworten, die dieselbe LDabs besitzen. Für die absoluten Levenshtein-Distanzen 10, 11 und 12 sind 100% der Antworten nur einmal angegeben worden. Die Antworten, die von mindestens zwei

Personen ausgewählt wurden, sind eher jene mit niedrigen absoluten Levenshtein-Distanzen. Ohne die einmaligen Levenshtein-Distanzen unterscheiden sich nun 53,92% der Antworten durch eine absolute Levenshtein-Distanz von 1 von den Kognaten\_f.

Die absoluten gewichteten Levenshtein-Distanzen zeigen eine ähnliche Verteilung (**Graf.20**). Da die einzelnen phonologischen Distanzen sich jedoch durch Kommawerte unterscheiden können, ist natürlich eine höhere Anzahl verschiedener Werte möglich. Hier sind wieder sehr viele Antworten vorhanden, die eine Distanz von 0 aufweisen. Dies kann in diesem Fall aber zwei verschiedene Ursachen haben. Zum einen sind natürlich wieder jene Wörter in dieser Grafik repräsentiert, die mit dem Kognaten\_f übereinstimmen. Zum anderen sind gewichtete Levenshtein-Distanzen von 0 aber durchaus möglich, ohne dass die Wörter orthographisch übereinstimmen, da die Kognatenpaare ja aufgrund ihrer orthographischen Unterschiede ausgesucht wurden. Ein solches Kognatenpaar ist z.B. *kännen* (Kognat f) – *kennen* (Kognat dt). Für die phonologischen Distanzen besteht eine Korrelation von 0,458 (statistisch signifikant für ein Signifikanzniveau von 0,001,  $p$ -Wert: 1,43E-17).

**Grafik 20:** Anzahl der Antworten (Anzahl), die sich in einem gegebenen absoluten gewichteten LD-Wert (gewLDabs) vom Kognat f unterscheiden.



Die häufigsten absoluten phonologischen Levenshtein-Distanzen, die in den Antworten zu finden sind, sind 0,5 (15,74%) und 1 (13,72%). Da viele der Editier-Operationen, abhängig von

den betroffenen Phonemen und der Art der Operationen, durch kleine Werte repräsentiert sind, kann man auf Basis dieser Grafik keine eindeutigen Rückschlüsse auf die Anzahl der Operationen schließen, die die Antworten von den Kognaten\_f trennen. Eine gewichtete Levenshtein-Distanz von 1 ist entweder der Austausch eines Konsonanten durch einen anderen oder eine Kombination von zwei Editierschritten (mit mehreren möglichen Kombinationen; siehe Anhang für die vollständige Liste der gewichteten Distanzen). Eine Distanz von 0,5 ist entweder der Austausch eines Vokals durch einen anderen Vokal gleicher Länge, der sich im vereinfachten Vokaltrapez direkt neben dem ersten befindet oder wieder eine Kombination von zwei Austauschoperationen. Einige Beispiele für Antworten, die sich durch eine phonologische Levenshtein-Distanz von 0,5 oder 1 vom Kognaten\_f unterscheiden, sind in **Tabelle 14** zu finden.

**Tabelle 14:** Einige Beispiele für Antworten, die sich durch die am häufigsten vorkommenden phonologischen gewichteten Distanzen (0,5 und 1; gewLDabs) von den Kognaten\_f unterscheiden

Kognat_f	Kognat_dt	gewLDabs
(1) beven	beten	1
(2) gade	gase	0,5
(3) tegel	tigel	0,5

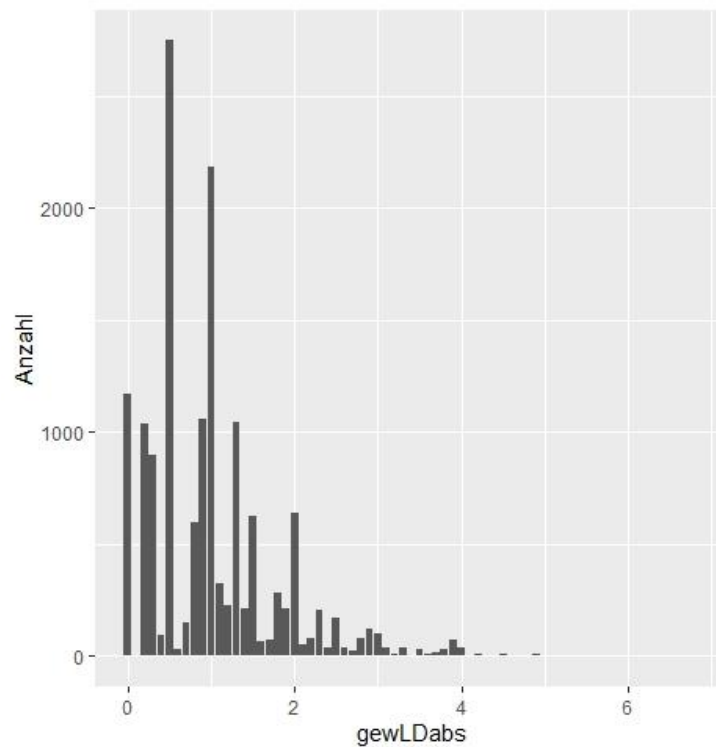
In Beispiel (1) wird ein Konsonant durch einen anderen ersetzt und diese Operation hat einfach eine phonologische Distanz von 1. In Beispiel (2) wird ebenfalls ein Konsonant durch einen anderen ersetzt, aber in diesem Fall hat die Operation einen Wert von 0,5, weil beide Konsonanten den selben Artikulationsort haben. Beispiel (3) zeigt einfach nur den Austausch eines Vokals durch einen anderen, der sich durch eine Position im Vokaltrapez unterscheidet.

Man erkennt insgesamt, dass auch bei der phonetischen Distanz vor allem kürzere Distanzen auftauchen. So unterscheiden sich 60,68% der Antworten in einer Editier-Distanz von den Kognaten\_f, die kleiner oder gleich 1 sind. Die Anzahl der Antworten nimmt, wie bereits zuvor die orthographischen Distanzen, exponentiell mit der Zunahme der phonologischen LD ab.

Für die absoluten gewichteten Levenshtein-Distanzen wurden in **Graf.21**, wie bereits zuvor für die einfache LD, die Einzelantworten entfernt. Es wurden ebenfalls die Antworten, die identisch zum Kognaten\_f sind, entfernt. Die Wörter, deren Orthographie sich von jener der Kognaten\_f unterscheidet, die aber eine phonologische Levenshtein-Distanz von 0 haben, wurden nicht entfernt. Die am häufigsten in den Antworten vertretenen Distanzen sind auch hier noch immer

0,5 und 1. Ohne die einmaligen Antworten sind nun aber 66,91% der Antworten Wörter, deren gewLDabs-Werte kleiner oder gleich 1 sind.

**Grafik 21:** Anzahl der Antworten (Anzahl), die sich in einem gegebenen absoluten gewichteten LD-Wert (gewLDabs) vom Kognat *f* unterscheiden, mit Ausnahme der den Kognaten *f* identischen und der nur einmal angegebenen Antworten.



Von den insgesamt 14 900 in dieser Grafik repräsentierten Werten haben 1170 (7,85%) eine absolute gewichtete Levenshtein-Distanz von 0. Das bedeutet, dass diese Wörter dieselbe (oder zumindest eine sehr ähnliche) Aussprache aufweisen wie die Kognaten *f*. Man erkennt außerdem an **Graf.20**, dass das Herausnehmen der einmaligen Antworten die Anzahl der Antworten mit einer phonologischen Levenshtein-Distanz von mehr als 2 stark abgenommen hat.

Die phonologischen und orthographischen Distanzen haben insgesamt eine sehr ähnliche Verteilung, bevor und nachdem die einzelnen Antworten herausgenommen wurden. Die Antworten der Testpersonen haben jedoch in nur 46,6% der Fälle (bzw. 53,92% ohne die einmaligen Antworten) eine absolute orthographische Levenshtein-Distanz von 1. Für die phonologische gewichtete Levenshtein-Distanz liegen die Distanzen, die kleiner oder gleich 1 sind, bei 60,68% (bzw. 66,91%). Die Antworten der Testpersonen weisen also häufiger eine größere phonologische Ähnlichkeit auf als eine orthographische. Die Korrelation ist bei der

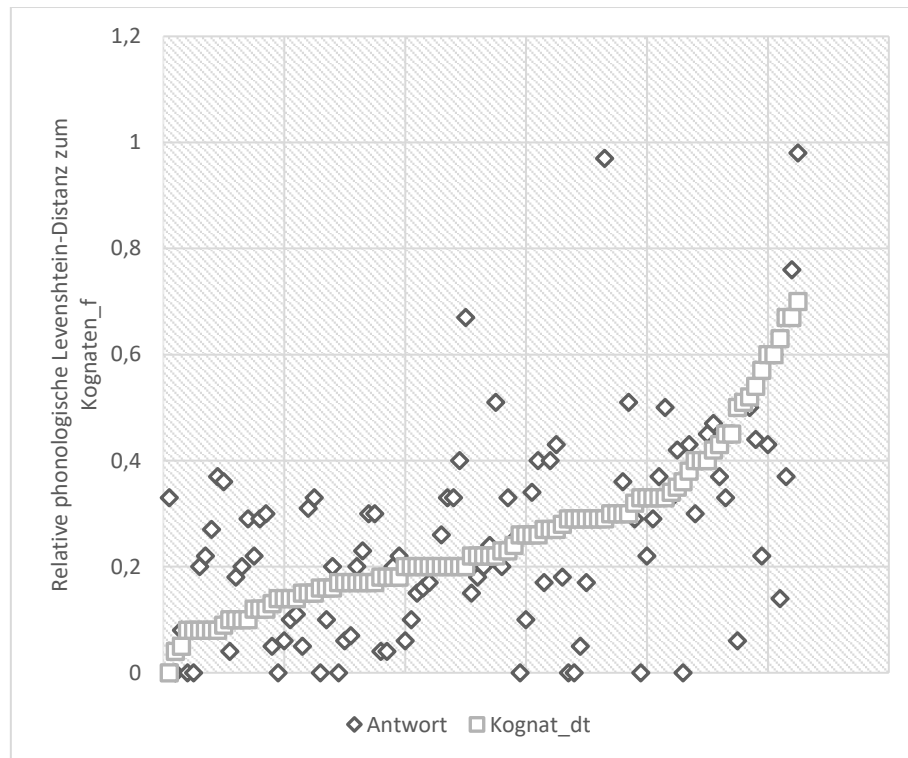
absoluten Levenshtein-Distanz jedoch größer (LDabs: 0,522, gewLDabs: 0,458). Das bedeutet, dass die einfachen Levenshtein-Distanzen zwischen den Antworten und den Kognaten\_f eher mit jenen der Kognatenpaare übereinstimmen, als dies für die phonologischen Distanzen der Fall ist. Dabei wird jedoch nicht berücksichtigt, dass die einfachen Levenshtein-Distanzen alle Operationen als gleich ansehen und so ist eine höhere Korrelation nicht indikativ für eine größere Ähnlichkeit zwischen den Antworten und den Kognaten\_f. Wenn ein <e> durch ein <i> ersetzt wird, dann ist dies eine wahrscheinlichere Operation als der Austausch eines <e> durch ein <k>. Dies wird jedoch nicht bei der einfachen Levenshtein-Distanz berücksichtigt.

#### 10.5.2 Nicht-Kognat\_dt-Antworten

Bei 105 von den 312 abgefragten Kognaten\_f, also etwas mehr als 30%, waren die häufigsten Antworten nicht mit den Kognaten\_dt identisch. Hier werden einige dieser Fälle genauer untersucht, um zu überprüfen, weshalb die Testpersonen die Kognaten\_dt zu Gunsten dieser anderen Antworten verwerfen – oder die Kognaten\_dt vielleicht nicht einmal in Betracht ziehen. Es wurden zuerst aber einige der oben bereits verwendeten Faktoren verwendet, um die Kognaten\_dt mit den Antworten der Testpersonen zu vergleichen.

Die relative phonologische Levenshtein-Distanz wurde verwendet, um zu überprüfen, ob die Antworten phonologisch näher an den Kognaten\_f liegen als die Kognaten\_dt. Dies hat ergeben, dass nur bei 52 von 105 Kognaten\_f, also in 49,52% der Fälle, die Antworten der Testpersonen, die nicht die Kognaten\_dt sind, eine kleinere gewichtete Levenshtein-Distanz zum Kognaten\_f besitzen als der Kognat\_dt. Anhand der Verteilung der Datenpunkte in **Graf.22** ist dies auch erkennbar. In dieser sind die phonologischen Levenshtein-Distanzen der Kognatenpaare nach aufsteigenden Werten sortiert, um den Vergleich mit den relativen phonologischen Distanzen zwischen den Kognaten\_f und den Antworten der Testpersonen zu erleichtern.

**Grafik 22:** Vergleich zwischen den relativen phonologischen Levenshtein-Distanzen der Wortpaare „Kognat\_dt - Kognat\_f“ und „Antwort - Kognat\_f“.

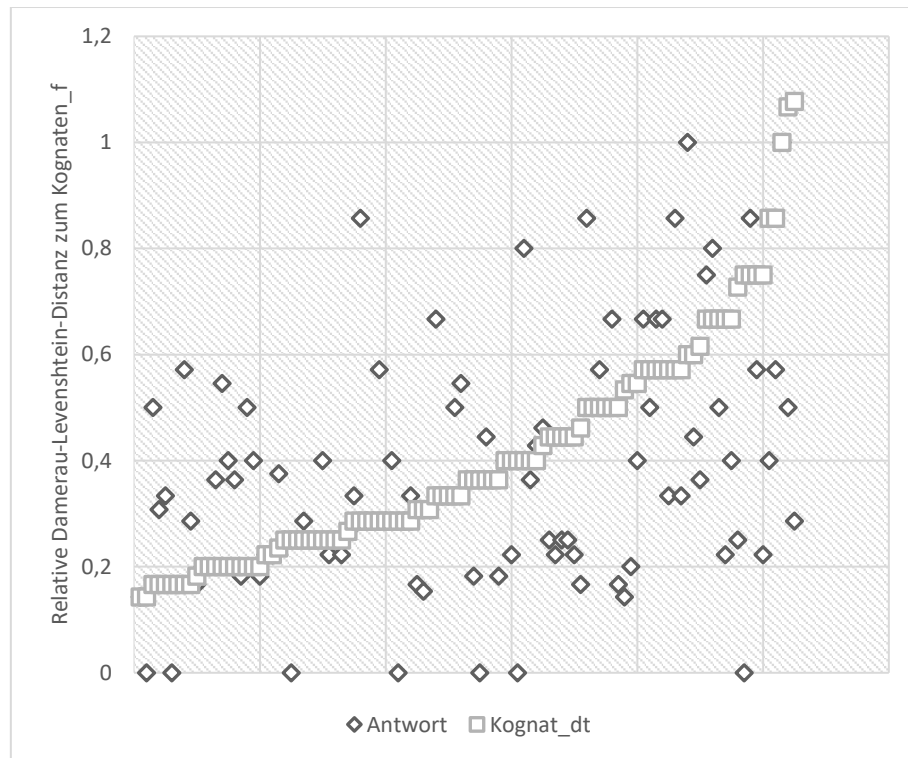


Die relativ gleichmäßige Verteilung der phonologischen Levenshtein-Distanzen zwischen den Antworten und Kognaten\_f ober- und unterhalb der Distanzen für die Kognatenpaare zeigt, dass die Antworten der Testpersonen nicht strikt höhere oder niedrigere Distanzen zum Kognaten\_f aufweisen als die Kognaten\_dt. Diese Antworten werden also nicht deshalb präferiert, weil diese den Kognaten\_f ähnlicher sehen als die richtigen Kognaten\_dt. Es besteht dennoch eine relativ hohe Korrelation zwischen den phonologischen Distanzen der Kognatenpaare und jenen zwischen den Kognaten\_f und den Antworten. Der Korrelationskoeffizient beträgt 0,494 und ist statistisch signifikant für ein Signifikanzniveau von 0,001 ( $p$ -Wert: 8,18E-05). Für die Kognaten\_f, die sehr ähnlich aussehende Kognaten\_dt haben, werden also häufig auch sehr ähnlich aussehende Antworten angegeben.

Für die orthographischen Distanzen (**Graf.23**) kann man ähnliche Schlüsse ziehen. Die Berechnung der relativen Damerau-Levenshtein-Distanzen für die Kognatenpaare und die Wortpaare „Kognat\_f – Antwort“ hat ergeben, dass nur 43 der von den Testpersonen abgegebenen Antworten eine kleinere orthographische Distanz aufweisen (41% der Antworten). Die anderen 59% der Antworten haben entweder eine größere (37 Antworten) oder eine identische (25 Antworten) Damerau-Levenshtein-Distanz zum Kognaten\_f.



**Grafik 22:** Vergleich zwischen den relativen phonologischen Levenshtein-Distanzen der Wortpaare Kognat\_dt - Kognat\_f und Antwort - Kognat\_f.



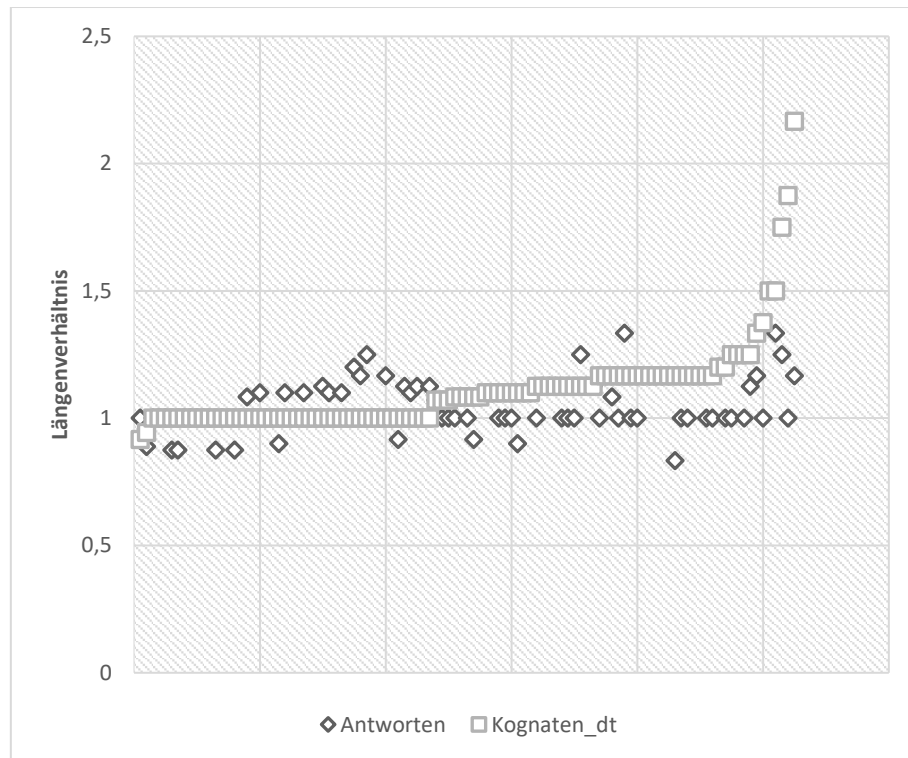
In diesem Fall ist die Korrelation zwischen den Distanzen der Kognatenpaare und den Distanzen zwischen den Kognaten\_f und den Antworten etwas schwächer, aber trotzdem noch immer sehr hoch: Korrelationskoeffizient 0,379, statistisch signifikant (Signifikanzniveau: 0,001,  $p$ -Wert: 6,56E-05). Die Korrelationskoeffizienten und die Grafik für die relativen einfachen Levenshtein-Distanzen sind fast identisch (Korrelationskoeffizient: 0,380, statistisch signifikant für ein Signifikanzniveau von 0,001). Die Werte der relativen LD wurden aufgrund ihrer Ähnlichkeit zu den Damerau-Levenshtein-Werten nicht graphisch dargestellt. Diese Ähnlichkeit der Korrelationskoeffizienten deutet darauf hin, dass Transpositionsoperationen entweder keinen Unterschied machen oder in diesen Antworten nur sehr selten vorkommen (oder beides).

Die Tatsache, dass ebendiese 106 Wörter anstelle der den Kognaten\_f entsprechenden Kognaten\_dt ausgewählt werden, kann aber definitiv nicht durch eine systematisch kürzere phonologische oder orthographische Distanz erklärt werden. Das bedeutet jedoch nicht, dass diese keinen Einfluss bei der Auswahl einzelner Antworten hat. Es ist davon auszugehen, dass mehrere Faktoren miteinander interagieren und dies dann zur Auswahl eines bestimmten Wortes als Antwort führt.

Einer dieser anderen Faktoren könnte die Frequenz der Kognaten\_dt und der Antworten sein. 62% (65) der Antworten, die häufiger als die jeweiligen Kognaten\_dt ausgewählt wurden, sind höherfrequenter Wörter im Korpus der Universität Leipzig (Wortschatz Leipzig). Wenn nun ein Wort aber nur 1,001 mal häufiger vorkommt als der Kognat\_dt, dann kann man wohl davon ausgehen, dass dieser Frequenzunterschied bei der Erkennung nicht sehr relevant sein sollte. Von den 65 Antworten, die höhere Frequenzen als die Kognaten\_dt aufweisen, sind 53 jedoch mindestens zweimal so häufig in den verwendeten Korpora vorhanden wie die Kognaten\_dt und 34 Antworten kommen sogar mindestens zehnmal häufiger vor. Die Kognaten\_dt sind nur in 29 Fällen mindestens doppelt so frequent und kommen nur in 15 Fällen mindestens zehnmal so häufig vor wie die häufigsten abgegebenen Antworten. Dass die Frequenz in diesem Fall eine Rolle spielt, ist durchaus plausibel, da diese Werte sehr asymmetrisch sind, d.h. zugunsten der Antworten verschoben sind, die nicht den Kognaten\_dt entsprechen.

Im ersten Analysedurchgang wurde überprüft, ob die Längenverhältnisse, d.h. die Länge des Kognaten\_dt geteilt durch die Länge des Kognaten\_f, der Kognatenpaare einen Einfluss auf die Erkennungsfrequenz haben. Die Hypothese war, dass man bei einer Korrelation zwischen dem Längenverhältnis und der Trefferquote beim Kognatentest sehen konnte, ob die Testpersonen eher Buchstaben an den Kognaten\_f anhängen oder Buchstaben wegnehmen, um den Kognaten\_dt zu bilden. Im ersten Analysedurchgang sind jedoch keine statistisch signifikanten Korrelationen erhalten worden. Hier wird nun überprüft, ob bei Kognatenpaaren, bei denen ein oder mehrere Buchstaben hinzugefügt werden müssen, um den Kognaten\_dt zu bilden, Antworten abgegeben wurden, die auch mehr Buchstaben enthalten als der Kognat\_f (und umgekehrt). Dies würde bedeuten, dass gewisse Kognaten das Hinzufügen von Buchstaben begünstigen. Es besteht in der Tat eine relativ hohe Korrelation (0,616, statistisch signifikant für ein Signifikanzniveau von 0,001,  $p$ -Wert: 5,57E-34) zwischen den Längenverhältnissen der Kognatenpaare und jenen der Wortpaare „Antwort – Kognat\_f“. Man erkennt in **Graf.23**, dass die Längenverhältnisse zwischen den Antworten und den Kognaten\_f nur sehr selten unter 1 sind. Wenn dieser Wert unter 1 liegt, dann bedeutet dies, dass die Antwort weniger Buchstaben enthält als der abgefragte Kognat\_f. Es zeigt, dass die Testpersonen nur sehr selten nach Lösungen suchen, die kürzer sind als der Kognat\_f. Man erkennt jedoch auch, dass insgesamt nur in sehr seltenen Fällen die Kognaten\_dt tatsächlich kürzer als die Kognaten\_f sind.

**Grafik 23:** Vergleich zwischen den Längenverhältnissen der Wortpaare „Kognat\_dt - Kognat\_f“ und dem gemittelten Wert der Wortpaare „Antwort - Kognat\_f“.



Es ist außerdem ersichtlich, dass die Längenverhältnisse zwischen den Antworten und den Kognaten\_f ab einem Wert von ungefähr 0,3 nicht weiter ansteigen, auch wenn die Längenverhältnisse der Kognatenpaare größer sind. Die Testpersonen sind anscheinend nicht bereit, zu viele Buchstaben an den Kognaten\_f anzuhängen, um ihre Antworten zu bilden.

## 10.6 Interpretation des Analysedurchgangs 2

Dieser zweite Analysedurchgang hat gezeigt, dass die Testpersonen durchaus versuchen, die Wörter als mögliche Kognaten\_dt auszuwählen, die eine gewisse Ähnlichkeit zum Kognaten\_f besitzen. Das Aussortieren einzelner Antworten hat gezeigt, dass die systematischen Antworten besonders nahe an den Kognaten\_f liegen und die meisten Wörter mit sehr großen phonologischen und orthographischen Distanzen personenspezifische Antworten waren, die über Faktoren ausgewählt wurden, die nichts mit der Ähnlichkeit zum Kognaten\_f zu tun haben.

Die phonologischen Distanzen zeigen, dass 2/3 der abgegebenen Antworten durch Operationen von den Kognaten\_f getrennt sind, die sehr kleine Unterschiede zwischen den Antworten und den Kognaten\_f repräsentieren. Da die phonologische Levenshtein-Distanz auf zutreffenden

Intuitionen der Testpersonen, was die Lautkorrespondenzen zwischen den anderen germanischen Sprachen und dem Deutschen betrifft, basiert, sind diese mehrheitlich niedrigen Distanzen ein Hinweis darauf, dass plausible Wörter als Kognaten\_dt ausgewählt wurden.

Die Tatsache, dass häufig als Antwort auf den Kognaten\_f einfach nur dieser angegeben wurde, kann zwei Gründe haben. Entweder lag es an fehlender Inspiration von Seiten der Testpersonen, oder sie dachten, dass der Kognat\_dt mit dem Kognaten\_f identisch ist. In diesem Fall glaubten die Testpersonen möglicherweise, dass es sich um ein deutsches Wort handelt, das ihnen unbekannt war. Dies ist natürlich nur Spekulation, da es keine Testpersonen gab, die sich dazu in einem Kommentar geäußert haben.

Die Analyse der Antworten, die häufiger als die richtigen Kognaten\_dt als Antwort auf den Kognatenerkennungstest angegeben wurden, zeigt, dass die phonologischen und orthographischen Distanzen durch die Distanzen zwischen den Kognaten\_und Kognaten\_dt beeinflusst werden: größere Distanzen der Wortpaare „Kognat\_dt – Kognat\_f“ führen zu größeren Distanzen im Wortpaar „Antwort – Kognat\_f“. Da die Distanzen in den Antworten, die anstelle der richtigen Kognaten\_dt angegeben wurden, jedoch nicht systematisch kleiner sind, sind die phonologische und orthographische Distanz nicht alleine für die Auswahl dieser anderen Wörter verantwortlich.

Sowohl die Frequenzen als auch die Längenverhältnisse scheinen die Auswahl der Antworten zu beeinflussen. Die Frequenzen der Antworten sind meistens bedeutend größer als die der Kognaten\_dt. Dies deutet darauf hin, dass diese Wörter ausgewählt wurden, weil sie den Testpersonen nicht nur als gute Entsprechungen für die Kognaten\_f erschienen, sondern auch häufig genug vorkommen. Dadurch werden die Wörter von den Testpersonen dann als wahrscheinlichere Kognat\_dt-Kandidaten angesehen. Der Vergleich der Längenverhältnisse zeigt vor allem, dass die Testpersonen sich bei der Wahl der Antworten nie zu weit von der Länge der Kognaten\_f entfernten. Es wurden außerdem nur in einem kleinen Teil der Fälle Antworten ausgewählt, die kürzer als der Kognat\_f sind.

Diese Faktoren zeigen deutlich, dass man die Erkennung von Kognaten nicht an einem einzigen Element festmachen kann und dass jeder Kognat\_f besondere Eigenschaften hat, die dessen Erkennung beeinflussen. Im nächsten Teil werden Faktoren untersucht, die teilweise sehr von den Kognaten\_f abhängen und dadurch eine verallgemeinernde Analyse der Faktoren erschweren.

## 11. Nicht-systematische Faktoren der Kognaten-Erkennung

Bei den bisher besprochenen Faktoren, wie z.B. den verschiedenen Levenshtein-Distanzen, wurde untersucht, ob diese einen systematischen Einfluss auf die Erkennung von Kognaten haben. Das bedeutet, dass diese unabhängig von den einzelnen Testitems einen Effekt auf die Kognaten-Erkennung im Allgemeinen haben. In diesem Teil geht es nun um Faktoren, die von den jeweiligen Kognaten\_f und von den Testpersonen abhängig sind. Drei Faktoren wurden identifiziert, die die Wahl der jeweiligen Antworten unterschiedlich stark beeinflussen: Kenntnisse des Englischen, semantische Assoziationen zu anderen deutschen Wörtern und Reime. Diese Faktoren werden an dieser Stelle anhand von konkreten Beispielen und stichprobenartig beschrieben, da eine statistische Analyse nicht möglich ist. Obwohl nur ein recht kleiner Anteil der Testpersonen Kommentare hinterlassen hat, kann man diese verwenden, um deren Gedankenwege nachzuvollziehen. Anhand der Kommentare einzelner Testpersonen kann man so die Validität der Vermutung, dass diese Faktoren tatsächlich eine Rolle spielen, überprüfen.

### 11.1 Das Englische

Obwohl sich keine Kognaten\_f, die über das Englische einfacher erkannt werden können, in den abgefragten Items befanden, sind häufig Wörter auf Basis einer scheinbaren Verbindung zum Englischen ausgesucht worden. In 62 Kommentaren wurde erwähnt, dass eine Antwort ausgewählt wurde, weil der Kognat\_f einem englischen Wort ähnlich sieht. Diese Kommentare stammen insgesamt von 36 verschiedenen Personen, was zeigt, dass diese Verbindungen häufig gemacht werden und sich nicht auf nur ein paar Personen beschränken. Natürlich ist davon auszugehen, dass sehr viele der Testpersonen, die keine Kommentare hinterlassen haben, auch ihre Antworten über ein englisches Lexem hergeleitet haben.

Häufig wird bei der Auswahl der Antworten auf der Basis englischer Lexeme völlig außer Acht gelassen, dass die deutsche Übersetzung dieses englischen Wortes dem Kognaten\_f überhaupt nicht ähnlich sieht. So wurde z.B. für das Wort *smal* das deutsche Wort *klein* als Antwort angegeben. Dazu schrieb die betreffende Testperson „small aus dem Englischen“. Interessanterweise wäre der Kognat\_dt *schmal* viel näher am Kognaten\_f dran gewesen, aber die Person hat sich für *klein* entschieden. *klein* ist die geläufigste Übersetzung für *small* und hat sich wohl deswegen gegenüber *schmal* durchgesetzt. Diese Assoziation von *smal* mit *klein* ist

der Testperson wohl sehr spontan eingefallen. Dies erkennt man daran, dass die Antwort in 3 Sekunden abgegeben wurde und dies der erste Versuch war. Insgesamt wurde die Antwort *klein* 15 mal für den Kognaten *small* angegeben (eine weitere Person hatte ebenfalls „Englisch“ im Kommentar stehen).

Es stellt sich die Frage, ob die Testpersonen über das Englische nach der richtigen Antwort suchen, da nicht sofort ein ähnliches deutsches Wort ersichtlich ist. Da die Testteilnehmenden keine direkte Verbindung zu einem deutschen Wort herstellen können, suchen sie vielleicht in ihrem mentalen Lexikon nach einem Wort einer LX. Dies trifft sicherlich häufig zu, aber das nächste Beispiel zeigt, dass nicht immer intensiv nach einem Wort der L1 gesucht wurde, bevor eine Verbindung zum Englischen geknüpft wurde.

Für den Kognaten\_f *deer* wurde ein weniger wahrscheinliches Wort anstelle des Kognaten\_dt ausgewählt. Der entsprechende Kognat\_dt ist *Tier*, aber eine Person hat das Wort *Wild* mit dem Kommentar „engl. deer“ angegeben. Abgesehen davon, dass wieder der plausiblere und richtige Kognat\_dt zugunsten der Übersetzung des englischen Wortes verworfen wurde, zeigt sich hier einer der Störfaktoren bei der Kognaten-Erkennung: *Falsche Freunde*. Das Wort *deer* ist tatsächlich Kognat mit *deer* und *Tier* und geht auf dasselbe germanische Wort *\*deuza* zurück (*Tier*, DWDS). Im Laufe der Zeit hat sich jedoch die Bedeutung des englischen Wortes geändert, sodass es heute nicht mehr *Tier* sondern (*Rot*-)*wild* bedeutet. Solche identischen oder zumindest sehr ähnlichen Wörter, die jedoch nicht dieselbe Bedeutung besitzen, werden als *false friends* oder *falsche Freunde* bezeichnet. Durch die Existenz solcher Wörter wird natürlich die alltägliche Anwendung von Kognaten-Erkennung erschwert. Selbst wenn der Ko- und Kontext dabei helfen, zu erkennen, dass der Kognat\_f nicht dieselbe Bedeutung wie der Kognat\_dt hat, bedeutet das noch nicht, dass die Person dann auch weiß, welche die Bedeutung des Kognaten\_f stattdessen ist. Diese *falschen Freunde* sind jedoch nicht so zahlreich, als dass sie die Verwendung von Kognaten-Erkennung als Basis für die Interkomprehension unmöglich machen.

Anhand einiger Beispiele für solche Übersetzungen von englischen Wörtern, die für Kognaten gehalten werden, kann man erkennen, wie vielfältig der Einfluss des Englischen auf die Erkennung von Kognaten ist (**Tab.15**).

**Tabelle 15:** Beispielliste von Kognaten\_f und dazugehörigen Antworten, die auf englischem Wortschatz basieren, mit den dazugehörigen Kommentaren und den Kognaten\_dt.

Kognat_f	Antwort	Kommentar	Kognat_dt
(1) sken	Haut	sken - skin - Haut (englisch)	schön
(2) scharp	scharf	englisch "sharp"	scharf
(3) kauh	kuh	durch die englische aussprache von cow [kau]	Kuh
(4) line	Linie	line bedeutet im Englischen Linie	Leine

In Beispiel (1) wurde einfach das englische Wort gewählt, weil es dem Kognaten\_f ähnlicher sieht. Die relative Levenshtein-Distanz beträgt 0,25 und die relative phonologische Distanz sogar nur 0,125. Das zweite Beispiel ist besonders interessant, da die Testperson tatsächlich den Kognaten\_dt gefunden hat. Dies ist jedoch über engl. *sharp* geschehen. Hier scheinen Kenntnisse des Englischen also als Übergang bei der Erkennung gedient zu haben. In Beispiel 3 wurde ebenfalls der richtige Kognat\_dt angegeben und hier hat die betreffende Person sogar explizit in ihren Kommentar geschrieben, dass die Erkennung auf phonologischem Wege stattgefunden hat. In Beispiel 2 kann der Kognat\_f auch über die Orthographie erkannt worden sein. Der Kognat\_f *line* in Beispiel 4 und die von der Testperson gegebene Antwort *Linie* sind sich genauso ähnlich wie der Kognat\_f und der Kognat\_dt (1 Einfügung). Die Testperson hat sich aber für das Wort *Linie* entschieden, weil es die am häufigsten verwendete Übersetzung des englischen Wortes *line* ist. Engl. *line* kann aber auch mit ‚Leine‘ übersetzt werden und wäre in diesem Fall dann der richtige Kognat\_dt. Dies kann zwei Gründe haben. Entweder hat sich die Person für *Linie* als Übersetzung entschieden, weil es die geläufigere Übersetzung für engl. *line* ist, oder die Person wusste nicht, dass *line* auch ‚Leine‘ bedeuten kann. Von allen hier genannten Beispielen ist *sken* das einzige englische Wort, das nicht mit den Kognaten\_f und Kognaten\_dt verwandt ist. Dies kommt jedoch in den Testergebnissen recht häufig vor (Tab.16).

**Tabelle 16:** Beispielliste von Kognaten\_f und dazugehörigen Antworten, die auf englischem Wortschatz basieren und nicht mit den Kognaten\_f etymologisch verwandt sind, mit den dazugehörigen Kommentaren und den Kognaten\_dt.

Kognat_f	Antwort	Kommentar	Kognat_dt
(5) ár	Pfeil	Englisch Arrow= Pfeil	Jahr
(6) pek	Schwein	ähnlich wie engl. pig -> Schwein	Pech
(7) steel	stahl	vgl zum engl.	Stiel
(8) smiden	lächeln	wegen englisch "smile"	schmeißen
(9) sin	Sünde	wegen englisch "sin"	Sinn

In solchen Fällen wäre das Finden des Kognaten\_dt natürlich reiner Zufall. Das ausgewählte englische Wort, das etymologisch nicht mit dem Kognaten\_f und Kognaten\_dt verwandt ist, muss diesen aber ähnlich genug sehen, um anhand des Kognaten\_f erkannt zu werden und es muss zusätzlich noch dieselbe Bedeutung haben. Die Beispiele in **Tabelle 16** zeigen außerdem, dass die Erkennung der Kognaten sowohl auf orthographischem als auch auf phonologischem Wege erfolgt.

Die Antwort *Pfeil* über engl. *arrow* für den Kognaten\_f *ár* (5) ist ein Beispiel für eine Assoziation, die sehr personenspezifisch ist. Diese Antwort wurde nur von einer einzigen Person angegeben. Das finden des richtigen Kognaten\_dt wurde vermutlich durch die „Abwesenheit“ des Buchstabens <j> erschwert. Dies in Kombination mit der sehr geringen Auswahl an deutschen Wörtern, die mit <ar> beginnen, Substantive sind und nicht zu viele Buchstaben enthalten, hat vermutlich dazu geführt, dass die Testperson sich für das englische Wort entschieden hat.

Die Lautform des Kognaten *pek* (6) ist dem englischen Wort *pig* hingegen recht nahe. Aufgrund der deutschen Auslautverhärtung wird <g> am Ende des Wortes wie ein /k/ ausgesprochen. Dadurch unterscheiden sich *pek* und *pig* phonologisch nur noch durch den Vokal in der Mitte des Wortes. Wie bereits von Möller & Zeevaert (2015:328) erklärt wurde, bevorzugen die Testpersonen die Übergänge von einem Vokal zum nächsten, wenn diese sich im vereinfachten Vokaltrapez nebeneinander befinden. Wenn man hier nun wieder das *interactive activation model* (IAM) betrachtet, dann ist die Wahl von *Schwein* als Kognat\_dt für *pek* durchaus nachvollziehbar. Da Wörter laut des IAM sowohl in ihrer graphischen als auch in ihrer phonologischen Form in ihrem Wortknoten gespeichert sind, wird hier vermutlich aufgrund der phonologischen Ähnlichkeit zwischen engl. *pig* und dem Kognaten\_f *pek* das englische Wort



aktiviert. Anhand dieses Beispiels kann man ebenfalls ein anderes Resultat aus Möller & Zeevaert (2015) erkennen: Testpersonen tendieren dazu, die Anfangsbuchstaben der Kognaten\_f nicht zu verändern. Eine naheliegende Antwort wäre das Wort *Speck* gewesen (selbst wenn der Wandel von /sp/ zu /p/ lauthistorisch nicht sehr plausibel ist). Diese Testperson hat sich jedoch dagegen entschieden und hat den „Umweg“ über das englische Wort gewählt. Und dies war kein Einzelfall: Für den Kognaten\_f wurde insgesamt 6 mal das Wort *Schwein* (und 1 mal *Ferkel*) angegeben. Die Antwort *Gipfel*, die wohl über das englische Wort *peak* gefunden wurde, war ebenfalls einmal unter den Resultaten.

Der Kognat\_f *steel* (7) kann auf verschiedene Arten zu einer Lösung führen. Der schriftliche Ausdruck ist ein Homograph zum englischen Wort *steel* („Stahl“). Dadurch ist es natürlich möglich, dass die Testpersonen das Wort *Stahl* als Kognat\_dt angeben. Die Aussprache des Kognaten\_f *steel* nach deutschen Lautgesetzmäßigkeiten liefert jedoch auch ein Wort, dessen Vokal sich nur durch eine einzige Position im Vokaltrapez von jenem im Wort *Stahl* unterscheidet. Der richtige Kognat\_dt weist dieselbe gewichtete phonologische Levenshtein-Distanz zum Kognaten\_f auf wie das Wort *Stahl*. Dennoch wird *Stahl* mehr als doppelt so häufig als Antwort angegeben als *Stiel* (*Stahl*: 28; *Stiehl*: 12). Diese übermäßige Präferenz für das Wort *Stahl* als Kognat\_dt ist sehr wahrscheinlich auf die Homographie des englischen Wortes und des Kognaten\_f zurückzuführen.

Die beiden Testpersonen, die für (8) *smiden* das deutsche Wort „lächeln“ angegeben haben, sind sehr wahrscheinlich nicht über die Phonologie zu diesem Schluss gekommen. Die Aussprache des <i> in als /aɪ/ ist im Deutschen vollkommen unüblich und die Verbindung zum englischen Wort *smile* kann so nur über den schriftlichen Ausdruck geschehen sein. Bei (9) *sin* ist dies ebenfalls plausibel, da das <s> vor Konsonant am Wortanfang im Deutschen immer stimmhaft ist, müsste der Kognat\_f von den Testpersonen [ˈsɪn] und nicht wie engl. [ˈzɪn] ausgesprochen werden. Wenn also als Antwort auf diesen Kognaten\_f *Sünde* angegeben wird, dann hat sich die Person vermutlich eher an der Schreibweise als an der Aussprache orientiert.

Anhand der hier beschriebenen Beispiele ist ersichtlich, dass das Englische sehr unterschiedliche Einflüsse auf die Kognaten-Erkennung haben kann. Das Englische führt manchmal dazu, dass die richtigen Kognaten\_dt gefunden werden, manchmal führt es zur Auswahl der falschen Antworten und manchmal wird zwar nicht der Kognat\_dt, aber doch eine plausible Alternative ausgewählt. Dies findet sicherlich nicht immer bewusst statt. In den hier genannten Fällen wurden Kommentare durch die Testpersonen abgegeben, die die Erschließung

der Kognaten\_dt durch das Englische zweifelsfrei belegen. In all jenen Fällen, in denen keine Kommentare abgegeben wurden, kann man sich jedoch nicht sicher sein, ob diese Verbindungen zu englischen Wörtern immer bewusst geknüpft wurden.

Von den 61 Datensätzen, in denen im Kommentar erwähnt wurde, dass das Wort über das Englische erkannt wurde (oder zumindest vermeintlich erkannt wurde), waren 42 Erstversuche. Dies zeigt, dass die Personen nicht erst nach Umwegen über das Englische suchen, nachdem sie einige Fehlversuche mit deutschen Wörtern hatten. Sie tendieren sofort dazu, das englische Wort dem deutschen Kognaten vorzuziehen. Dank der Instruktionen zu Beginn des Kognatentests mussten die Testpersonen eigentlich wissen, dass es für sie möglich ist, die Kognaten\_f direkt, aufgrund deren Ähnlichkeit zu einem Kognaten\_dt, mit diesem in Verbindung zu bringen. Dass sie es trotzdem über das Englische versuchen, könnte entweder an fehlender Aufmerksamkeit beim Lesen der Instruktionen liegen (möglicherweise wurde dort auch nicht alles gelesen) oder das englische Lexem hat sich einfach bei der Erkennung des Kognaten\_f durchgesetzt. Ersteres ist m.E. sehr unwahrscheinlich, da einzelne dieser Antworten teilweise von sehr vielen Personen angegeben wurden. So haben 22 Testpersonen *Haut* als potentiellen Kognaten\_dt für *sken* gegeben und es ist damit die häufigste Antwort für diesen Kognaten\_f. Wenn nur einzelne Personen solche Antworten abgegeben hätten, dann wäre das Ignorieren der Instruktionen eine plausible Erklärung. Bei so vielen Antworten, die auf englischen Lexemen basieren, ist es jedoch wahrscheinlicher, dass die Wortknoten für die betreffenden englischen Wörter im mentalen Lexikon so stark aktiviert wurden, dass alle anderen möglichen Antworten verdrängt wurden.

## 11.2 Semantische Assoziationen

Semantische Assoziationen sind den englischen Einflüssen ziemlich ähnlich. Der Unterschied besteht darin, dass die Personen hier nicht über ein Wort in einer anderen Sprache zu ihrer Antwort gelangen, sondern semantische Assoziationen zwischen dem Kognaten\_f und einem deutschen Wort bilden. Auf die semantischen Assoziationen wird hier nur kurz eingegangen, da dieses Thema bereits sehr detailliert in Möller & Zeevaert (2010, 2015) behandelt wurde.

Da nicht so häufig Kommentare zu den verschiedenen Antworten abgegeben wurden, ist es manchmal schwierig, die Gedankengänge der Testpersonen nachzuvollziehen. In anderen Fällen kann man diese Verbindungen zwischen den Antworten und den Kognaten\_f durchaus erklären. In **Tabelle 17** sind einige Beispiele für semantische Assoziationen enthalten.

**Tabelle 17:** Kognaten\_f und durch semantische Assoziationen gefundene Antworten, die nicht dem Kognaten\_dt entsprechen

Kognat_f	Antwort	Kognat_dt
(10) barsten	schreien/schimpfen	bersten
(11) ár	ohne/von	Jahr
(12) löschen	fangen	löschen

*schreien* und *schimpfen* für *bersten* hängt vermutlich damit zusammen, dass eine Person „vor Zorn bersten“ kann und dies sich häufig durch Schreien und Schimpfen äußert. Für die Beispiele (11) und (12) ist es nicht gelungen, den Gedankengang der Testpersonen nachzuvollziehen. Solche Assoziationen sind natürlich sehr hinderlich, wenn es um die richtige Erkennung von Kognaten\_f geht. Wie schon bei den Einflüssen durch das Englische wird auch hier die Ähnlichkeit zwischen dem Kognaten\_f und der ausgewählten Antwort vollkommen ignoriert.

### 11.3 Reime

Sehr häufig scheinen die Testpersonen die Antworten alleine deshalb auszuwählen, weil diese sich mit dem Kognaten\_f reimen. Dabei ist es natürlich manchmal schwierig zu sagen, ob eine Antwort deswegen ausgewählt wurde, weil die phonologische Distanz zwischen diesem und dem Kognaten\_f besonders niedrig ist oder weil die beiden Wörter sich reimen, da diese beiden Aspekte natürlich eng miteinander verbunden sind. Zwei Wörter können sich natürlich nur reimen, wenn diese eine gewisse Anzahl an Phonemen gemeinsam haben und diese sich am Wortende befinden. Aufgrund der gewichteten phonologischen Levenshtein-Distanz, die ja auf den Präferenzen von Testpersonen aufbaut, kann man jedoch sagen, wann eine bestimmte Antwort nicht aufgrund der phonologischen Ähnlichkeit zum Kognaten\_f ausgewählt wurde. Für den bereits zuvor besprochenen Kognaten\_f *pek* wurden zum Beispiel die folgenden Antworten abgegeben:

- (13) Deck
- (14) *Fleck*
- (15) *Gepäck*

Dass diese Wörter alleine aufgrund ihrer phonologischen oder graphemischen Ähnlichkeit zum Kognaten\_f ausgewählt wurden, scheint nicht sehr wahrscheinlich. Das Ersetzen des <p>, eines

stimmlosen bilabialen Plosivs, durch ein <d>, ein stimmhafter alveolarer Plosiv, wird durch die gewichtete phonologische Levenshtein-Distanz mit einem Preis von 1 verbunden. Den Kognat\_dt *Pech* hingegen trennt nur eine phonologische Distanz von 0,4 vom Kognaten\_f: Austausch eines Konsonanten durch einen anderen, der den selben Artikulationsort und dieselbe Phonation (stimmhaft oder stimmlos) aufweist (siehe Werte der gewichteten phonologischen LD im Anhang). Dies ist also kein Austausch, der typischerweise von Testpersonen präferiert wird. Die Lautformen von *pek* und *Fleck* unterscheiden sich sogar in zwei Editierschritten. *Gepäck* unterscheidet sich ebenfalls in zwei phonologischen Editierschritten vom Kognaten\_f *pek*, aber der große Unterschied zwischen den schriftlichen Ausdrücken sollte dieses Wort als potentiellen Kognaten\_dt eigentlich ausschließen. Dieses letzte Beispiel ist natürlich besonders markant wegen der fast vollständigen Nicht-Übereinstimmung der Buchstabenfolgen. Solche Worte, die keine semantische Verbindung zum Kognaten\_f haben und sich reimen, kommen immer wieder in den Testresultaten vor. Der Kognat\_f *skaft* (Kognat\_dt *Schaft*) lässt keinen Zweifel zu, dass die Reime eine Rolle bei der Auswahl der Antworten spielen, da diese nur sehr geringe Ähnlichkeit zum Kognaten\_f aufweisen: *Luft* und *Stift*.

#### 11.4 Fazit: Nicht-systematische Faktoren

Diese nicht-systematischen Faktoren zeigen sehr deutlich, dass es sehr schwierig ist, die Kognaten-Erkennung auf eine Handvoll von bestimmten Faktoren festzulegen. Alleine das Englische hat einen sehr starken Einfluss auf die Identifikation der Kognaten\_f. Man kann jedoch nicht einfach sagen, dass das Englische ein Störfaktor ist und als potentielle Hilfe bei der Erkennung der Kognaten\_f keinen Wert hat. Die im theoretischen Teil dieser Arbeit beschriebenen Unterschiede zwischen dem Englischen und den anderen germanischen Sprachen machen zwar die Interkomprehension anhand von Kognaten schwieriger, aber in einigen Fällen kann das Englische die Brücke zwischen der Muttersprache und einer anderen germanischen Sprache schlagen, wie es z.B. bei dem Kognaten *sharp* der Fall gewesen ist. Die englische Sprache sollte aber nur dann verwendet werden, wenn anhand des Deutschen alleine, die Kognaten nicht mehr identifiziert werden können. Ansonsten kann es häufiger vorkommen, dass englischer Wortschatz Kognaten\_f, die sich eigentlich nicht sehr von den entsprechenden Kognaten\_dt unterscheiden, schwieriger erkennbar macht. Dabei muss man jedoch darauf Acht geben, dass man durch das Erwähnen des Englischen vor dem Test nicht den umgekehrten Effekt hervorruft („Denken Sie auf keinen Fall an einen rosa Elefanten!“).

M.E. ist es aber vor allem wichtig, dass den Personen, die an solchen Tests teilnehmen, genau erklärt wird, was genau Kognaten sind. Das bedeutet, dass die Testpersonen verstehen müssen, dass die Wörter sich, ob nun auf graphemischer oder phonologischer Ebene, zu einem gewissen Grad ähneln müssen. Viele der von den Testpersonen abgegebenen Antworten zeigen, dass dieser Faktor häufig vollkommen missachtet wird. Im Falle eines solchen Tests, bei dem die Testpersonen wissen, dass sie einen bestimmten Kognaten erkennen können, müssten sie sich eigentlich bewusst sein, dass eine Antwort wie *Pfeil* für den Kognaten\_f *ár* eigentlich nicht möglich ist. Dies ist jedoch offensichtlich nicht so und solche Antworten, die keine Kognaten sein können, wurden, wie bereits oben deutlich gezeigt wurde, immer wieder als Antworten auf diesen Test abgegeben.

Die Antworten der Personen, die an diesem Test teilgenommen haben, zeigen deutlich, dass die semantische Komponente häufig der graphemisch-phonologischen vorgezogen wird. Es werden so starke Assoziationen aufgebaut, dass die Testpersonen, ob nun bewusst oder unbewusst, die Erkennung der Kognaten\_f an ebendiese Assoziationen binden, die teilweise sehr personenspezifisch und häufig von anderen nicht nachvollziehbar sind.

Natürlich können die meisten dieser Interferenzerscheinungen durch Assoziationen und englischen Wortschatz wahrscheinlich durch Kognaten-Erkennung im Ko- und Kontext ausgeschlossen werden. In einer realen Interkomprehensionssituation sollte keine Verwechslung zwischen Wörtern wie *Jahr* und *Pfeil* (Kognat\_f *ár*) oder *Pech* und *Schwein* (Kognat\_f *pek*) stattfinden. Aber selbst hier kann es vorkommen, dass sich der Kognat\_dt und eine alternative Antwort nicht nur ähnlich sehen, sondern sehr ähnliche Bedeutungen haben (Kognat\_f *deer*). Es ist natürlich niemals möglich, absolut alle Faktoren, die die richtige Erkennung von Kognaten negativ beeinflussen, zu kontrollieren und einige Fehler sind nur sehr schwer zu vermeiden.

## **12.Diskussion und Fazit**

Diese Arbeit hat versucht ein Licht auf die verschiedenen Arten von Faktoren zu werfen, die einen Einfluss auf die Erkennung von germanischen Kognaten haben und sie hat gezeigt, wie komplex diese sind. Die beiden orthographischen und die phonologische Levenshtein-Distanzen sind sehr ausschlaggebend für die letztendliche Erkennung der Kognaten\_f. Insbesondere die phonologische Komponente der Kognaten\_f scheint dabei eine sehr große

Rolle zu spielen. Die Ergebnisse haben immer wieder gezeigt, dass die Testpersonen über die (vermutete) Phonologie der Kognaten\_f versuchen, die Kognaten\_dt zu finden. Der sehr hohe Einfluss der phonologischen LD, der gemessen wurde, zeigt, dass diese Levenshtein-Variante sehr gut geeignet ist, um die Trefferquoten der Kognaten-Erkennung vorauszusagen.

Die Damerau-Levenshtein-Distanz hat teilweise bessere Resultate geliefert als die einfache Levenshtein-Distanz. Weitere Studien sollten der Frage nachgehen, ob die Damerau-Levenshtein-Distanz möglicherweise die einfache Levenshtein-Distanz als Standardmesswert für die Ähnlichkeit zwischen Kognaten\_f und Kognaten\_dt in der Kognaten-Erkennung ersetzen sollte. Es wäre außerdem interessant zu überprüfen, ob eine gewichtete phonologische Damerau-Levenshtein-Distanz ebenfalls bessere Resultate als die phonologische Levenshtein-Distanz liefern kann.

Trotz dieser sehr vielversprechenden Resultate für die Levenshtein-Varianten wurde dennoch gezeigt, dass mehrere Faktoren die Erkennung von Kognaten beeinflussen. Dazu gehören sowohl die systematischen Faktoren, wie z.B. die Wortfrequenz, als auch Faktoren, wie z.B. der Einfluss des Englischen, der die Kognaten-Erkennung entscheidend beeinflussen kann. Das Englische ist m.E. ein Faktor, der intensiver untersucht werden sollte. Aufgrund der Tatsache, dass das Englische in Europa (und der ganzen Welt) so weit verbreitet ist, gibt es nicht sehr viele Menschen, die keinerlei Kenntnisse des Englischen haben. Dadurch ist das Risiko natürlich groß, dass englischer Wortschatz bei der Interkomprehension interferiert und die Kommunikation erschwert. Durch eine Sensibilisierung der Testpersonen für den besonderen Status des Englischen kann meiner Meinung nach nicht nur die Kognaten-Erkennung, sondern auch die Interkomprehension im Allgemeinen erleichtert werden.

### 13.Literaturverzeichnis

- Arntz, R. (1997): "Passive Mehrsprachigkeit – eine Chance für die "kleinen" Sprachen Europas". In: Ammon, Ulrich / Mattheier, Klaus J. / Nelde, Peter H. (Hrsg.): *Einsprachigkeit ist heilbar - Überlegungen zur neuen Mehrsprachigkeit Europas*. Tübingen: 166–183.
- Balota, David et al. (2008): "Moving beyond Coltheart's *N*: A new measure of orthographic similarity", in: *Psychonomic bulletin and review* Band 15(5); S.971-979.
- Berthele, R. / Vanhove, J. (2017): "Interactions between formal distance and participant-related variables in receptive multilingualism", in: *Interactions in Receptive Multilingualism* Band 55(1); S.23-40.
- Dewaele, Jean-Marc (2017): "Why the Dichotomy 'L1 Versus LX User' is Better than 'Native Versus Non-native Speaker'", in: *Applied Linguistics* Band 39; S.236-240.
- Elspaß, S. / Möller, R. (2015): *dtv-Atlas Deutsche Sprache*, München.
- Gooskens, C. et al. (2015): "Mutual intelligibility of Dutch-German cognates by children: The devil is in the detail", in: *Linguistics* Band 53(2); S.255-283.
- Gooskens, C. / Swarte, F. (2017): „Linguistic and extra-linguistic predictors of mutual intelligibility between Germanic languages“, in: *Nordic Journal of Linguistics* Band 40(2); S.123-147.
- Heeringa, W. (2014): "Modeling Intelligibility of Written Germanic Languages: Do We Need to Distinguish Between Orthographic Stem and Affix Variation", in: *Journal of Germanic Linguistics* Band 26(4); S.361-394.
- Keipert, H. (2009): „Slavische Interkomprehension. Eine Einführung. (narr studienbücher)“ - Rezension, in: *Zeitschrift für Slavische Philologie*, Band 66(1); S.234-241.
- Lerner et al. (2014): "What can we learn from learning models about sensitivity to letter-order in visual word recognition?", in: *Journal of Memory and Language* Band 77; S.40-58.
- Mader, M. (2008): *Lateinische Wortkunde für Alt- und Neusprachler: Der lateinische Grundwortschatz im Italienischen, Spanischen, Französischen und Englischen*, Stuttgart.
- Marx, N (2012): "Reading across the Germanic languages: Is equal access just wishful thinking?", in: *International Journal of Bilingualism* Band 16(4); S.467-483.
- McClelland, James / Rumelhart, David E. (1981): "An interactive activation model of context effects in letter perception: I. An account of basic findings", in: *Psychological review* Band 88(5); S.375-407.
- McClelland, James et al. (2014): „Interactive Activation and Mutual Constraint Satisfaction in Perception and Cognition“, in: *Cognitive Science* Band 38; S.1139-1189.
- Möller, R. / Zeevaert, L. (2010): "‘Da denke ich spontan an *Tafel*‘- Zur Worterkennung in verwandten germanischen Sprachen“, in: *Zeitschrift für Fremdsprachenforschung* Band 21(2); S.217-248.
- Möller, R. (2011): "Wann sind Kognaten erkennbar? Ähnlichkeit und synchrone Transparenz von Kognatenbeziehungen in der germanischen Interkomprehension“, in: *Linguistik online* Band 46(2); S.79-101.

- Möller, R. / Zeevaert, L. (2015): „Investigating word recognition in intercomprehension: Methods and findings”, in: *Linguistics* Band 53(2); S.313-352.
- Riionheimo, H. (2017): “Introduction: Receptive multilingualism”, in: *Nordic Journal of Linguistics* Band 40(2); S.117-121.
- Schepens, J. (2012): “Distribution of cognates in Europe as based on Levenshtein distance”, in: *Bilingualism: Language and Cognition* Band 15(1); S.157-166.
- Serva, M. / Petroni, F. (2008): “Indo-European languages tree by Levenshtein distance”, in: *A Letters Journal Exploring the Frontiers of Physics* Band 81; S.1-5.
- Stedje, A. (2007): *Deutsche Sprache gestern und heute*, Paderborn.
- Zeevaert, L. / Thijs, J. D. ten (2007): “Introduction”, in: J.D. ten Thijs / L. Zeevaert (Hrsg.): *Receptive Multilingualism. Linguistic analyses, language policies and didactic concepts*, S.1-25.

### Verwendete Programme

R, in Kombination mit den Ergänzungen *vwr* (*visual word recognition*) und *ggplot2* (zum Zeichnen einiger Grafiken)

### Verwendete Korpora

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995): The CELEX lexical database (release 2) [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Wortschatz Universität Leipzig:

URL:[https://corpora.uni-leipzig.de/de?corpusId=deu\\_newscrawl\\_2011](https://corpora.uni-leipzig.de/de?corpusId=deu_newscrawl_2011)  
[Letztes Abrufdatum: 15.07.2020].



## 14. Anhang

A. Werte zur Berechnung der gewichteten phonologischen Levenshtein-Distanz: Diese Erklärungen stammen aus dem für die automatische Berechnung verwendeten VBA-Skript von Robert Möller.

- Pro Segment normal 1 (gleich f. Einfügung, kompl. Tausch und Tilgung)
  - außer: Tausch Kons. - Vokal: => 2; Vokal - Vokal: => 0.9
- Tausch bei gleich ausgesprochenen Graphemen (ä/e usw): 0 (rein phonetische Rechnung!)
- Konsonanten:
  - normal: 1
  - verbilligt:
    - Obstruenten mit nur gl. Art.-Ort: 0.5
    - sch-s und sch-ch: auch 0.5
    - Obstruenten mit gl. Art.-Ort PLUS gl. Art.-Art (weite Def., s.u.) (= nur Stimmhaftigkeit versch.): 0.2
    - Obstruenten mit gl. Art.-Ort PLUS gl. Phonation (= nur Art-Art versch.): 0.3
    - Affrikaten: mit stl. Plos./Frik 0.3 - sth. Plos./Frik 0.5
    - Nasale mit versch. Art.-Ort: 0.5
- Vokale:
  - Tausch normal: 0.9
  - Reduktionen:
    - nur 1 Stufe im vereinfachten Vokaldreieck : 0.5 (auch Rundung/Ent-)
    - Umlaut: 0.3
    - i/y und ü/y: #
    - e/ä: 0
    - äu/eu: 0
    - Längenunterschied: 0.3
- Mono-Diphthong: wenn Monophth. graphisch Teil d. Diphth.s oder in der Mitte: 0.5
- Einfügung / Tilgung:
  - normal: 1
  - Schwa: 0.2
  - #d nach n oder l: 0.6

## B. Fragebogen am Ende des Kognatenerkennungstests

1. Ist Deutsch Ihre Muttersprache ?  
☐ ja ☐ nein, ich habe aber  Jahre Deutsch gelernt
2. Beherrschen Sie neben Hochdeutsch/Umgangssprache einen deutschen Dialekt ?  
☐ ja ☐ ein bisschen ☐ nein
3. Welche der folgenden Sprachen/Dialekte haben Sie einigermaßen zu verstehen gelernt (im Rahmen von Unterricht oder durch häufigeren Kontakt damit):  
☐ Englisch  
☐ Schwedisch  
☐ Norwegisch  
☐ Dänisch  
☐ Isländisch  
☐ Niederländisch  
☐ Luxemburgisch  
☐ Niederdeutsch / nieder rheinisches Platt  
☐ Kölsch / rheinisches Platt  
☐ Schweizer Dialekt  
☐ anderen mitteldeutschen oder süddeutschen/österreichischen Dialekt  
 - Fällt es Ihnen normalerweise eher leicht, fremde Dialekte zu verstehen ?  
☐ ja ☐ eher nein
4. Sind Sie  
☐ Schüler/in  
☐ Student/in der Germanistik (oder ehemaliger/r ...)  
☐ Student/in einer anderen Sprache / der Allgemeinen Sprachwissenschaft (oder ehemaliger/r ...)  
☐ keins davon, aber in Beruf oder Freizeit häufig mit Texten beschäftigt  
☐ keins davon
5. Falls Sie Student/in der Germanistik oder einer anderen Sprache bzw. der Allgemeinen Sprachwissenschaft sind oder waren:  
 - Zu welchen der folgenden Themen haben Sie Veranstaltungen besucht (einschließlich laufendes Semester, wenn schon mindestens zur Hälfte vorbei):  
☐ Sprachgeschichte (Deutsch) - falls auch Lautwandel besprochen wurde  
☐ Sprachgeschichte (andere germanische Sprache) - falls auch Lautwandel besprochen wurde  
☐ Sprachgeschichte (nichtgermanische Sprache) - falls auch Lautwandel besprochen wurde  
☐ Phonetik/Phonologie allgemein  
☐ Althochdeutsch/Mittelhochdeutsch (Texte)  
☐ ältere Texte (andere germanische Sprache)  
☐ ältere Texte (nichtgermanische Sprache)  
☐ Dialektologie (Deutsch)  
☐ Dialektologie (andere Sprache)  
 - Haben Sie sich intensiver mit einem dieser Themen befasst ?  
☐ ja ☐ eher nein

## C. Kognatenpaare, die im Test verwendet wurden

Kognat_f	Kognat_dt	Kognat_f	Kognat_dt	Kognat_f	Kognat_dt
----------	-----------	----------	-----------	----------	-----------

akse	Achse	kleppe	Klappe	skudden	schütten
aksel	Achsel	kleppel	Klöppel	skuld	Schuld
alg	Alge	kletsen	klatschen	slag	Schlag
ár	Jahr	klitte	Klette	slange	Schlange
as	Asche	kneden	kneten	slapp	schlapp
barsten	bersten	kogel	Kugel	slät	schlecht
batter	bitter	korf	Korb	smal	schmal
ben	Bein	kort	kurz	smiden	schmeißen
berk	Birke	koss	Kuss	smieten	schmeißen
bett	Biss	kraan	Kran	sminke	Schminke
beven	beben	kraven	Kragen	smoren	schmoren
bieden	bieten	kreft	Kraft	söm	Saum
blad	Blatt	krom	krumm	spel	Spiel
blie	Blei	kruid	Kraut	spett	Specht
bok	Bock	kruit	Kraut	spied	Spieß
boll	Ball	kussen	Kissen	spoken	spuken
boom	Baum	kwaal	Qual	staf	Stab
boren	bohren	laden	lassen	stald	Stall
braden	braten	lam	lahm	ställen	stellen
braken	brechen	lan	Land	stauhl	Stuhl
brännen	brennen	läschen	löschen	staute	Stute
bratzel	Bretzel	laven	laben	steel	Stiel
breckelen	bröckeln	leed	Leid	steken	stechen
bred	breit	leer	Leder	stelen	stehlen
breed	breit	leggen	legen	stem	Stimme
brengen	bringen	leiden	leiten	sterk	stark
bruken	brauchen	lessen	löschen	stigen	steigen
bruse	Brause	liggen	liegen	stillen	stellen
buik	Bauch	lim	Leim	stjern	Stern
buk	Bauch	line	Leine	stof	Staub
dak	Dach	ljocht	Licht	stom	stumm
dal	Tal	long	Lunge	stoof	Staub
dauchen	tauchen	looch	Lauch	stoten	stoßen
deck	dick	loog	Lauge	stralen	strahlen
deeg	Teig	löv	Laub	stram	stramm
deig	Teig	loven	loben	sträng	streng
del	Teil	lus	Laus	streek	Strich
dief	Dieb	lüven	lügen	streid	Streit
dier	Tier	mand	Mann	streng	Strang
dierfen	dürfen	mars	Marsch	streven	streben
disch	Tisch	mäten	messen	struik	Strauch
djur	Tier	maven	Magen	sugen	saugen
dom	dumm	mees	Meise	suigen	saugen
dorp	Dorf	menen	meinen	suisen	sausen
draven	traben	mens	Mensch	suken	suchen

druk	Druck	merke	Marke	supen	saufen
duch	Tuch	misse	Messe	susen	sausen
dugen	taugen	muil	Maul	svans	Schwanz
dun	dünn	muis	Maus	svår	schwer
erf	Erbe	murer	Maurer	swaar	schwer
erven	erben	mus	Maus	swart	schwarz
fald	Fall	nat	Nacht	swerm	Schwarm
fast	fest	naven	nagen	täcke	Decke
finen	finden	nevel	Nebel	tahlen	zahlen
finster	Fenster	nicht	Nichte	tam	zahn
fjeder	Feder	njur	Niere	tand	Zahn
fjern	fern	nod	Not	tange	Zange
fles	Flasche	not	Note	tas	Tasche
fletten	flechten	oge	Auge	tass	Tatze
flieden	fließen	orm	Wurm	tegel	Ziegel
flier	Flieder	osse	Ochse	teken	Zeichen
flieven	fliegen	paren	paaren	tenken	denken
floom	Flaum	pek	Pech	ticht	dicht
fris	frisch	pensel	Pinsel	tistel	Distel
gade	Gasse	pols	Puls	tjener	Diener
gebit	Gebiss	pronk	Prunk	tjock	dick
geit	Geiß	pukkel	Pickel	toll	Zoll
genieten	genießen	raden	raten	torn	Dorn
gieden	gießen	reiden	reiten	tovern	zaubern
gieten	gießen	reiken	reichen	trappe	Treppe
glad	glatt	ren	rein	trüf	trüb
glud	Glut	renne	Rinne	trycken	drücken
greipen	greifen	rett	Recht	tunge	Zunge
gripen	greifen	riep	reif	tunn	dünn
grof	grob	riip	reif	tvång	Zwang
had	Hass	ropen	rufen	twilling	Zwilling
hamel	Hammel	rotte	Ratte	tydlich	deutlich
hamer	Hammer	roven	rauben	und	wund
han	Hand	rusten	rasten	under	Wunder
hark	Harke	saad	Saat	ung	jung
hauhn	Huhn	salig	selig	urn	Urne
hausten	husten	sänden	senden	vägen	wiegen
håven	heben	såven	sägen	vare	Ware
håxe	Hexe	schap	Schaf	växel	Wechsel
hede	Heide	scharp	scharf	veder	Feder
heem	Heim	schau	Schuh	vee	Vieh
heeß	heiß	scheef	schief	veel	viel
heler	Hehler	scherm	Schirm	veksel	Wechsel
herfst	Herbst	schrick	Schreck	verk	Werk
horen	hören	schuim	Schaum	vild	wild

hosten	husten	simmer	Sommer	vind	Wind
huid	Haut	sin	Sinn	vinken	winken
hun	Hund	sind	Sinn	wang	Wange
is	Eis	sjuden	sieden	wassen	wachsen
kande	Kanne	skäckig	scheckig	wegen	wiegen
känner	kennen	skaffen	schaffen	wenen	weinen
kärn	Kern	skaft	Schaft	wenken	winken
kauh	Kuh	skam	Scham	werpen	werfen
kaup	Kauf	sken	Schein	wierzen	würzen
kelder	Keller	skepp	Schiff	wip	Wippe
kelk	Kelch	skev	schief	wochs	Wachs
keren	kehren	skön	schön	wogn	Wagen
kim	Keim	skräck	Schreck	woll	Wall
kinnen	können	skreck	Schreck	yrken	wirken