EUROPHRAS · 2012 MARIBOR

ZORA ◦ 97

Phraseologie im Wörterbuch und Korpus

Phraseology in Dictionaries and Corpora

Phraseologie *im* Wörterbuch, Phraseologie *im* Korpus – das klingt weniger komplex als es de facto ist. Denn in Wirklichkeit und bei näherer Betrachtung ist die Angelegenheit vielschichtiger, und zwar nicht nur, wenn man den Begriff der Phraseologie im Sinne einer wissenschaftlichen Disziplin, also im Sinne von Phraseologie*forschung*, sondern auch und gerade dann, wenn man ihn als auf den Objektbereich ‚Phraseologie' bezogen versteht. Denn es ist ja keineswegs so, dass die Phraseologie quasi von sich aus in einem Wörterbuch bzw. Korpus enthalten ist und damit gewissermaßen „einfach so" für weitere Zwecke – angefangen von einfachem Interesse für den privaten Gebrauch über Lehr- und Lehrzwecke bis hin zu phraseologischen Forschungen – zur Verfügung steht. Vielmehr unterscheiden sich Wörterbuch und Korpus (genauer gesagt: verschiedene Arten von Wörterbüchern und Korpora) in dieser und verschiedener anderer Hinsicht voneinander und stehen zudem möglicherweise in komplexen Wechselbeziehungen miteinander.

Die Beiträge gehen auf die traditionelle zweijährige Tagung der Europäischen Gesellschaft für Phraseologie (EUROPHRAS) zurück, die in der Zeit vom 27. 8. 2012 bis 31. 8. 2012 an der Universität Maribor stattfand.

ZORA
97

Phraseologie im Wörterbuch und Korpus
Phraseology in Dictionaries and Corpora

ZORA
97

Phraseologie im Wörterbuch und Korpus
Phraseology in Dictionaries and Corpora

Mit finazieller Unterstützung der Slowenischen Forschungsagentur (ARRS)
und der Europäischen Gesellschaft für Phraseologie (EUROPHRAS)

With financial support from the Slovenian Research Agency (ARRS)
and the European Society of Phraseology (EUROPHRAS)

ZORA
97

Phraseologie im Wörterbuch und Korpus

Phraseology in Dictionaries and Corpora

Herausgeber / Editors
Vida Jesenšek
Peter Grzybek

Maribor, Bielsko-Biała, Budapest, Kansas, Praha
2014

ZORA
97

Phraseologie im Wörterbuch und Korpus
Phraseology in Dictionaries and Corpora

Herausgeber/Published by
Mednarodna založba Oddelka za slovanske jezike in
književnosti, Filozofska fakulteta, Univerza v Mariboru

Verantwortlich/For the Publisher
Marko Jesenšek

Titelbild/Cover image
Arhiv Europhras 2012

Design
Grafični atelje Visočnik

Auflage/Print Run
200

Druck/Print
Tiskarna Saje d.o.o.

CIP – Kataložni zapis o publikaciji
Univerzitetna knjižnica Maribor

81'373.7(082)

PHRASEOLOGIE im Wörterbuch und Korpus = Phraseology in dictionaries and corpora / Herausgeber, editors Vida Jesenšek, Peter Grzybek. – V Mariboru : Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta, 2014. – (Mednarodna knjižna zbirka Zora ; 97)

ISBN 978-961-6930-05-5
1. Jesenšek, Vida
COBISS.SI-ID 80434177

Inhaltsverzeichnis / Table of Contents

Phraseologie im Wörterbuch und Korpus. Einführende Bemerkungen

Phraseologie *im* Wörterbuch, Phraseologie *im* Korpus – das klingt weniger komplex als es de facto ist. Denn in Wirklichkeit und bei näherer Betrachtung ist die Angelegenheit vielschichtiger, und zwar nicht nur, wenn man den Begriff der Phraseologie im Sinne einer wissenschaftlichen Disziplin, also im Sinne von Phraseologie*forschung*, sondern auch und gerade dann, wenn man ihn als auf den Objektbereich 'Phraseologie' – also als Menge phraseologischer Einheiten, die in einem definierten Sprachmaterial oder in einer Sprache als Ganzes vorkommen – bezogen versteht. Denn es ist ja keineswegs so, dass die Phraseologie quasi von sich aus in einem Wörterbuch bzw. Korpus enthalten ist und damit gewissermaßen „einfach so" für weitere Zwecke – angefangen von einfachem Interesse für den privaten Gebrauch über Lehr- und Lernzwecke bis hin zu phraseologischen Forschungen – zur Verfügung steht. Vielmehr unterscheiden sich Wörterbuch und Korpus (genauer gesagt: verschiedene Arten von Wörterbüchern und Korpora) in dieser und verschiedener anderer Hinsicht voneinander und stehen zudem möglicherweise in komplexen Wechselbeziehungen miteinander. Vor allem aber ist phraseologisches Material nicht einfach so „gegeben" – weder in einem Korpus noch in einem Wörterbuch.

Auch wenn eine Reihe dieser Bemerkungen für den Großteil eines phraseologisch ausgerichteten Publikums mehr oder weniger selbstevident sein mögen, scheint es in Anbetracht der Vielfalt der hier vorgeschlagenen und eingeschlagenen Wege dennoch nicht unangebracht, die angesprochene Problematik in allgemeinster Weise anzusprechen, um die einzelnen Beiträge in einen solchen allgemeinen Rahmen einordnen zu können.

Wenn wir bei einem Korpus davon ausgehen, dass es sich um ein Textkorpus handelt, so wird der Begriff *Korpus* in diesem Sinne in der Regel als eine mehr oder weniger systematische Zusammenstellung von sprachlichem Material verstanden, die dazu dient, empirische Beobachtungen zu machen, d. h. spezifische linguistische Daten zu erheben. Diese können sich entweder im Sinne von single case studies auf individuelle oder aber auf eine bestimmte Menge (in unserem Fall: phraseologischer) Einheiten beziehen. Jedenfalls sollen in der Regel auf der Basis des betreffenden Materials sodann Aussagen über das Vorkommen bestimmter Elemente getroffen oder auch verallgemeinernde Aussagen angestrebt werden, die mitunter über die Gültigkeit des untersuchten Korpus hinausgehen sollen.

Im einfachsten Fall haben wir es dabei mit einer symptomatischen Beschreibung zu tun: etwas kommt im Untersuchungsmaterial vor, oder auch nicht – im Grunde genommen ein einfaches binäres, dichotomes Beschreiben. In weiterer Folge führt die Angabe von Vorkommenshäufigkeiten zur Kategorie der Quantität bzw. Gradualität, insofern beschrieben wird, wie oft etwas (mehr oder weniger) vorkommt. Dies kann, muss aber nicht über den Status des Symptomatischen hinausgehen, nicht zuletzt in Abhängigkeit davon, ob die Häufigkeiten auf bestimmte Art und Weise (in Bezug zu einem Gesamtvorkommen oder dem Vorkommen im Hinblick auf vergleichbare Daten) relativiert werden. Wenn dies der Fall ist und dann Aussagen über Vorkommen bzw. Vorkommenshäufigkeiten getroffen werden, haben wir es freilich immer noch mit einem deskriptiven Vorgehen zu tun – und da muss man sich im Grunde nach wie vor kaum Gedanken über die Beschaffenheit des Korpus machen: es ist, wie es ist, und die Aussagen betreffen nicht mehr und nicht weniger als die zur Verfügung stehenden und analysierten Daten, gegebenenfalls eben in Relation zu einer Gesamtdatenmenge oder zu anderen Daten. Will man hingegen darüber hinausgehende Schlussfolgerungen ziehen (oder, zunächst einmal, vorsichtiger und richtiger gesagt: Vermutungen in dieser Richtung anstellen bzw. diesbezügliche Hypothesen aufstellen), die über den begrenzten Objektbereich der beobachteten Daten hinausgehen, geht man üblicherweise davon aus, dass das Korpus so etwas wie eine „repräsentative" Stichprobe für eine (angenommene) Grundgesamtheit oder gar für eine Sprache in ihrer Gesamtheit darstellt. Gerade diese letzte Annahme ist jedoch aus theoretischer Sicht kaum haltbar, denn so etwas wie „die" Sprache – ein in sich heterogenes, dynamisches und kontinuierlich evoluierendes System – als eine geschlossene Menge gibt es nicht: Weder ist ‘Sprache’ die Summe aller jemals produzierten noch aller möglicherweise jemals produzierten Texte – wenn, dann ist ‘Sprache’ nur als abstraktes Konstrukt greifbar, basierend auf sprachlichen Beobachtungen und Verallgemeinerungen, und damit nur als abstraktes System.

Statt dessen kann man freilich, und das wäre ein theoretisch abgesichertes Vorgehen, auf der Grundlage beobachteter Daten Modelle erstellen, deren Gültigkeit in weiterer Folge an anderem (u. a. auch umfangreicheren) Sprachmaterial in Form von Hypothesen und deren Überprüfung zu testen ist. In diesem Fall ist man jedoch bereits bei inferentiellen Verfahren, was auch schon für den (statistischen) Vergleich von zwei Stichproben gilt, insofern hier getestet wird, ob beide ein und derselben Grundgesamtheit entstammen. Auf jeden Fall ist hier die qualitative und quantitative Formulierung von Hypothesen, deren empirische Überprüfung und abschließende Interpretation Voraussetzung.

Im Prinzip betreffen diese Punkte die Arbeit mit Wörterbüchern in gleichem Maße wie die mit Korpora. Wenn man bei einem Textkorpus von einem (hier nicht näher zu definierenden) Text als konstitutiver Minimaleinheit ausgeht, dann stellt die prinzipiell begrenzte Menge der miteinander kombinierten Texte T_1, T_2, … T_n das Korpus dar. Die Texte können im Prinzip genuin mündlicher, genuin schriftlicher oder verschriftlichter bzw. transliterierter Art sein, und sie können im Grunde genommen – was heute eher üblich ist – in elektronischer (digitaler) Form vorliegen oder nicht. In jedem Fall aber müssen Korpora kompiliert werden, und je nach Beschaffenheit des Korpus wird unterschiedliches Material in unterschiedlichem Maße „enthalten" sein – was, wie gesagt, weniger für deskriptive Vorgangsweisen ein Problem darstellt als für verallgemeinernde Ambitionen von Relevanz ist.

Unter anderem werden in einem solchen Korpus – wenn es denn umfangreich genug ist – auch phraseologische Einheiten verschiedenen Typs vorkommen, die in diesem Fall dann zwar in gewissem Sinne „gegeben", damit aber noch lange nicht aufgefunden sind. Um solche in einem Korpus zu finden (bzw. zu identifizieren und aus dem Korpus zu extrahieren), müssen Suchstrategien eingesetzt werden, und deren Art richtet sich wiederum nach der Beschaffenheit bzw. Aufbereitung des gegebenen Korpus. Ist das Korpus in elektronischer Form gegeben, und das ist heute der Regelfall, dann hängen die möglichen Such- und Auffindungsstrategien wesentlich von der Qualität des jeweiligen Korpus ab, doch freilich geht die Suche auch hier nicht ohne Humanressourcen vonstatten. Ist nämlich das Korpus (noch) nicht spezifisch aufgearbeitet oder annotiert, dann können diese Suchstrategien sich primär auf objektsprachliches Benutzerwissen stützen: man kennt eine phraseologische Einheit und sucht nach dieser oder nach einzelnen ihrer Komponenten in Form von sog. String-Recherchen, also einfachen Zeichenketten, bestehend aus Elementen der gesuchten Einheit. Hier gilt also: man kann nur suchen, was man kennt. Dies gilt auch für die Erhebung von Konkordanzen, also sprachliche Vorkommnisse in ihrem unmittelbaren sprachlichen Kontext. Allerdings ergibt sich auch die Möglichkeit der automatischen, d. h. computerbasierten Suche nach phraseologischen Einheiten und deren Extraktion aus dem Korpus. In diesem Fall verlässt man sich auf das phraseologische Kriterium der Festigkeit, demzufolge eine phraseologische Einheit aus mehr als einer lexikalischen Einheit besteht, wobei diese Einheiten stereotyp miteinander verbunden werden: Berechnet werden hier üblicherweise die Vorkommenshäufigkeiten der einzelnen Komponenten und verschiedene daraus ableitbare Zusammenhangsmaße, die allerdings meistens mit statistischen Problemen verbunden sind, auf die hier nicht im Detail einzugehen ist. So lassen sich Kookkurrenzen feststellen, die in der Regel nicht mehr und nicht weniger als phraseologische „Kandidaten" sind, und die es in weiterer Folge von festen Wortverbindungen, allgemeinen Kollokationen usw. zu unterscheiden sowie auf die eine oder andere Art, aber üblicherweise wiederum nicht ohne Bezugnahme auf Humanressourcen, zu verifizieren bzw. klassifizieren gilt. Die Identifikation spezifisch phraseologischer Einheiten kann entweder auf der Basis von Introspektion geschehen – ein aufgrund der Subjektivität womöglich recht unzuverlässiges Verfahren – oder in Form von Informantenbefragungen – was je nach Art der Befragung und der Befragten eigene Probleme nach sich ziehen kann, aber die Entscheidung zumindest auf eine breitere Basis stellt. Natürlich kommt auch der Vergleich mit spezifischen Datenbanken (so sie denn existieren) oder der Vergleich mit Wörterbüchern verschiedener Art in Frage – doch hier hängt das Ergebnis von deren jeweiliger Beschaffenheit ab. Ist das Korpus annotiert, was in der Regel wiederum den vorherigen Einsatz von Humanressourcen der einen oder anderen Art voraussetzt, und sind dabei phraseologische Einheiten durch metasprachliche Tags ausgezeichnet (was im Rahmen der Phraseologieforschung allerdings auch heute noch nach wie vor ein allgemeines Desiderat ist), ergeben sich – bei Vorhandensein entsprechender Interfaces, die zwischen dem Datenmaterial und dem Benutzer „vermitteln" – weitere, u. a. eben auch metasprachliche Suchstrategien. Doch dies setzt bereits die Identifikation und Annotation phraseologischen Materials in einer früheren Phase der Korpus-Aufbereitung voraus, und von diesen Zielen ist die Phraseologieforschung auf breiter Ebene noch weit entfernt.

Im Vergleich zu einem Korpus ist sprachliches Material in einem Wörterbuch nicht einfach im oben dargestellten Sinne „enthalten", sondern es muss zuvor auf besondere Art und Weise erhoben, selegiert, zusammengestellt und aufbereitet worden sein, bevor es in der letztendlichen Form Eingang in ein Wörterbuch findet. Insofern ergibt sich im Hinblick auf den Status von Korpus und Wörterbuch eine essentielle zeitliche und qualitative Differenz.

Allgemein kann man davon ausgehen, dass ein Wörterbuch im engeren Sinne den Wortschatz einer Einzelsprache bzw. einen definierten Teilwortschatz dieser Sprache erfassen und abdecken soll. Dabei kann das Material eines Wörterbuchs im Prinzip genau so auf einem einzigen Text wie auf einem Textkorpus beruhen, und das auf dieser Basis erstellte Wörterbuch kann sowohl allgemeiner Natur sein (d. h. nicht nur phraseologische Einheiten enthalten), oder es kann ein spezifisch phraseologisches Wörterbuch sein. Dabei zeichnet sich ein Wörterbuch, wenn es nicht gerade ein Wortformen-Wörterbuch ist, durch die Lemmatisierung der Einträge aus, begleitet und ergänzt durch weitere (meta-)sprachliche Informationen, angefangen von Schreibung und Aussprache über grammatische Angaben (Wortart, Genus, usw.) bis hin zu erklärenden Angaben zu Herkunft, Bedeutung, Verwendungsweise, Äquivalenzen, usw. Je nach Art der Angaben (und in Abhängigkeit davon, ob die Angaben in derselben oder einer bzw. mehreren anderen Sprache/n erfolgt), hat man es mit unterschiedlichen Typen von Wörterbüchern zu tun.

Die Aufnahme phraseologischen Materials in ein solches Wörterbuch im engeren Sinne kann also eigentlich nur unter der (heute eher nicht mehr vertretenen bzw. vertretbaren) Annahme der Wortäquivalenz von Phrasemen geschehen, ansonsten hat man es bereits mit Wörterbüchern in einem allgemeineren Sinne zu tun, in welchen *inter alia* vorkommende phraseologische Einheiten wieder spezifische Suchstrategien erfordern, was gleicherweise auch für spezielle phraseologische Wörterbücher gilt. Diese Suchstrategien sind ihrerseits wiederum nicht zuletzt von der Beschaffenheit des Materials abhängig: Handelt es sich um elektronisches (digitalisiertes) Wörterbuch, sind auch elektronische Abfragen möglich. Doch auch elektronisches Material bzw. elektronische Suchen und Abfragen erfordern spezifische Suchstrategien und in der Regel muss man zuerst wissen, wonach man sucht, bevor man die Suche ausführt. Der Erfolg hängt dann nicht zuletzt von der Art des Wörterbuchs (einsprachige vs. zwei- oder mehrsprachige, allgemeine vs. Lernwörterbücher, spezifische phraseologische Wörterbücher, usw. usf.) ab.

Die aufgezeigte Komplexität gilt im Prinzip – wenn auch auf teilweise andere Art und Weise – ebenso für spezifische Datenbanksysteme, die für spezifische Zwecke mitunter die Funktion von (phraseologischen) Wörterbüchern übernommen haben. Denn auch die Erstellung und Nutzung von üblicherweise aus zwei Teilen – der eigentlichen Datenbank bzw. Datenbasis einerseits, dem Datenbank-Managementsystem andererseits – bestehenden Datenbanksystemen setzen Humanressourcen voraus, unter Einschluss spezifischer, zwischen den Benutzern und der Datenbank und ihrer Struktur vermittelnden Interfaces.

In der Zusammenschau ergibt sich so eine Reihe von Unterschieden zwischen der Phraseologie in Wörterbüchern und in Korpora. Wörterbücher, die ja nicht einfach nur Wortlisten (bzw. Wortformenlisten) sind, und die bereits das Ergebnis linguistischer Bearbeitungen beinhalten, setzen die Analyse sprachlichen Materials und in weiterer Folge dessen lexikographische bzw. (abhängig davon, ob es sich um spezifisch phraseologische Wörterbücher handelt) dessen spezifisch phraseographische Aufbereitung voraus. Im Vergleich dazu können Korpora – und damit sind nicht spezifische (elektronische) phraseologische Datenbanken gemeint, sondern eben Textkorpora mit Kompilationen von Sprachdaten – zu unterschiedlichem Grad vorverarbeitet, annotiert usw. sein. Ungeachtet dieser allgemeinen Unterschiede, was den Status phraseologischen Materials in Wörterbüchern und Korpora anbelangt, gilt es im Hinblick auf die Frage möglicher Beziehungen zwischen Wörterbüchern und Korpora zu beachten, dass auch diese äußerst vielfältig und komplex sein können: Einerseits kann die Arbeit an und mit Korpora Voraussetzung für die Erstellung von Wörterbüchern, für die Erhebung, Verifizierung und Quantifizierung phraseologischen Einheiten sein, andererseits kann die (manuelle oder auch automatisierte) Suche nach phraseologischen Einheiten in Korpora sich an Lexikonmaterial orientieren und dieses für die Suchen nutzen.

Vor dem Hintergrund dieser vielfältigen und vielschichtigen Differenzierungen und wechselseitigen Beziehungen scheint es uns angebracht, die im vorliegenden Band zusammengeführten Beiträge nicht in einzelne Rubriken zu unterteilen, sondern schlicht und einfach alphabetisch anzuordnen. Einzelnen Beiträgen sind Abstracts vorangestellt, wobei auch die folgenden Kurzdarstellungen eine erste Orientierung geben mögen. Die Beiträge gehen auf die lexikographische Sektion der traditionellen zweijährigen Tagung der Europäischen Gesellschaft für Phraseologie (EUROPHRAS) zurück, die in der Zeit vom 27. 8. 2012 bis 31. 8. 2012 an der Universität Maribor stattfand.

In TORBEN ARBOE's (Aarhus) Beitrag *Phraseology – Central Parts of Culture Treated in a Dictionary* wird über ein jütländisches Dialektwörterbuch berichtet, das Kollokationen und Phraseme verzeichnet; der Autor diskutiert ausgewählte Beispiele aus dem semantischen Bereich der Haustiere.

ELENA BERTHEMET's (Brest) Beitrag *Colidioms. A Contribution to Cross-Cultural Research* gibt einen Überblick über ein Projekt mit dem Ziel der Gestaltung einer mehrsprachigen phraseologischen Datenbank für phraseologische Systeme unterschiedlicher Sprachen unter Einschluss semantischer und syntaktischer Aspekte diskutiert.

Ausgangspunkt der Studie *Corpus-Driven Phraseology Assessment: an Experiment* von JEAN-PIERRE COLSON (Brussels/Louvain-la-Neuve) ist die empirisch überprüfte Beobachtung, dass die Verwendung von Phrasemen (in einer breiten Auffassung dieses Begriffes) bei Nicht-Muttersprachlern in doppelter Hinsicht gekennzeichnet ist: einerseits sind bestimmte Strukturen im Sprachgebrauch unterrepräsentiert, andererseits werden gleichzeitig bestimmte Strukturen bevorzugt.

COSIMO DE GIOVANNI (Cagliari) zielt auf eine Revision der Beziehung zwischen Synonymie und Kollokation auf der einen Seite und zwischen Korpus und zweisprachigem Wörterbuch auf der anderen; in seiner Studie *La synonymie collocationnelle. Entre corpus et dictionnaire bilingue* werden relevante Beispiele analysiert, um den Unterschied zwischen korpusempirischen Sprachdaten und lexikographischer Behandlung zu zeigen.

Im Beitrag von MARCEL DRÄGER, RENÉ FRAUCHIGER, MARLÈNE LINSMAYER und ALESSANDRA WIDMER (Basel) mit dem Titel *Kollokationenlexikographie. Ein Bericht aus der Praxis* werden praktische Probleme der Erstellung eines Kollokationswörterbuchs diskutiert; wie sie überzeugend zeigen, müssen quantitativ-statistische Kookkurrenzanalysen durch anschließende qualitative Bearbeitungsverfahren ergänzt werden.

In der Studie *Étude du figement dans les Curiositez françoises (1640) d'Antoine Oudin* von CLAIRE DUCARME (Liège) wird das Problem der festen Wortverbindungen in diachroner Perspektive dargelegt, wenn die Festigkeit phraseologischer Einheiten nicht durch intuitiv-introspektive Verfahren oder die Sprachkompetenz heutiger Sprecher belegt werden kann; bei ihrer Analyse des phraseologischen Sprachmaterials aus einem französischen Wörterbuch des 17. Jhd.s konzentriert sie sich deshalb auf interne (implizit oder explizit durch den Lexikographen gegebene) und externe (durch die Konsultation anderer lexikographischer und literarischer Quellen gewonnene) Hinweise, um gefrorene phraseologische Strukturen identifizieren und deren Status bestimmen zu können.

Der Beitrag von PETER GRZYBEK (Graz) und DARINKA VERDONIK (Maribor) mit dem Titel *General Extenders: From Interaction to Model* basiert auf der Annahme, dass allgemeine Erweiterungsformeln wie *etc.*, *usw.*, u. a. m. eine eigene Kategorie im sprachlichen und phraseologischen System einer Sprache darstellen; es wird gezeigt, dass das Auftreten dieser formelhaften Wendungen regelhaft und gesetzmäßig organisiert und modellhaft als Ergebnis eines Diversifikationsprozesses erfasst werden kann.

AI INOUE (Yokosuka) konzentriert sich in ihrer Studie *Study of a new phraseological unit – 'be on against' as an example* auf ein jüngeres Phänomen im gegenwärtigen Englisch, das noch keinen Eingang in relevante Wörterbücher gefunden hat; analysiert wird ein Phänomen, bei dem zwei als „komplex" bezeichnete Präpositionen zu einer neuen phraseologischen Einheit mit einer einheitlichen Bedeutung zusammengefügt werden.

EMMERICH KELIH (Wien) behandelt in seinem Beitrag *Paarformeln und Binomiale im Slowenischen: Ein korpusbasierter Ansatz* reversible Paarformeln im Slowenischen. Nach einer einleitenden Diskussion über die Gründe für die Reihenfolge der einzelnen Komponenten einer Paarformel geht er auf Aspekte der Variabilität ein und plädiert für eine stärkere Berücksichtigung der Vorkommens- und Verwendungshäufigkeit bei der Untersuchung dieses Phrasem-Typs.

From Dictionary to Corpus lautet der Titel des Beitrags von MARIE KOPŘIVOVÁ und MILENA HNÁTKOVÁ (Prag); hier wird die Frage aufgeworfen, wie man idiomatische Wendungen in großen Korpora identifiziert, wenn intelligente Suche und Suchstrategien nötig sind; sie diskutieren und zeigen die Möglichkeit des Gebrauchs speziellen phraseologischen Wörterbuchmaterials für automatisierte Suchstrategien, die geeignet sind, Informationen bezüglich der Häufigkeit und Verteilung phraseologischen Materials in den Texten eines Korpus zu verschaffen.

Nataša Kralj (Maribor) kritisiert in ihrer Studie *Digitalisierung der Phraseologie und der Benutzer-Aspekt* den nach wie vor bestehenden Mangel an benutzerorientierten Untersuchungen elektronischer phraseologischer Ressourcen insbesondere im Hinblick auf Materialien für den Fremdsprachenerwerb.

Claudia Lückert, geb. Aurich (Münster) will im Beitrag *Prosodic Aspects of Proverb Change in English: Panini's Principle* zeigen, dass nicht nur Phraseme mit Wortgruppenstatus, sondern auch Sprichwörter zu einer semantischen und phonologischen Musterhaftigkeit in ihrer Abfolge neigen, und dass diese Tendenzen besonders in einer diachronen Perspektive von Bedeutung sind; im Detail untersucht sie den Einfluss von Panini's (auch als Behaghelsches Gesetz bekanntem) Prinzip in Sprichwörtern, betont aber, dass zu Analysezwecken gleichzeitig auch weitere phonologische Prinzipien zu berücksichtigen sind.

Jasmina Markič's (Ljubljana) Beitrag *Acerca de la (in)traducibilidad de las unidades fraseologicas en la interpretacion de conferencias* behandelt phraseologische Einheiten in translatorischen Zusammenhängen (genauer: im Konferenzdolmetschen); exemplarisch werden Probleme der Übersetzung von Kollokationen und Sprichwörtern aus dem Spanischen ins Slowenische analysiert.

Matej Meterc (Ljubljana) stellt ein laufendes Projekt zur Bekanntheit slowenischer Sprichwörter dar (*Online Questionnaire Providing Information on most Well-known and Well-understood Proverbs in Slovene Language*); die online-Umfrage enthält 918 Sprichwörter aus zwei lexikographischen Quellen, weiterführende Studien sind geplant.

Vesna Mikolič (Koper) diskutiert ausgewählte phraseologische Einheiten aus dem slowenisch-englischen terminologischen Wörterbuch im Bereich Tourismus (*Večbesedni termini v turističnem terminološkem slovarju*); die Autorin betont die Wichtigkeit der typologischen Unterscheidung zwischen Kollokationen, nicht-phraseologischen Mehrwort-Fachtermini und Phrasemen in Bezug auf ihre lexikographische Behandlung.

Piotr Pęzik's (Łodź) Artikel *Graph-based Analysis of Collocational Profiles* bespricht ausgewählte Aspekte der automatischen Erarbeitung und Benutzung von Kollokationswörterbüchern; ein Schwerpunkt liegt dabei auf graphenbasierten Methoden der Erforschung und Visualisierung des analysierten (phraseologischen) Sprachmaterials.

Liezl Potgieter (Stellenbosch) thematisiert die Behandlung der Phraseologie in Übersetzungswörterbüchern (*Idioms in Dictionaries for Translators*); ausgehend von der Annahme, dass die gängigen zweisprachigen Wörterbücher für Zwecke der professionellen Übersetzung phraseologischer Einheiten un-geeignet sind, unterbreitet sie Vorschläge für eine angemessenere und übersetzerfreundlichere lexikographische Behandlung phraseologischer Einheiten in zweisprachigen Wörterbüchern.

MAKOTO SUMIYOSHI (Osaka) konzentriert sich auf die Analyse der Präsentation von Valenzen in einsprachigen Lernerwörterbüchern und Valenzwörterbüchern und vergleicht sie mit empirischen Sprachdaten (*Valency Patterns in Dictionaries*); es wird gezeigt, dass Phraseologen in Zusammenarbeit mit Lexikographen durch solche Vergleiche bestimmte Sprachwandelprozesse besser nachvollziehen und zu einer besseren lexikographischen Kodifizierung beitragen können.

CLAUDIA MARIA XATARA (São Paulo) stellt ein brasilianisch-portugiesisches phraseologisches Online-Wörterbuch vor (*Un projet phraséographique: critères et choix*), welches neben Definitionen und illustrativen Beispielen auf Synonyme und Äquivalente im portugiesischen Portugiesisch und in drei Varietäten des Französischen (in Frankreich, Belgien und Kanada) enthält.

Peter Grzybek, Vida Jesenšek

Phraseology in Dictionaries and Corpora. Introductory Remarks

Phraseology *in* the dictionary, phraseology *in* the corpus – this sounds less complex than it is: upon closer examination, we are faced with a multi-faceted matter. Not only can we understand ‘phraseology’ as a scientific discipline (i. e., in terms of phraseological research), but we can also think of it as an object of study (i. e., the treasure of phraseological units occurring in clearly defined linguistic material or even in a language as a whole). After all, phraseology is not contained per se in a dictionary or a corpus, and thus “simply placed”, at our disposal for other purposes – starting from general interests for private use, through instructional and educational purposes, right up to phraseological research. Rather, the dictionary and corpus (or, more correctly: different kinds of dictionaries and corpora) are different in this and other respects and, additionally, stand in multiple complex interrelations. But, first and foremost, phraseological material is not “simply given” – not in a corpus or in a dictionary.

In order to provide a general framework for the individual contributions to this volume, and with regard to the variety of paths and ways suggested and pursued here, it seems reasonable to discuss the problems mentioned, albeit in a most general manner, notwithstanding the fact that some, or even most, of the following remarks may appear to be more or less self-evident to a phraseologically oriented audience.

Assuming that we are concerned with a corpus of text, ‘corpus’ is usually understood as a more or less systematic compilation of linguistic material, serving to make empirical observations, i. e., to collect specifically linguistic data, which may refer either to an individual phraseological unit (in this case resulting in what might be termed a single case study) or to a particular set of phraseological units. In any case, the objective is a statement about the presence or occurrence of particular elements, or a generalizing statement, the validity of which may be intended to go beyond the corpus under study.

In the simplest case, we are merely concerned with symptomatic description: some occurs in the material under study, or it does not occur; this is, after all, a simple binary, dichotomous classification. The subsequent information about frequency of occurrence leads to the categories of quantity or degree, providing information on how often, or to what degree, a particular phenomenon occurs. This may, but need not, go beyond the symptomatic state, not least depending on whether the frequencies are relativized in one way or another, i. e., related to some totality, or with regard to some comparable amount of data. In this case, we are, however, still concerned with descriptive procedures, and in principle there is still no need to reflect on the quality of the given corpus: it is as it is, and all statements concern no more and no less than the available data analyzed. In case one intends to make far-reaching conclusions (or, to be more cautious, assumptions or hypotheses), which go beyond the data observed, one usually regards the corpus as a "sample" for a (assumed) totality, or even for (a given) language as a whole. From a theoretical point of view, this last assumption is difficult to sustain because "language as a whole" does not exist, regarding language to be an inherently heterogeneous, dynamic and continuously evolving system: 'language' is neither the sum of all texts ever produced, nor is it the sum of all texts ever to be produced – 'language' is tangible only as an abstract construct, based on linguistic observations and generalizations.

What can be done instead – and this would be a theoretically substantiated procedure – is to create models on the basis of observed data, the validity of which can subsequently be applied to other (including larger) data sets of linguistic material, in the form of hypotheses and testing. However, this employs inferential processes, which is true even for the (statistical) comparison of two samples, insofar as what is tested here is the assumption that both samples originate from one and the same population. In any case, the qualitative and quantitative formulation of hypotheses, empirical testing, and the final interpretation represent a sine qua non condition.

In principle, these remarks concern dictionaries as well as corpora. Given that a text (not necessarily to be specified here in detail) is a constitutive minimal unit of a given corpus, then the principally limited set of combined texts T_1, T_2, … T_n represents the corpus. The texts can be genuinely oral, written or transliterated, and they can be available in electronic (digital) form – which is the standard today – or not. In any case, corpora must be compiled, and depending on the nature of the corpus, various material will be represented to varying extent – which, as has been said above, constitutes a problem for generalizing ambitions of relevance, rather than for descriptive procedures.

If such a corpus is large enough, phraseological units of different kinds will occur; these are then, in a certain sense, "given", but still not detected. In order to identify these successfully (i. e., to identify and extract from the corpus), search queries are needed, the quality of which depends on the kind of corpus given. In the case of an electronic corpus (which is the standard case today), the possible search and retrieval strategies essentially depend on the pre-processing procedures, which, in turn, pre-suppose human resources of one kind or another. If the corpus is not, or not yet, specifically pre-processed, or annotated, (e. g., grammatically tagged, syntactically parsed), such search strategies can only rely primarily on object-language based user knowledge: knowing a given phraseological unit one can search for it, or for its individual components, by way of string searches, i. e., simple chains of characters as parts of the requested unit. It holds true here that you can only search for what you know. The same relates to searching for concordances, that is, the presence and occurrence of linguistic units in their immediate contexts. However, the possibility of an automatic, computer-based search for phraseological units and their extraction from corpora also exists: in this case, one relies on the phraseological criterion of frozenness, according to which a phraseological unit is composed of more than one lexical entity, stereotypically fused, or merged, into one whole. Works along this line usually count the individual components' frequency of occurrence and then calculate a measure of association, or correlation, most of which are not void of statistical problems, which need not be discussed in detail here. Lexical co-occurrences, which are usually only phraseological "candidates", can be detected this way and at the next step they should be separated from fixed lexical combinations, general collocations, etc., and be verified again (and, eventually, classified), necessarily relying on human resources. The identification of specifically phraseological units can be undertaken based on either introspection – a procedure rather unreliable due to the subjectivity involved – or in the form of interviews and surveys with informants – a procedure which, depending on the method chosen, may cause problems in its own right, but at least the results and decisions come from a broader base and a wider range. As a matter of fact, comparisons with specific databases (if these do exist), or with relevant dictionaries, are also of concern here, depending of course on these sources' quality. If the corpus is annotated – which in turn presupposes the prior employment of human resources – and if phraseological units are specifically marked by way of meta-linguistic tags (which is, however, to this day a general desideratum in phraseological research), then more promising search strategies (including meta-lingual) are at our disposal, given the existence of adequate interfaces, mediating between user and data. This, however, asks for the prior detection, identification, and annotation of phraseological material in earlier phases of corpus pre-pressing, and achieving this state is still a long way away for contemporary phraseological research.

As compared to a corpus, linguistic material in a dictionary has not simply been gathered within it (in the sense outlined above), but it must be specifically collected, compiled and processed, before it finds its way into the dictionary, in its ultimate form. This circumstance is a trivial but crucial difference between the status of a corpus and a dictionary, in both temporal and qualitative aspects.

Generally speaking, one can say that a dictionary is intended to cover a given language's lexical treasury, or a clearly defined part of it. In this framework, the dictionary's material can in principle be based on a single text as well as on a text corpus, and the dictionary derived from these sources can be of both a natural kind (i. e., contain not only phraseological units) and a specific phraseological dictionary. Moreover, a dictionary is usually characterized by lemmatization of its entries (unless we are concerned with a specific word form dictionary), accompanied and complemented by further (meta-lingual) information, starting from orthography and pronunciation, through grammatical (part of speech, gender, etc.), to explanatory information about origin, meaning, usage, translations, synonyms, equivalences, etcetera. Depending on the kind of information given (and depending on whether this information is given in one and the same or one or more other language/s), we are concerned with different kinds of dictionaries.

Strictly speaking, phraseological material can thus be included in such a dictionary in a narrow sense of the term only under the assumption of a phraseme's word equivalence – otherwise, we would be concerned with a dictionary in the broader sense of this term; these dictionaries contain *inter alia* phraseological units, asking for specific search and query strategies, which likewise holds true for specific phraseological dictionaries. Search strategies thus depend, among other features, on the dictionary's quality: If the dictionary is not given in an electronic form, one can only search manually, and the success rate will largely depend on the arrangement of the dictionary entries; but even in the case of electronically available material, specific search and query strategies are needed, and usually it is necessary to know what exactly one is looking for, the success often depending on the kind of dictionary one is using (mono- or multilingual, dictionaries for special purposes such as for language learning, specific phraseological dictionaries, etc.).

Most of these remarks likewise hold (albeit in a somewhat different way) for specific data base systems, which for specific purposes may take the function of (phraseological) dictionaries, which usually consist of two parts, the database itself, and the database management system – they, too, ask for human resources, including specific interfaces capable of mediating between users and the data base structure.

In conclusion, a variety of differences can be observed between phraseology in dictionaries and in corpora. Dictionaries, which are more than simple word lists (or word form lists), represent the result of linguistic processing, based on the analysis of linguistic material and its lexicographic (or, eventually, phraseographic) treatment. As compared to this, text corpora as specific compilations of linguistic material may be pre-processed in different ways and to different degrees. Notwithstanding, these differences as to the phraseological material's status in dictionaries and in corpora, manifold and possibly complex relations between these two must be taken into account: on the one hand, work with corpora can be a presupposition for the compilation of dictionaries, the collection, identification, verification, or quantification of phraseological units; on the other hand, the manual or automatic search for phraseological units in corpora may be oriented towards and based on specific dictionary material.

Given these multiple and multi-layered distinctions and manifold relations and interrelations, it seems reasonable to organize the contributions to this volume in a simple and straightforward alphabetical way, rather than in thematically defined sections. All contributions are preceded by short abstracts, but the following short summaries may also be helpful for the reader and serve as a first orientation. Presentations based on these contributions were held at the EUROPHRAS conference (a traditional biannual conference organized by the European Society of Phraseology – EUROPHRAS) hosted by the University of Maribor between the 27th and the 31th of August I 2012.

In TORBEN ARBOE's (Aarhus) contribution *Phraseology – Central Parts of Culture Treated in a Dictionary*, we find a report about a Jutlandic dialectal dictionary, which contains collocations and idiomatized set phrases; by way of an illustration, the author discusses selected examples from the domain of domestic animals.

ELENA BERTHEMET's (Brest) *Colidioms. A Contribution to Cross-Cultural Research* presents an overview of a project aiming at the design of a multilingual phraseological database for phraseological systems of different languages, including semantic as well as syntactic information.

The starting point of JEAN-PIERRE COLSON's (Brussels/Louvain-la-Neuve) study *Corpus-Driven Phraseology Assessment: an Experiment* is the observation that the use of phraseology (in a broad understanding of this term, including multi-word expressions) by non-native speakers is characterized by the underuse of some structures and simultaneous overuse of others.

COSIMO DE GIOVANNI (Cagliari) aims at a revision of the relation between synonymy and collocation, on the one hand, and between corpus and bilingual dictionaries, on the other. In his study *La synonymie collocationnelle. Entre corpus et dictionnaire bilingue*, relevant examples are analyzed in order to demonstrate the difference between corpus evidence and lexicographic treatment.

Marcel Dräger, René Frauchiger, Marlène Linsmayer und Alessandra Widmer (Basel), in their article *Kollokationenlexikographie. Ein Bericht aus der Praxis* show how quantitative-statistical co-occurrence analyses must necessarily be complemented by subsequent qualitative editing procedures in compiling collocation dictionaries.

In her study *Étude du figement dans les Curiositez françoises (1640) d'Antoine Oudin*, Claire Ducarme (Liège) raises the problem of frozen structures of old languages, or of past conditions of languages, when it is not possible to study the frozenness of phraseological items using modern speakers' competence or intuitive/introspective methods; analyzing material from a French 17th century dictionary, she focuses on both internal (implicitly or explicitly given by the lexicographer) and external (obtained through the consultation of other lexicographic and literary sources) indications, aiming to identify frozen structures and to determine their status.

Peter Grzybek's (Graz) und Darinka Verdonik's (Maribor) contribution *General Extenders: From Interaction to Model* is based on the assumption that general extenders represent a separate category in the linguistic and phraseological system of a given language; their study attempts to show that the frequency of occurrence of general extenders is organized in a regular and law-like manner, as the result of a diversification process, and presents a relevant theoretical model.

Ai Inoue's (Yokosuka) *Study of a new phraseological unit – 'be on against' as an example* concentrates on a recent phenomenon to be observed in present-day English that has not yet found entrance into relevant dictionaries, when two prepositions (termed 'complex prepositions') are put together with a single meaning, becoming established as new phraseological units.

Emmerich Kelih (Vienna), in his *Paarformeln und Binomiale im Slowenischen: Ein korpusbasierter Ansatz*, studies reversible binomials; subsequent to a short synopsis about reasons for the order of a binomial's components, he studies aspects of phraseological variability and argues in favor of taking into account frequency as an important factor related to a binomial's linguistic form.

From Dictionary to Corpus is the title of Marie Kopřivová's und Milena Hnátková's (Prague) article; the authors raise the question of how to identify idiomatic expressions in large corpora, when intelligent quest and search strategies are needed; they discuss and demonstrate the possibility of using special phraseological dictionary material for automatic search strategies apt to yield information of frequency and distribution of phraseological material in a corpus' texts.

NATAŠA KRALJ (Maribor), in her Study *Digitalisierung der Phraseologie und der Benutzer-Aspekt*, criticizes that there are no sufficient user-oriented studies on the usability of electronically based phraseological material, particularly as far as foreign learners, or foreign language learning material, is concerned.

CLAUDIA LÜCKERT, geb. AURICH (Münster), in her contribution *Prosodic Aspects of Proverb Change in English: Panini's Principle*, shows that not only set phrases, but also proverbs, tend to show specific patterns of semantic and phonological sequences, and that these tendencies are important from a diachronic perspective; in detail, she studies the influence of Panini's Principle (also known as Behaghel's Law), in proverbs, showing that this principle is at work, but that further phonological principles may need to be taken into account at the same time.

JASMINA MARKIČ's (Ljubljana) article *Acerca de la (in)traducibilidad de las unidades fraseologicas en la interpretacion de conferencias* deals with phraseological units and conference interpreting, and analyses examples of collocations, locutions, paremias and formulas which appear in Spanish speeches and are translated into Slovene.

MATEJ METERC (Ljubljana) presents an on-going project on the familiarity of Slovene paremiological units (*Online Questionnaire Providing Information on most Well-known and Well-understood Proverbs in Slovene Language*); the questionnaire, consisting of 918 units from two lexicographic sources, is presented online as a full text presentation. Further follow-up studies are planned.

VESNA MIKOLIČ (Koper) studies idiomatic expressions, using selected examples to be included in a Slovenian-English dictionary of tourism terminology. As the author argues in her contribution *Večbesedni termini v turističnem terminološkem slovarju*, the distinction between collocations, non-phraseological multi-word terms and phrasemes turns out to be important, since only the latter are included as entries (including idiomatic or non-idiomatic expressions), whereas terminological collocations are stated in the dictionary entry at one of the entry words.

PIOTR PĘZIK's (Łodź) *Graph-based Analysis of Collocational Profiles* focuses on the study of distributional characteristics of phraseological units; he discusses selected aspects of generating and using automatic collocation dictionaries in phraseological studies, with particular emphasis on graph-based methods of exploring and visualizing the (phraseological) material under study.

According to LIEZL POTGIETER (Stellenbosch), bilingual dictionaries are an inadequate resource for professional translators when translating idioms; the author makes some suggestions for improving the treatment of idioms in bilingual dictionaries and making them more user-friendly for translators.

MAKOTO SUMIYOSHI (Osaka) focuses on the analysis of *Valency Patterns in Dictionaries*; studying valency patterns in monolingual learners' dictionaries, valency pattern dictionaries, and authentic data, the author shows that the comparison of descriptions in these dictionaries, phraseologists, in collaboration with lexicographers, can contribute to the clarification of language change, phraseology thus being able to play a role in language research, especially lexicography, that is more important than usually assumed.

CLAUDIA MARIA XATARA (São Paulo) presents a Brazilian-Portuguese online dictionary of idioms (*Un projet phraséographique: critères et choix*), which contains definitions, additional information, illustrative examples, indications of synonymy (if any), and equivalents in Portuguese of Portugal and in the three variants of French (France, Belgium and Canada).

Peter Grzybek, Vida Jesenšek

Phraseology – Central Parts of Culture Treated in a Dictionary

Torben Arboe (Aarhus)

Abstract

The dialectal dictionary behind the study is shortly introduced, and basic terminological notions are defined; further terms are shortly mentioned. The material is presented: phrasemes with lexemes for domestic animals and animals of the near environment, followed by an analysis that shows different phraseological contexts and meanings of a single lexeme. Some groups of phrasemes are discussed, the largest one being the comparative phrasemes (analogies, similes), as to both collocations and idioms. Idioms together with metaphors and proverbs are dealt with, one section showing idioms used in different speech acts, another pointing to certain types of proverbs. Collocations are discussed together with polysemy, and a few smaller groups of phrasemes are mentioned. Concludingly is pointed to further perspectives, e. g. a broader European context.

1 Introduction

The *Jysk Ordbog* (i. e. *Dictionary of the Jutland Dialects*) which we are editing these years is based on a collection of 3.1 millions of dictionary slips with phonetic, semantic etc. informations from the year 1700 until today. Further, it is based on answers to a wide range of questionnaires (about 150, each with 30–40 questions of dialectal words) from a large net of informants from the middle of the 1950es until today (5–600 persons in the best years), also yielding presumably about 3 millions of informations, most of which have been transferred to databases, a project still under way. We also make use of the informations given in a comprehensive dictionary for the Jutland dialects from about 1900, namely Feilberg (1886–1914). The new Jutland dictionary is an online dictionary, and the section A–H is accessible on the internet <www.jyskordbog.dk>; the section I–J has been edited and will be uploaded as soon as economy permits.

The dialects of Jutland are divided in three main dialects, West, East and South Jutlandic. These can further be divided into 9 primary dialects, and about 20 secondary dialects. The phrasemes in the following are taken from the whole area and most often rendered in Standard Danish as the dialectal versions would complicate matters unnecessarily.

Geographically, Jutland forms the western part of Denmark. A parallel dictionary for the dialects of Danish islands is being edited at the University of Copenhagen, named *Ømålsordbogen* (i.e. *Dictionary of the Insular Dialects*).

2 Terminology

The material consists of mainly two types of phrasemes: collocations and idioms respectively, or at least more or less idiomatized set phrases. Each of the two terms have been defined in different ways during the years of research in the field of phraseology. As to *collocation* I basically rely on the definition by Hausmann (2004): two (or a few) words that usually are used together in a way specific for the language in question, the one named *basis*, the other *collocator*. In the verb phrase *to brush one's teeth* we have *teeth* as the basis and *brush* as the collocator; other languages use other verbs as collocators, French *se laver des dents* (literally, 'to wash one's teeth'), German *die Zahne putzen* ('to polish one's teeth'), whereas Danish in *børste tænder* uses a collocating verb with the same meaning as the English one. This linguistic definition of collocation is chosen in opposition to the statistical definition which in general is used for corpus analysis: two words that just happens to appear next to each other, the more interesting pairs appearing relatively often, especially in a statistically significant number. The linguistic definition thus yields only a subset of the collocations found by use of the statistical definition, but certainly an interesting subset. My linguistic definition hereby follows the same principle as used in the *Oxford Dictionary of Collocations* (2002).

As to *idiom* I use the most common definition, i.e. a string of words or a set phrase that means more than the sum of the words in it; most often it has a *figurative* meaning. For instance, *take to one's heels* 'to run away in a hurry'; in Danish this is expressed by another idiom: *tage benene på nakken*, literally, 'take the legs up upon the neck'. The same may be expressed by the metaphor, *bruge harens gevær*, literally, 'use the gun of the hare'. Not all the following examples are full idioms as these ones, some are only partly idioms, or to use a German term, *Teilidiom* (part idiom) and the corresponding adjective, *teilidiomatiziert*, partly idiomatized.[1] The range of phrasemes to be considered here thus stretches from collocations over 'part idioms' to full idioms.

[1] These concepts of idiomaticity have earlier been sharply (*äusserst kontrovers*) discussed in German phraseological research; later on focus has shifted to the relations between phraseologisms and metaphors (Kühn 2007: 623f.).

3 The material: Phrasemes with lexemes for domestic animals etc.

The material of this investigation consists of collocations and idioms with lexemes denoting domestic animals and other animals of the near environment. These phrasemes are chosen because they form central parts of language and culture, especially the earlier, rural culture, much of it surviving as preserved in the dialects, but some also in Standard Danish.[2] Other languages have parallel circumstances as to "animal phrasemes" in the dialects and standard language.[3] We may use the word *ko* (*cow*) to illustrate some of the differences hinted at in section 1.

3.1 Collocation, figurative meaning, metaphor, polysemy

We have straightforward free word combinations as *rød ko*, *sort ko* [red cow, black cow], but also the sentence, *de er som to røde køer* [they are as two red cows], used about two persons who are very good friends of each other. *To røde køer* here represents a collocation or a set phrase – you cannot say *to sorte køer* [two black cows] in this sentence, nor use other colour adjectives, and maintain the same meaning. The background for the set phrase is presumably an observation of red cows often following each other in pairs, perhaps more often than cows of other colours; further considerations of these possibilities belong to agricultural history.

However, the noun phrase *sort ko* [black cow] is found as a collocation on its own, namely as designating a bottle for aquavit; thus it has a *figurative meaning* and may be called an idiom, or in a way, a *metaphor*. This idiomatic use can be stretched out to a whole sentence, e. g. *jeg er da så ked af, at vor sorte ko er blevet 'sen'* [I am really sorry that our black cow has turned dry], i. e. that the bottle of aquavit has been emptied. However, the material also include sentences as, *jeg skal til tyr med min ko* [I have to take my cow to the bull], which in a perhaps rather rustic or direct manner expresses that the bottle of aquavit must be filled. This sentence shows that the noun *ko* [cow] also without the adjective *sort* [black] may be used in the sense mentioned,

[2] For instance, Toftgaard Andersen (2001) gives an overview of "animal lexemes" including domestic animals etc. (ibid.: 252ff.), refering to the alphabetically ordered phrasal lexicon on the preceding pages of the book.

[3] As to German, e. g. Friedrich (1960: 414ff.); as to English, e. g. *Dictionary of English Idioms* (2002: 52ff.).

which means that perhaps we have a figurative or metaphorical meaning of the noun *ko* [cow] itself and thus really a case of *polysemy*.[4]

The material also includes the noun phrase *den trebenede ko* [the three-legged cow], used of a milk jug with three small, stabilizing "legs" or "feet". This suggests that in the future lexical entry we shall have to establish a special meaning, *ko = beholder* [cow = container] with its own redactional number, and also to decide if this meaning has to be marked as figurative. Not quite an easy task as we are faced with partly idiomatized noun phrases, or *Teilidiome*, to use the German term. Whether we mark the meaning as *figurative* or not we shall have to adduce the sentences with the idiom or metaphor *sort ko* [black cow] there. In the old dictionary of Jutlandic dialects these figurative senses are just mentioned in the last part of the first meaning of the article *ko* [cow], after 3 1/2 pages (7 columns) of other examples and folkloristic informations (Feilberg 1886–1914, 2: 242); in the new *Jysk Ordbog* (*Dictionary of the Jutland dialects*) we shall give them a more conspicuous placing in the article.

3.2 Idiom, metaphor, transparancy

The word *ko* [cow] also offers an example of a full (or almost full) idiom or metaphor, *den blå ko* [the blue cow], meaning the sea, i.e. the North Sea, or in Danish, *Vesterhavet* [the Western Sea]. From North Jutland we have the sentence *den blå ko b(r)øler i aften* [the blue cow is lowing tonight], i.e. the sea, the surf is making noise.[5] This idiom or metaphor is at least partly transparent, and in the sentence *den blå ko malker godt i år* [the blue cow yields much milk this year] one would perhaps expect a meaning like, 'there is a good fishery this year', and this is in fact the usual meaning of it in West Jutland. But, surprisingly, in North Jutland the meaning is more specialized, 'there has been a good many ship wrecks this year'! Correspondingly, the metaphorical sentence *han er blevet rig ved at malke den blå ko* [he has become rich by milking the blue cow] refers to the making of a fortune by stealing wreckage in former times.[6] This must, at least outside the region, be termed a fully opaque and thus idiomatic or metaphorical use of the collocation and its basis

[4] As synonyms we have the birdnames *lærke* [lark] and *kukkemand* [cuckoo] used for the bottle of aquavit, but no relevant set phrases with these lexemes. (The word *kukkemand* is dialectal for Standard Danish *gøg*).

[5] The dialectal verb is *bøle* [to low], Standard Danish has *brøle* [to roar] in stead.

[6] In fact, the bay of the North Sea and the Skagerrak in this region has the name *Jammerbugten* [Misery Bay], presumably because of the large amount of ship wreckages during the centuries.

(or rather the whole verb phrase, *to milk the blue cow*). As to fishery, Jutlandic has the phrase *bruge havet* [to use the sea] as the neutral collocation to indicate fishery as an occupation.

4 Comparative phrasemes, analogies: similes

A frequent type of phrasemes in the material is comparisons between animals and human beings,[7] or more precisely, analogies from the behaviour of animals to the behaviour of human beings. This we have already seen exemplified above, as to the relationship between two good friends, 'like two red cows'; other similes with *cow* as the basis more often focus on this animal's ignorance or slow comprehension. The same holds to some phrasemes with *får* [sheep], e. g. *så glad som fem får i et tøjr* [as glad as five sheep in a tether], i. e. very hilarious. Usually, you had only two sheep in one tether which then split up into two parts, one for each sheep, but this could be enough for the two of them to entangle the whole thing by bouncing and running around, in a hilarious way it may seem to the onlooker, and presumably all the happier if there are as many as five sheep in a tether.

A quick, impatient approach to something can be expressed with the set phrase, *fare i det som hund i hed grød* [rush into it like a dog into hot porridge], i. e. this action is too quick and raises problems, one has better wait a little or slow down. Or perhaps better still, you have to act like a cat in stead, cf. *gå omkring det som katten om den varme grød* [go around it as the cat around the hot porridge], i. e. be hesitant or shy to confront a certain matter. This is the Standard Danish phrase; in Jutlandic also another phrasal verb is used, *kimse ad det som katten ad de varme grød* [turn up one's nose at it as the cat at the hot porridge],[8] i. e. act as if something is not good enough. Something like this may also be expressed by another phrase about the same animal, *så ked af det som kat af sennep* [so sick of it as a cat of mustard]. As another couple of phrasemes with *cat* as the basic lexeme may be mentioned, *se ud som en kat i torden / en gal kat i blæst* [look like a cat in thunder / a mad cat in strong wind], i. e. to look scared, frightened or confused, bewildered respectively.

[7] As in other studies, e. g. Piirainen (2000, 1: 389 et passim).

[8] In most Jutlandic dialects the noun *grød* [porridge] is conceived of as being a collective or plural noun (informally termed *stof-pluralis*, 'matter/substance plural'), hence the definite article *de* of plural forms here as opposed to Standard Danish *den* of singular forms (which yields the phrase *den varme grød* in the preceding example). This grammatical speciality is further discussed in Arboe (2003).

Jutlandic has a similar wide range of phrasemes or idioms as to people's appearances by use of "animal lexemes" as the basis in comparisons. As to someone looking ill or not well e. g. the phrase, *hænge med næbbet som en syg kylling* [let the beak hang down as an ill chicken], and the phrase, *se ud, som om kragerne skulle flyve med ham* [he looks as if the crows should fly away with him]. In passing it may be noted that this use of a characteristic part of an animal as the basis of a collection is in the German tradition called a *Tiersomatismus* ('animal somatism'; cf. Piirainen 2000: 422ff.). As to someone looking very angry Jutlandic among others has the phrasemes, *så olm som en tyr* [as angry as a bull] and, again with a bird's name as the basic lexeme, *så ond som en terne* [as wicked as a tern]. Being opposed to others may also be expressed by the phrase, *altid være imod verden ligesom tudserne* [always to be against the world as the toads], referring to an observation of the toads' way of looking or gazing with large open eyes in an ugly and seemingly wicked manner.

Birds are also used as the point of reference in the sentence, *der var en sladder af dem som syv skader på en gavlende* [there was a chattering of them like seven magpies on a gable], said of a noisy party with much talking. Further it is found in the diminishing phrase, *ikke så stor en fugl, som fjerene bruser til* [not as big a bird as its feathers are ruffled to]. A parallel expression here is, *ræven er ikke så stor, som halen bruser til* [the fox is not as big as its tail stands out], which also tries to diminish a person's boasting.

We may conclude this section with a couple of more or less ironic phrasemes. Firstly, *være så glad som en enøjet hund* [be as happy as a one-eyed dog], i. e. not happy at all. And a more robust analogy, *grine som en død ræv* [to grin like a dead fox]. A little milder, but still ironic, is the phrase, *se så mildt som en tudse* [look as mildly as a toad], i. e. angry or wicked as a toad – which are always against the world, as we have just seen. And, as the last example, *der er ikke mere godt i ham, end der er honning i en skrubtudse* [there is no more good in him than there is honey a common toad], i. e. nothing at all.

5 More on idioms and metaphors

The metaphor *familiens sorte får* [the black sheep of the family] is found in Standard Danish as well as equivalents exist in other languages.[9] It is also found in Jutlandic, but not as commonly there as one might expect. The theme

[9] Cf. *Duden* (11: 650), for instance. The use of this idiom is also described and mapped in the extensive investigation of European idioms, Piirainen (2012: 182ff.).

is reflected in another saying, *skal jeg have skyld for alle de sorte lam, der bliver gjort?* [am I to blame for all the black lambs which are made?], i.e. all the cases of bad luck; a sentence uttered by someone who feels himself unjustly accused of causing bad things. Besides these ones Jutlandic has the phraseme, *være på fårenes side* [to be on the sheeps' side], i.e. to be silly. Another idiom, *have svin på skoven* [to have swine in the wood], means 'to be distraught, not to have one's thoughts gathered', presumably reaching back to times where swine were taken to the woods to gather their food, and the peasant in the village therefore often thinking of his swine in the woods in stead of things more nearby. However, a corresponding, but negated sentence, *ikke have mange svin i skoven* [not to have many swine in the wood], has quite a different meaning, namely 'not to be very rich'.

Also here *cat* and *dog* show up as lexemes, e.g. in idioms like *leve som hund og kat* [to live as dog and cat], i.e. in an unfriendly manner. *Cat* is also used as the basis in *købe katten i sækken* (literally, 'to buy the cat in the sack'), or 'to buy a pig in poke', to use the correct corresponding English idiom. Jutlandic offers another idiom with *cat* and *sack*, *nu har vi katten i sækken* [now we have the cat in the sack], i.e. now the matter is finished. With *dog* as the basic lexeme we have the idiom, *have en hund i rebet* [to have a dog in the rope], presumably derived from the similar phrase, *have en bjørn i rebet* [to have a bear in the rope], both meaning 'to be a little intoxicated'. The last of these idioms relates back to times (at least the 19th century) where a bear and its acquired skills could be presented at markets by a bear leader, who sometimes was a little drunk and walked insecurely. As to *cat*, we further have the phrase, *når katten er 'sat', er mælken besk* [when the cat is satisfied the milk is bitter], which has a parallel with *mouse* as the lexeme: *når musen er mæt, er melet besk* [when the mouse is satisfied, the flour is bitter]. Finally may be mentioned that *bird* is used as the lexical basis in two idioms with opposing meanings, on the one side, *der fløj en fugl op for mig* [a bird flew up for me], i.e. 'I got an idea'; on the other, *han har en fugl* [he has a bird], which means 'he is batty, he is nuts'.

We now turn to idioms with different animal lexemes, but parallel meanings, and the use of them in *speech acts*.

6 Idioms in speech acts: parallel ideas, different animal lexemes

When a person has been irritated by another one's questions or remarks for some time he or she may say, *du kan træde en hund så længe på halen, at den bider* [you may tread a dog for so long on its tail that it bites], or correspondingly, *du kan træde en kat så længe på halen, at den viser kløer* [you may tread a cat for so long on its tail that it shows its claws], i.e. it is going to scratch. Both idioms are meant as a warning that the one person is getting angry with the other one and as an advice to stop talking in that way, further perhaps meant as a threat of a punishment. In other words, the idioms are used in a *speech act*, the kind of which relies on the circumstances of its uttering.

In another idiom the lexical basis is a bird, *flyvende fugl får noget, den liggende intet* [the flying bird will get something, the sitting one nothing], i.e. 'you do not get anything (food etc.) by doing nothing'. A parallel idiom is, *løbende hund får noget, den liggende ikke* [the running dog will get something, the lying one will not]. Also here the idioms may be used in different speech acts, as just comments to why someone has obtained good results or no results, or as requests of trying harder.

In fact there is a good deal of such idioms with parallel animal lexemes, and here I can only mention a few more, e.g. *du kigger så højt, er dine bier* (*duer*) *fløjet hen?* [you are looking so high in the air, have your bees (or pigeons) flown away?]. A person may utter one of these sentences when meeting another if this one does not seem to notice him or her, i.e. the illucutionary act in this is to imply that the other one seems to be a bit haughty or arrogant this day. Here one may also use analogies with other bird or animal lexemes, e.g. *se så højt som skader i blæst* [to look so high in the air as magpies in wind].

If you have something unpleasant to settle with another person it may both in Standard Danish and in Jutlandic be expressed by the idiom, *have en høne at plukke med nogen* [have a hen to pluck with someone], i.e. 'have a bone to pick with someone'. But in Jutlandic you may also use an idiom with the verb in the negated participle form, *uplukket* (literally, 'unplucked', meaning 'not yet plucked'), together with other bird lexemes, e.g. *vi to har en gås uplukket* [the two of us have a goose not yet plucked] and *jeg har en krage uplukket med dig* [I have a crow not yet plucked with you].

Standard Danish also has the idiom, *én fugl i hånden er bedre end ti på taget* [one bird in the hand is better than ten at the roof], i.e. '… is worth two in the bush', whereas Jutlandic may use the lexeme *fisk* [fish] in stead, *en fisk på land er bedre end to i vandet* [one fish on land is better than two in the water].

The idiom, *hunden er slem til at løbe med hans mål* [the dog often runs away with his measure] may be uttered as a humorous or ironic comment to inaccurate work by an artisan; the same is implied by the parallel idiom, *katten er rendt med målet* [the cat has run away with the measure].

These examples of two or more idioms expressing the same underlying concept or idea show that idioms, psycholinguistically speaking, are not just reproduced, but perhaps stored in a special way in the mental lexicon, as a "summary" of knowledge, which can be unfolded in different ways (Piirainen 2007: 535). This principle may also be observed in some of the following idioms.

7 Idioms and proverbs: 'one and all', 'one and (always) the same'

The material offers a few idioms about one individual as an example to be followed by all others, e. g. *når et får løber til vands, løber de alle* [when one sheep runs to the water they all run], referring to the behaviour of sheep in a flock and also transferred to a group of people. Correspondingly with cattle as basis, *når en ko bisser, bisser de alle* [when one cow stampedes they all stampede], originally used about a herd attacked by botflies, but later on presumable just opaque to non-dialectal speakers and also to younger dialect speakers, botflies being extirpated decades ago. Longterm problems are hinted at by the sentence, *et skabet får kan fordærve den hele hjord* [a scabby sheep may spoil the whole flock], figuratively used about the impact of bad behaviour by one member of a group; this is an idiom also found in Standard Danish and having a parallel in German (*ein räudiges Schaf steckt die ganze Herde an*).

Another aspect is hinted at by a sentence which presumably is drawing on shepherds' experiences, *man kan snakke med et får, så længe man vil, det siger alligvel 'mæh' om aftenen* [you can talk with a sheep for as long as you like, it will still say 'meh' in the evening], i. e. you cannot make it say or do new things, a saying also used in transferred meaning about people's stubborn or stupid behaviour. The theme of always beeing the same, also despite attempts of embellishment, is further found in the idiom, *en kan sende en gris til Paris, og så er det endda en gris* [one can send a pig to Paris, it is pig all the same], i. e. a pig will not be changed even if you send it to a glamorous town like Paris. Also the wolf is used as basis in an idiom here, *hvor længe en endda prædiker for ulven, den siger endda "lam!" om aftenen* [irrespective of for how long you preach to the wolf it will still say "lamb!" in the evening], i. e. there is no chance of altering the nature of a beast of prey, neither of the person the idiom is used about.

8 Collocations, polysemy, metaphors

As mentioned in sections 2 and 3.1, some collocations are partly idiomatized; this may hold to e. g. *fange / få / stikke en gedde* [catch / get / prick a pike], meaning 'to get wet feet', which can also be expressed by *fange en torsk / en god fisk* [catch a cod / a good fish]. This may be seen as examples of polysemy because these meanings can be mentioned in the dictionary as special meanings of the lexemes *fisk, gedde, torsk* [fish, pike, cod]. In the new *Dictionary of Jutlandic Dialects* we have placed the idiom with *fisk* as "figurative" at the end of word meaning 1 in the article *fisk*, whereas the idioms with *gedde* are placed as a word meaning 2 in their own right in the entry *gedde*. A full idiom is found in *bruge harens gevær* [use the gun of the hare], i. e. 'run away when frigthened'; it is placed in the section *faste forbindelser* (set phrases) in the last part of word meaning 1 in the article *1. hare*. The sentence, *er du sådan en hare* [are you such a hare], may correspondingly be uttered to a person who is afraid of something. This is making use of the metaphor *hare* 'scared person', and in the new dictionary it is marked as figurative and placed as a self-contained word meaning 2 in the article *1. hare*; in doing this we (also) present it as an example of polysemy. Similarly, idiomatic use of a noun is found in the sentence, *de hvide bier sværmer* [the white bees are swarming], which means 'it is snowing', thus implying that *hvid bi* [white bee] means 'snowflake'; this is also placed in a section of set phrases towards the end of the article *1. bi* in the dictionary.

9 Names of plays

Some collocations or part idioms (*Teilidiome*, cf. section 2) are used as names of children's plays and games, e. g. *lege / hviste / skyde krage* [play / throw / shoot a crow], which means to turn a somersault. *Tage ulv efter får* [play wolf after sheep] is the name of a catching play outdoors, whereas *hund efter hare* [dog chasing hare] and *ræv og gås* [fox and goose] are names of board games.

10 Mild oaths, subjunctives

Further, a few animal lexemes are used in mild oaths, e. g. the subjunctives *hunden sparke mig* [may the dog kick me], *katten rive mig* [may the cat scratch me], i. e. as a surrogate for *djævelen, fanden* [the devil] and other names for the devil in combination with some other verbs. These lexemes are also found in the genitive in phrases like, *det var hundens / kattens* [blame it on the dog / the cat], again as weak synonyms for the devil.

11 Literary proverbs

As a last point may be mentioned that most of the idioms with literary sources, including biblical phrases, are seldomly found in Jutlandic as an oral dialect.[10] For instance, you do not find the Standard Danish phrase, *den, der ager med stude, kommer også med* [he who is driving with bullocks will also get forward], i. e. ‘the slow or old-fashioned will also do’. But a similar idea is shown in the phrase, *den bagre ko kommer også til by, og somme tider allerbedst* [the rear cow in a herd will also come to town, and sometimes best of all], where the slowness is accepted and indeed praised a little. A biblical idiom like *en ulv i fåreklæder* [a wolf in a sheep’s clothing] is not found in dialectal speech because you cannot use *klæder* [clothes, clothing] figuratively here; the relation must be directly connected to the animal to be semantically acceptable.

12 Conclusion

We have seen that Jutlandic has a rich amount of phrasemes referring to domestic animals and other animals from the nearest environment (and the above mentioned are of course only a selection), so that they may be said to form a central part of the language and culture in this region. Many of these phrasemes are also found in Standard Danish according to the dictionary *Ordbog over det Danske Sprog* (Dictionary of the Danish Language), but several are unique to Jutlandic. It might be worthwhile to single these ones out, and the same could be done for the domestic animal phrasemes in *Ømålsordbogen* (the Dictionary of the Danish Insular Dialects). Or one could make comparisons to dialects of other standard languages, e. g. to *nynorsk* (New Norwegian), a standard language based on dialects, and perhaps also to Low German, because Middle Low German has had a great impact on Danish and Jutlandic. And generally, we may look for parallels to these phrasemes in other languages and dialects in the comprehensive investigation *Widespread Idioms in Europe and beyond* (cf. Piirainen 2012).

[10] Cf. Piirainen (2000: 71f.) as to the lack of *geflügelte Worte* (literally, ‘winged words’) in Westmünsterländisch and other dialects without literary sources.

Literature

ARBOE, Torben, 2003: Jyske kollektivformer – “stof-pluralis”. Akselberg, Gunnstein et al. (eds.): *Nordisk Dialektologi*. Oslo: Novus forlag. 235–247.

Dictionary of English Idioms, 2002. London etc.: Penguin Books.

Duden 11, 2002: *Redewendungen und sprichwörtliche Redensarten*. 2. Aufl. Mannheim etc.: Dudenverlag.

FEILBERG, Henning F., 1886–1914: *Ordbog over jyske almuesmål*. 1–4. København: Thieles Bogtrykkeri.

FRIEDRICH, Wolf, 1966: *Moderne deutsche Idiomatik*. München: Max Hueber Verlag.

HAUSMANN, Franz Josef, 2004: Was sind eigentlich Kollokationen? Steyer, Kathrin (ed.): *Wortverbindungen – mehr oder weniger fest*. Berlin/New York: Walter de Gruyter. 309–334.

Jysk Ordbog, 2000ff.: <www.jyskordbog.dk>. Aarhus: Peter Skautrup Centret for Jysk Dialektforskning, Aarhus Universitet.

KÜHN, Peter, 2007: Phraseologie des Deutschen: Zur Forschungsgeschichte. Burger, Harald et al. (eds.): *Phraseologie / Phraseology*. 2. Berlin/New York: de Gruyter. 619–643.

Ordbog over det danske Sprog, 1918–1956. 1–28. København: Gyldendal.

Oxford Collocations Dictionary for students of English, 2002. Oxford: Oxford University Press.

PIIRAINEN, Elisabeth, 2000: *Phraseologie der westmünsterländischen Mundart*. 1–2. Baltmansweiler: Schneider Verlag Hohengehren.

PIIRAINEN, Elisabeth, 2007: Dialectal phraseology: Linguistic aspects. Burger, Harald et al. (eds.): *Phraseologie / Phraseology*. 1. Berlin/New York: de Gruyter. 530–540.

PIIRAINEN, Elisabeth, 2012: *Widespread Idioms in Europe and Beyond: Toward a Lexicon of Figurative Units*. New York etc.: Peter Lang.

TOFTGAARD ANDERSEN, Stig, 2001: *Talemåder i dansk: Ordbog over idiomer*. 2. oplag. København: Gyldendal.

Ømålsordbogen: En sproglig-saglig ordbog over dialekterne på Sjælland, Lolland-Falster, Fyn og omliggende øer, 1992–. København: Nordisk Forskningsinstitut, Københavns Universitet; C. A. Reitzels forlag; Syddansk Universitetsforlag.

Colidioms. A Contribution to Cross-Cultural Research

ELENA BERTHEMET (Brest)

Abstract

This article is an overview of a multilingual phraseological database called *Colidioms*. It is designed to describe phraseological systems of different languages. That means that *Colidioms* is aimed to record meanings, as well as semantic and syntactic combinatorial properties of phrasemes. The aim of *Colidioms* is to reconcile the complex nature of phrasemes and compile a pertinent easy-to-use tool. The way to produce such tool is not evident, because many of different aspects have to be taken into consideration. *Colidioms* is based on recent technological advances. It combines tradition and innovation. It is a thin client application. As of today, however, it contains very few entries and still needs improvements, such as modifications to make it more user-friendly. Within the scope of this article, four points are considered. After the introduction, the central organizing principle, based on the concept of *notions* is studied. A more detailed description of *Colidioms* is given in sections 3 and 4, respectively. Finally, some conclusions are made based on our findings.

1 INTRODUCTION

The idea to build a multilingual phraseological database comes from the author's PhD thesis *Compared Lexicology of Phraseological Units* (*Zoomorphisms in French, English, German and Russian*). In this work, the task was to find phraseological equivalents in French, English, German and Russian and then to compare them. Two main methods for finding equivalents were used: the consultation of dictionaries and questioning of native speakers.

The first problem observed was the time waste when searching for an equivalent in monolingual and bilingual dictionaries. Often, the information found was fragmented and insufficient for the following two reasons:

(1) equivalents which were supposed to have the same meaning were often only partly equivalent;

(2) the proposed equivalents were removed from their contexts, lacking information about their usage, image and frequency.

The second problem was technical. In fact, the information gathered presented a great number of non-coordinated files. This coordination, or rather its absence, had three main shortcomings:

(1) it was time-consuming to find any information;

(2) the risk to forget or to miss something important;

(3) and, it was very difficult to collaborate with native speakers on these files.

So, it was necessary to elaborate a container where the information collected could find place. As we wanted to put all phrasemes into a single space, we needed an efficient organization method applicable to each phraseme of each language.

2 Central organizing principle, based on the concept of *notions*

The goal of the PhD thesis was to create a bridge between phrasemes in different languages. In order to link all the parts of the phraseme patchwork, we use *notions*. What do we mean by this term? In order to illustrate our proposal we have chosen five picturesque expressions meaning ‘to be extremely drunk’:

Language	Headword	Literal translation	Definition in source language
English	*(as) drunk as a skunk*	(as) drunk as a skunk	‘extremely drunk, to the point of behaving badly and without fear of repercussions’
Italian	*ubriaco fradicio*	wet drunk	‘completamente sbronzo’
Polish	*pijany jak bąk*	drunk as a bumblebee	‘zupełnie, kompletnie pijany’
Chinese	醉如烂泥	drunk as mud	‘喝醉了’
Japanese	鯨飲馬食	drink like a whale, eat like a horse	‘時折ある食べ過ぎ、飲み過ぎ’

Table 1: Phrasemes meaning ‘to be extremely drunk’.

As it can be observed, headwords as well as definitions are written in respective languages. All these phrasemes have approximately the same meaning, but a computer is not able to automatically find the link between them. To create this link between the five collocations, we labelled them with simple words. A number of notions are attributed to each phraseme. For example, the English collocation *(as) drunk as a skunk* is associated with notions *alcohol*, *behavior*, *drunkenness*, *indignity* and *misbehavior*:

Phraseme	Notions
(as) drunk as a skunk	*alcohol, behavior, drunkenness, indignity, misbehavior*
ubriaco fradicio	*alcohol, discomfort, drunkenness*
pijany jak bąk	*alcohol, drunkenness*
醉如烂泥	*alcohol, drunkenness, indignity, unconsciousness*
鯨飲馬食	*alcohol, drunkenness, excess, gluttony*

Table 2: Notions for *(as) drunk as a skunk*.

In order to guarantee compatibility and effectiveness of queries, notions are written in English. English was chosen due to its status as a very widely used international language. Notions are built on the empirical exploration of data. They are similar to descriptors in a thesaurus.

To simplify the choice of notions for the lexicographer and the search for the user, all notions are limited to the category of noun. For example, the adjective *unaware* (see Figure 3), in spite of the fact that it fits perfectly the English phraseme *(as) drunk as a skunk*, cannot be considered as a notion. Also, a group of words like *lack of awareness*, being composed from several words, cannot be used as a notion. Other notions like *negligence* and *disregard* are not appropriate to *(as) drunk as a skunk.*[1] In fact, it seems that *misbehavior* corresponds better.

Words that cannot be considered as notions	The reason
unaware	belongs to the category of adjectives
lack of awareness	is a combination of several words
negligence, disregard	do not correspond to the meaning

Table 3: Words that cannot be considered as notions.

[1] We aim to resolve this kind of issue during this research work.

It is difficult to determine foolproof notions:

(1) First, notions are not easy to state. Contrary to alphabetical order, which is neutral, the attribution of notions is a subjective, fastidious and time-consuming task.

(2) Second, since they are subjective, they risk being interpreted incorrectly and inadequately.

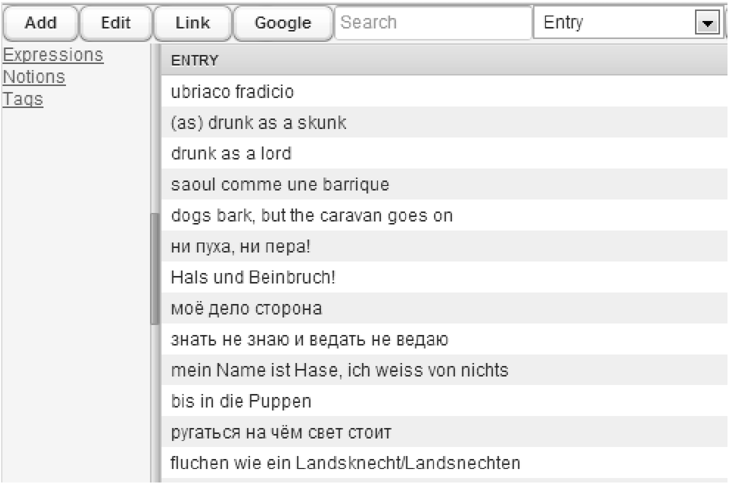
(3) Third, taking into account the phenomenon that notions have synonyms, it is not easy to define the minimal subset of notions.

Nevertheless, the use of notions homogenizes the corpus and orders the phraseological chaos in *Colidioms*. In order to illustrate how *Colidioms* works in practice, the following section is devoted to some aspects of its organization.

3 Lexicographer's view

Currently there are three different views in the application: *Expressions*, *Notions* and *Tags*. In the future version there will also be views to manage versioning, preferences (emails, language, and notifications), users, groups (students, experts, etc.), rights (admin, delete article, revert versions, etc.) and statistics.

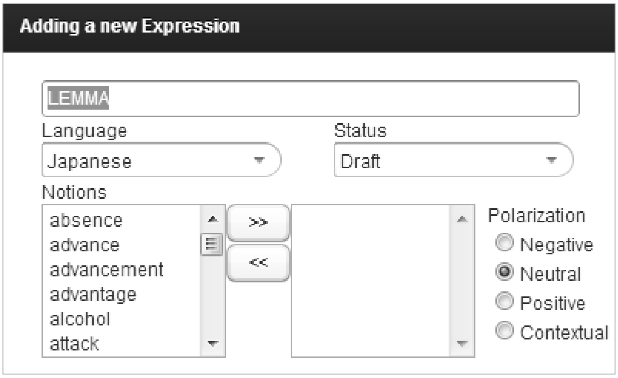
As it can be seen below, the *expression view* is the main view, it displays the last hundred expressions that have been added or modified:



ENTRY
ubriaco fradicio
(as) drunk as a skunk
drunk as a lord
saoul comme une barrique
dogs bark, but the caravan goes on
ни пуха, ни пера!
Hals und Beinbruch!
моё дело сторона
знать не знаю и ведать не ведаю
mein Name ist Hase, ich weiss von nichts
bis in die Puppen
ругаться на чём свет стоит
fluchen wie ein Landsknecht/Landsnechten

Figure 1: Phrasemes which have recently been added or modified.

To add a new expression the lexicographer clicks the corresponding *add* button in the top bar. The following field appears:



Adding a new Expression

LEMMA

Language: Japanese

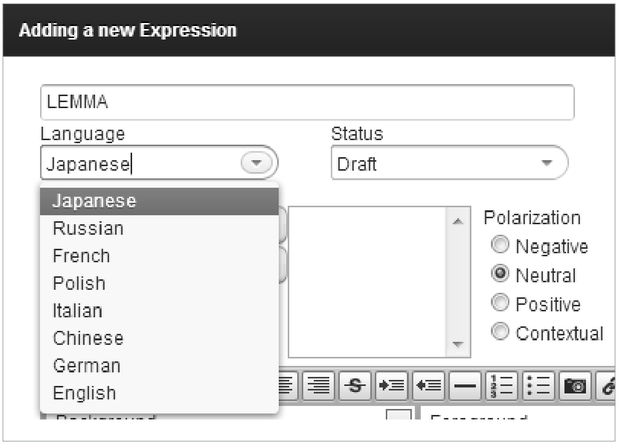
Status: Draft

Polarization: ○ Negative ● Neutral ○ Positive ○ Contextual

Figure 2: Adding a new phraseme.

As one can see, each phraseological article is identified by *lemma*, its initial, the most common form.

The next field is *language*. *Colidioms* supports eight languages: Chinese, English, French, Italian, Japanese, German, Polish and Russian. Multi-directional searches of phraseological equivalents are thus possible in any of these languages. This field appears as follows:



Adding a new Expression

LEMMA

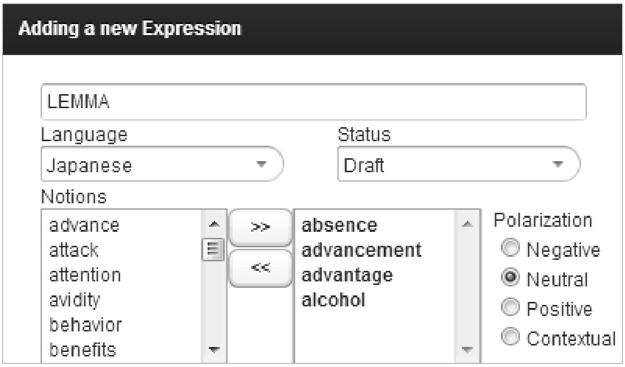
Language: Japanese

Status: Draft

Polarization: ○ Negative ● Neutral ○ Positive ○ Contextual

Figure 3: Languages supported by *Colidioms*.

After the language field, the lexicographer has a list of notions. He can choose an existing notion. Notions that are attributed to the phraseme are features in bold.



Adding a new Expression

LEMMA

Language: Japanese

Status: Draft

Notions

advance
attack
attention
avidity
behavior
benefits

absence
advancement
advantage
alcohol

Polarization

○ Negative
● Neutral
○ Positive
○ Contextual

Figure 4: List of available (left) and chosen (right) notions.

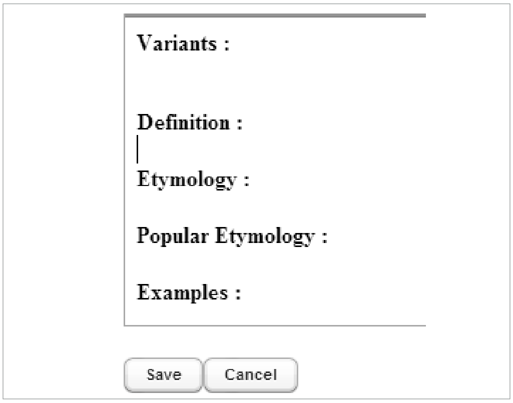
The list of notions is opened. The lexicographer can add a new notion. At present, notions are restricted to English. In the future version, it is planned 1) to propose a definition for each notion, 2) to offer a choice of synonyms for each notion, 3) to translate notions in order to give access to the tool to the people who don't speak English. Finally, it is planned to add a commentary field, containing syntactical position and combinatorial properties and also different usage constraints (stylistic, temporal, geographical).

On the right, *polarization* field appears. It can be observed that *polarization* can be *negative*, *neutral*, *positive* or *contextual*. It is true that on the one hand, polarization can depend on the context (for example, the French *copains comme cochons* 'friends like pigs' can be positive or negative) and in this case *polarization* field may appear useless. On the other hand, *polarization* could be useful to differentiate such expressions as *(as) drunk as a skunk* (negative polarization) and *drunk as a lord* (positive polarization).[2]

The bottom part is *a free form editor.* As long as the key words are retained, the application will recognize the content and will store the data accordingly.[3] The key words are *variants*, *definition*, *etymology*, *popular etymology*, and *examples*. Once completed, the lexicographer clicks save.

[2] The utility of *polarization* remains to be proven by further research.

[3] Most of the article can be copied and pasted from and to another application like *Microsoft Word*, *Open Office*, or *Outlook* for easy offline editing.



Variants :

Definition :

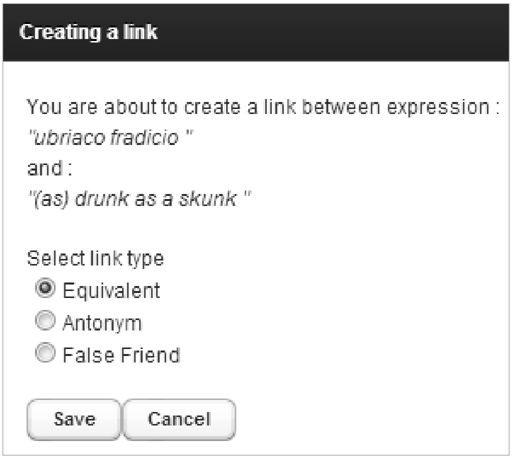
Etymology :

Popular Etymology :

Examples :

Figure 5: Expression view. Free form editor.

It is also possible to link two or more phrasemes. At present, there are three types of links: 1) equivalent,[4] 2) antonym, 3) false friend.



Creating a link

You are about to create a link between expression :
"ubriaco fradicio "
and :
"(as) drunk as a skunk "

Select link type
● Equivalent
○ Antonym
○ False Friend

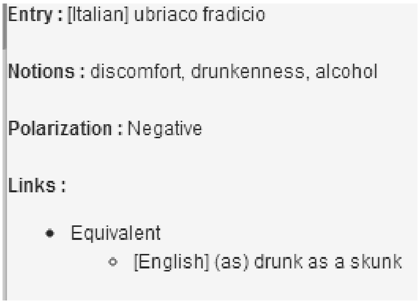
Figure 6: Creating a link.

To find a functionally adequate phraseme, the user clicks on the corresponding article displaying all information about the equivalent in question. The next section is devoted to the blueprint of an entry.

[4] The term *equivalent* is used here by convenience, because generally phrasemes have no complete equivalents, they are only approximations. Even identical equivalents might not fit the phraseme of the source language.

4 User's view

To help the user to find a functionally adequate phraseme, each article is rigorously structured as follows: *entry*, *notions*, *polarization*, *links*, *variants*, *definition*, *etymology*, *folk etymology*, and *examples*. As explained in the previous section, the first four fields are *lemma*, *notions*, *polarization* and *links*:



Entry : [Italian] ubriaco fradicio

Notions : discomfort, drunkenness, alcohol

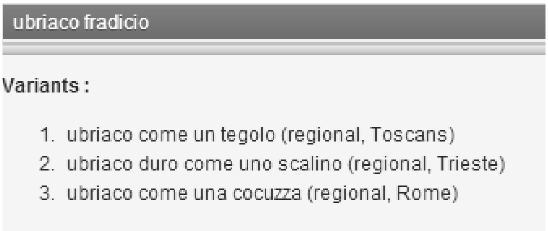
Polarization : Negative

Links :

- Equivalent
  - [English] (as) drunk as a skunk

Figure 7: First four fields of an entry: *lemma*, *notions*, *polarization*, and *links*.

As can be seen below, the Italian phraseme *ubriaco fradicio* 'wet drunk' has variants: *ubriaco come un tegolo* (regional, Tuscan) 'drunk like a tile', *ubriaco duro come uno scalino* (regional, Trieste) 'drunk as hard as a step is' and *ubriaco come una cocuzza* (regional, Rome) 'drunk as a pumpkin':[5]



ubriaco fradicio

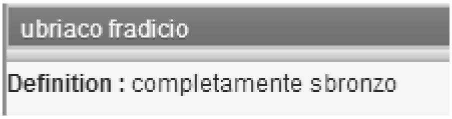
Variants :

1. ubriaco come un tegolo (regional, Toscans)
2. ubriaco duro come uno scalino (regional, Trieste)
3. ubriaco come una cocuzza (regional, Rome)

Figure 8: Geographical and dialectal variants for *ubriaco fradicio* (It.).

The next field is *definition*. Its function is to explain the actual meaning of the phraseme. In order to be as precise as possible, it is written in the original language of the phraseme:

[5] In fact, one of the main features of *Colidioms* is that it is flexible enough to account for geographical and dialectal variants within the same language.



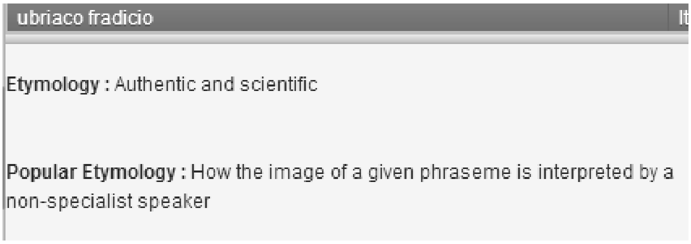
ubriaco fradicio

Definition : completamente sbronzo

Figure 9: Definition in respective language.

The definition is comprehensible for a non-professional, uses stylistically neutral words and contains relevant semantic information.[6]

After the definition, when it is known, the true, scientific, etymology is written. It is about the authentic, primary source of the phraseme:



ubriaco fradicio

Etymology : Authentic and scientific

Popular Etymology : How the image of a given phraseme is interpreted by a non-specialist speaker

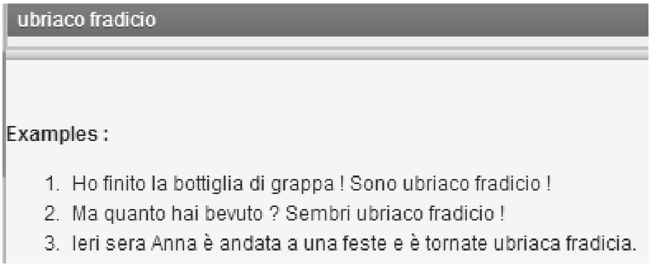
Figure 10: True and popular etymology.

Colidioms doesn't propose only the true etymology. In fact, it seems that to know how the phraseme' inner form is understood by speakers would help to understand and then to transmit in a source language all the various images and sentiments, contained within each phraseme. That is why the popular etymology is proposed next to the true one. Indeed, most native speakers interpret the phraseme in their own way, depending on their extralinguistic experience.

The last entry is *examples*. Two types of examples are proposed: 1) pedagogical examples and 2) corpus examples.

Pedagogical examples are selected and carefully chosen. They are minimal and concise and meant to help students understand the phraseme, rather than reproduce it.

[6] In the case when a phraseme has several meanings, all of them are indicated.



ubriaco fradicio

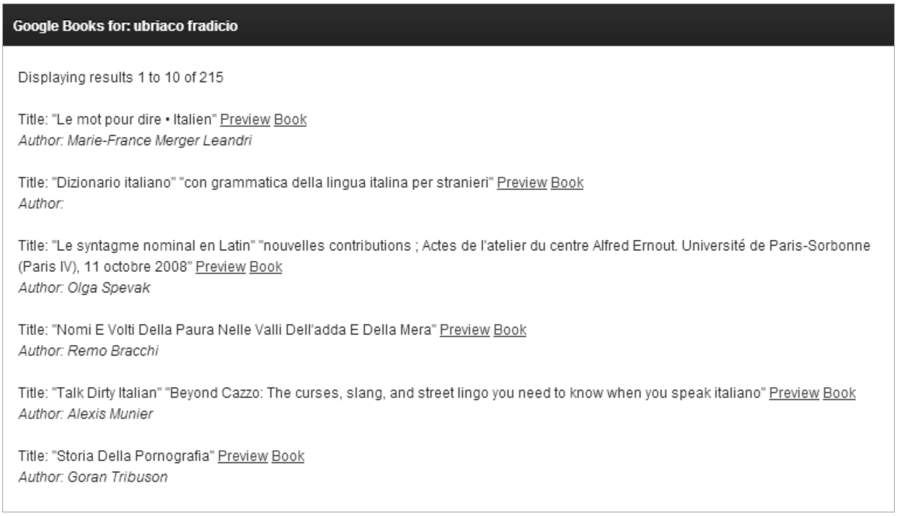
Examples :

1. Ho finito la bottiglia di grappa ! Sono ubriaco fradicio !
2. Ma quanto hai bevuto ? Sembri ubriaco fradicio !
3. Ieri sera Anna è andata a una feste e è tornate ubriaca fradicia.

Figure 11: Pedagogical examples.

In addition to the pedagogical examples, the user is provided with additional illustrative material: a link to *Google Books*. We distinguish two types of corpora: *Google Books* and *corpora*. In the current version, only the first one, *Google Books*, is available.

The application is connected to *Google Books* and displays *the Title* and *the Author* of the first ten books containing the searched-for phraseme:



Google Books for: ubriaco fradicio

Displaying results 1 to 10 of 215

Title: "Le mot pour dire • Italien" Preview Book
Author: Marie-France Merger Leandri

Title: "Dizionario italiano" "con grammatica della lingua italina per stranieri" Preview Book
Author:

Title: "Le syntagme nominal en Latin" "nouvelles contributions ; Actes de l'atelier du centre Alfred Ernout. Université de Paris-Sorbonne (Paris IV), 11 octobre 2008" Preview Book
Author: Olga Spevak

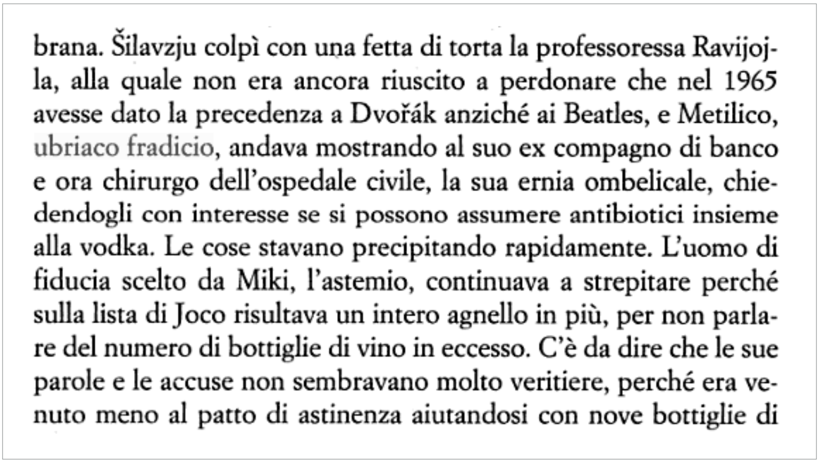
Title: "Nomi E Volti Della Paura Nelle Valli Dell'adda E Della Mera" Preview Book
Author: Remo Bracchi

Title: "Talk Dirty Italian" "Beyond Cazzo: The curses, slang, and street lingo you need to know when you speak italiano" Preview Book
Author: Alexis Munier

Title: "Storia Della Pornografia" Preview Book
Author: Goran Tribuson

Figure 12: Displaying results for *Google Books*.

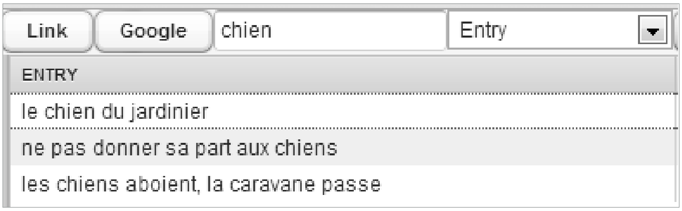
Selecting a book shows three possibilities depending on what the author or publishers allow them to do. Either: 1) *Google Books* displays the full book and the phraseme in question is highlighted, as shown in Figure 13, 2) parts of the book are displayed, 3) the user is notified that certain books contain the match.



brana. Šilavzju colpì con una fetta di torta la professoressa Ravijojla, alla quale non era ancora riuscito a perdonare che nel 1965 avesse dato la precedenza a Dvořák anziché ai Beatles, e Metilico, ubriaco fradicio, andava mostrando al suo ex compagno di banco e ora chirurgo dell'ospedale civile, la sua ernia ombelicale, chiedendogli con interesse se si possono assumere antibiotici insieme alla vodka. Le cose stavano precipitando rapidamente. L'uomo di fiducia scelto da Miki, l'astemio, continuava a strepitare perché sulla lista di Joco risultava un intero agnello in più, per non parlare del numero di bottiglie di vino in eccesso. C'è da dire che le sue parole e le accuse non sembravano molto veritiere, perché era venuto meno al patto di astinenza aiutandosi con nove bottiglie di

Figure 13: *Google Books* displaying a book extract with the phraseme highlighted.

Not all these features are needed for each user, but it is not a problem, because the program allows the user to decide which information to consult. By completing all these points, the lexicographer offers to the user a maximum of information about each phraseme. The user can find phrasemes in two ways. The first, *semasiological search*, is searching by lemma. The example below shows three phrasemes containing the word *chien* in the entry:



chien

Entry

ENTRY

le chien du jardinier

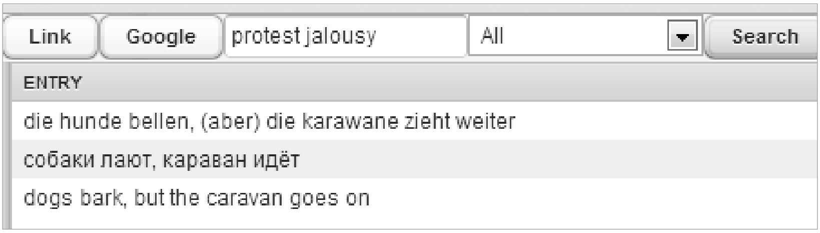
ne pas donner sa part aux chiens

les chiens aboient, la caravane passe

Figure 14: Semasiological search.

The second type of search, *onomasiological search*, is useful when the user cannot quite remember the expression exactly. It is about making complex queries based on the notions. For example, the user remembers only that there is an expression referring to *protests* and *jealousy*. He can type *protest* and *jealousy* and the software will return a list of results containing these notions.[7]

[7] It is possible to make one-notion and multi-notions queries.



protest jalousy

ENTRY
die hunde bellen, (aber) die karawane zieht weiter
собаки лают, караван идёт
dogs bark, but the caravan goes on

Figure 15: Onomasiological search.

Thus, *Colidioms* helps the user to make an informed choice for the phraseme he wishes to translate.

5 Concluding remarks

The decision to enter all this information in the same database turns out to have many advantages. First of all, notions seem to be the most important component of *Colidioms* as they make cross-lingual comparisons possible. The use of notions enables users to find expressions without knowing any of their key-words. Second, *Colidioms'* architecture makes it possible to inventory phrasemes and describe them systemically. It shows semantic relationships between phrasemes and registers the diversity of variants. Third, the software is descriptive and prescriptive at the same time. It provides a uniform detailed description of phrasemes, thereby reducing the possibility of wrong usage or conclusions. Finally, in *Colidioms* it is possible to incorporate other search engines in order to automatically explore online corpora.

The work is unfinished, and a lot of questions remain to be answered. We hope that *Colidioms*'s way of describing phrasemes is sufficient to accurately retain them and to reflect their real use.

Corpus-Driven Phraseology Assessment: an Experiment

JEAN-PIERRE COLSON (Brussels/Louvain-la-Neuve)

Abstract

Most studies devoted to the use of phraseology by non-native speakers have shown that the situation is very complex, because there may be underuse of some structures but also overuse of other ones. The borderline between phraseology in the broad sense and grammatical but unexpected constructions is particularly striking among advanced learners. This paper presents the results of an experiment carried out in collaboration with the Centre of English Corpus Linguistics at Louvain University. All n-grams (from bigrams to sixgrams) were extracted from the 2.6 million word ICLE corpus (University of Louvain) and from a comparable portion of 2.6 million words randomly selected from the BAWE corpus (University of Warwick), and the proportion of n-grams on a benchmark of 11,000 collocations was computed. The results indicate a very marked difference between the ICLE corpus and the native BAWE corpus and are therefore consistent with previous work on non-native English phraseology.

1 INTRODUCTION

Most foreign language learners have experienced the constant frustration caused by phraseology: no matter how gifted they are or how hard they try, they will after many years of considerable effort hardly reach the level of native competence. Simply on the basis of grammar and vocabulary, they will in many cases produce uncommon or unnatural utterances and sentences.

A commonly used explanation for this situation is that learners have an imperfect mastery of phraseology in the broad sense. It is not an easy task, however, to assess in what precise way this will be reflected in learner corpora, nor to use phraseology as a criterion for the assessment of learner production.

In this paper we report the results of an experiment carried out in collaboration with the *Centre of English Corpus Linguistics* at *Louvain University*. We will argue that translation corpora and advanced learner corpora play a key role in this debate. Traditional studies have indeed shown that those corpora are lacking in phraseology. Efficient methods for the extraction of phraseology at all levels should therefore be able to measure a different concentration of phraseology in learner corpora on the one hand, and in native corpora on the other.

2 Theoretical backgrounds

For the purpose of this paper, we will use phraseology "in the broad sense" (Burger 1998, Čermák 2007), covering all multi-word expressions (MWEs) that are common and fixed in a given language community.

Previous studies have attempted to provide an accurate description of phraseology in learner corpora: Granger (1998), Nesselhauf (2005), De Cock (2004, 2007), Barfield/Gyllstad's (2009), Paquot (2010). The picture that emerges from those studies is, however, far from being clear: what is apparently at stake is a complex interplay of overuse of some structures (e.g. frequent collocations and formulae) and underuse of others (especially less common collocations, phrasemes, and in general structures differing a lot from the learner's native language).

The literature also suggests that the language production of advanced learners may be of particular interest in this debate. Indeed, beginners will tend to rely too much on basic collocations or on the phraseology of their mother tongue, which makes it hard to draw the line between errors pertaining to vocabulary, grammar or phraseology. Advanced learners, on the other hand, have a much broader mastery of vocabulary and grammar but also of phraseology, and their production may therefore reveal much more clearly which subtle differences remain between native and non-native use of MWEs.

As far as the theoretical underpinnings of phraseology are concerned, research on learner corpora also lies at the crux of the matter. They indeed provide a very useful bottom-up approach to the identification of what phraseology is all about: which categories of MWEs will be absolutely necessary in the various stages of the learning process and maybe in which order, how do we define the scope of phraseology as opposed to grammar, are communicative MWEs more useful than purely idiomatic ones?

Without sufficient theoretical insights into the very nature and use of phraseology, it can be a daunting task to explain to students why their language production is not idiomatic enough. This is illustrated by the following two excerpts from comparable essay corpora. The first one, written by a non-native, comes from the *International Corpus of Learner English* (ICLE, Université catholique de Louvain), and the second from the *British Academic Written English Corpus* (BAWE, University of Warwick):

(1) Practical work will help us the time we need to work really, the next thing we will be sitting there with our degrees not knowing how to practice our knowledge and skills.

(2) He ran to keep the lift open for his wife and sister in law and the next thing he knew he was laying on the hospital foyer floor.

These few lines will serve to illustrate that in many cases, it is not so easy to distinguish the production of advanced learners as in (1) from that of natives as in (2). Indeed, (1) contains just one or two less common combinations. For instance, *need to work really* is problematic. If we look it up in a fairly large corpus (in this case a 330 million word corpus, part of the *ukWacky corpus*, Baroni et al. 2009), we only get 1 single occurrence, as opposed to 1184 for *have to work*, a much more common grammatical and communicative phrase.

The theoretical remark would therefore seem to be that frequency plays a part in idiomaticity and may help recognizing the lack of phraseology in language production. But if we now consider excerpt (2), we find a clear example of a MWE, *the next thing he knew*. If we check the frequency of this communicative formula in an even larger English corpus of 2 billion words, the *ukWacky corpus* (Baroni et al. 2009), we only find 11 occurrences in this exact form, and 97 for the formula in the first person (*the next thing I knew*). This confirms previous studies on the very scarce presence of many MWEs, especially phrasemes, in large corpora (Moon 1998; Colson 2007, 2008).

As frequency is not the sole criterion for recognizing phraseology in language production, a more promising method is to start from benchmarks of MWEs that are extracted from corpora but also checked by native speakers. This enables the researcher to assemble an objective collection of phrases that may be characteristic of a given domain or activity such as the production of essays.

It should also be pointed out that relying on the dictionary alone can be misleading when trying to measure phraseology in language production, because the dictionary rarely mentions in which precise variants, persons or tenses a MWE will be commonly used. This point will be crucial in identifying native vs. non-native production. Consider for example the following phrases with *go* that were, among many others, extracted automatically (and then manually checked) from the previously mentioned *ukWacky corpus*:

> *go hand in hand with*, *go it alone*, *go some way towards*, *go to great lengths to*, *how do you go about*, *there is a long way to go before*, *the list could go on*, *watch the world go by*, *things can go wrong*, *go that extra mile*, *would go so far as to*.

From the viewpoint of a general theory of phraseology, large linguistic corpora again pose the question of the classification of MWEs. Indeed, these few phrases are not always easy to situate between formulae (*would go so far as to*), phrasemes (*go to great lengths to*) or clichés (*watch the world go by*, *go that extra mile*). Also, the corpus reveals that most MWEs display many morpho-syntactic restrictions. *Go that extra mile*, for instance, is sometimes used in the past tense (*he went that extra mile*, *have gone that extra mile*), or in the gerund (*going that extra mile*), but the *ukWacky corpus* clearly shows that it is more frequent in the infinitive form: *have decided to go that extra mile, time to go that extra mile*, *we will go that extra mile*.

For all these reasons, corpus-based research on phraseology should derive maximum benefit from a bottom-up perspective, extracting combinations as they appear in large corpora and not on the sole basis of dictionary entries or pre-existing lists. The method therefore consists in assembling a huge collection of various types of MWEs, a "benchmark" of common and very fixed structures used by native speakers of a language.

Collecting such MWEs on a large scale is one of the crucial issues in corpus-based phraseology, as a manual method is time consuming and partly subjective. Within the fields of corpus linguistics and computational linguistics, there is now a growing body of literature on the thorny question of the automatic extraction of all MWEs (often called collocations in the broadest sense) from a corpus. A thorough discussion of all the statistical scores that have been used by those studies falls beyond the scope of the present paper. Suffice it to say that most of them (for an overview, see Evert 2004) are limited to two-word combinations, the "bigrams", that the theoretical starting point (the statistical foundation of word distributions) remains complex, and that the overall efficiency of the method has been seriously questioned. Evert and Krenn (2001) have shown that the results obtained by those statistical scores often come close to those gained by the observed absolute frequency in a corpus. As shown by Ďurčo (2010), there is also much variation between the results yielded by different scores.

3 Methodology

Within the framework of this experiment, a new method was used for the automatic extraction of a benchmark of multi-word expressions: the *PR*-score (Proximity score), corresponding to the average distance between the element parts of the multi-word expression. A full description is found in Colson (2010) for the Web and has been adapted here to be used on large text corpora. This method is not directly related to statistics, but rather to Information Retrieval, in which different clustering methods have been proposed for the extraction of semantically related words. Specifically, “metric clusters” (Baeza-Yates/ Ribeiro-Neto 1999) have been computed by measuring the distance between words in a document, an approach related to the methodology used here. As it is still experimental, a manual check was carried out in order to ensure that the benchmark consisted of common MWEs used by native speakers.

As a first step, a reference corpus was used (the academic part of the *British National Corpus*, 15 million words) in order to extract as many MWEs as possible. All n-grams (combinations of n words) ranging from bigrams (2 words) to sixgrams were extracted by a computer program (*Perl script*). The n-grams were then passed to a 200 million word corpus, the first section of the *ukWacky corpus* (Baroni et al. 2009). The absolute frequency of all n-grams was computed, as well as their *PR*-score: the average distance between the grams, given the window of a paragraph in which they occur. A manual check was carried out in order to eliminate irrelevant results.

The frequency thresholds were set experimentally according to the length of the n-grams: the combinations had to display at least 50 occurrences in the 200 million word corpus for bigrams, 10 occurrences for trigrams and four-grams, and 4 for fivegrams and sixgrams. This choice is justified by the sharp decrease in absolute corpus frequencies if the size of the n-gram is larger.

In our experiment, two comparable academic corpora were analysed by means of the benchmark of common MWEs mentioned above:

– a non-native corpus of essays: the 2.6 million word ICLE corpus (*International Corpus of Learner English*, University of Louvain)[1]

– a native corpus of essays: a portion of 2.6 million words randomly selected from the BAWE corpus (*British Academic Written English*, University of Warwick).[2]

[1] *The International Corpus of Learner English* (Université catholique de Louvain, Centre for English Corpus Linguistics, http://www.uclouvain.be/en-cecl-icle.html) is a collection of essays written by higher intermediate to advanced learners of English from several mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, Turkish).

[2] *The BAWE corpus* (University of Warwick, Centre for Applied Linguistics, http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/) contains a collection of proficient assessed student writing across four levels of study (undergraduate and taught master’s level).

Our main research question was: is there more overlap between the native corpus and the benchmark than between the learner corpus and the benchmark? In other words, will this objective benchmark of common MWEs reveal that there is more phraseology in the essays written by native as opposed to non-native speakers?

4 Results and discussion

The benchmark that was assembled in this way consisted of a total of 11 453 MWEs belonging to the domain of academic English. They included 5616 bigrams, 2770 trigrams, 1541 fourgrams, 1178 fivegrams, and 348 sixgrams. Below are some examples from the list (a few bigrams, trigrams and fourgrams that were automatically extracted by the algorithm).

> Bigrams: *anecdotal evidence, careful consideration, civilian population, closely tied, closer inspection, decision makers, devastating effect, enabling us, enormity of, enormous amount, experimenting with, flurry of, ideally suited, inevitable consequence, labour intensive, main thrust, major obstacle, notable exceptions, randomly selected, rigors of, sizeable proportion, stark contrast, successfully completed, thus avoiding, touched upon, united front, what emerges.*
>
> Trigrams: *achieve these goals, acutely aware of, alive and well, all reasonable steps, all walks of, allow him to, an early age, an equal number, an essential component, an excellent example, an extreme case, an imbalance between, an increasing emphasis, an informed decision, an open letter, an unfair advantage, are equally applicable, as yet unknown, at our disposal, at the forefront, attempts to explain, be better served, before embarking on, behind the scenes, bribery and corruption, bringing the total, buyers and sellers, calculated according to, call into question, came into conflict, can be gleaned, cannot be guaranteed.*
>
> Fourgrams: *against the wishes of, aided and abetted by, all intents and purposes, an equal chance of, answers to these questions, as quickly as possible, at a particular moment, at any given moment, at any time during, at opposite ends of, at the lower end, before the expiration of, both formal and informal, came to the fore, can only be achieved, cannot be held responsible, cannot help feeling that, cast doubts on the, come to the fore, comes to an end, comes to the conclusion, comparisons to be made, concentrated in the hands, considerable period of time, consideration should be given, contexts in which they, depend very much on, distinction is drawn between, does not intend to, emphasis was placed on, emphasized the importance of, even went so far, events leading up to, examined in more detail, failed to comply with, find out more about, for no better reason.*

There can be little doubt that the words contained in the examples above would be immediately recognized by native speakers of English as “belonging together” or as “frequently combined with each other”. This is precisely where phraseology in the broadest sense comes in: all types of associations that would be used by natives and far less often by foreign language learners. In the bigram list above, for instance, the common collocation *flurry of* was automatically extracted from the corpus by the algorithm. If we check out the paragraphs in which this MWE is used, we come across a variety of contexts such as: *a flurry of irate responses*, *a flurry of migrants*, *a flurry of revelations*, *a flurry of interest*, etc.

The observed absolute frequency of this MWE in the *ukWacky corpus* (200 million) words is, however, not so high for a bigram (320 occurrences), but the clustering algorithm makes it clear that this is indeed a very fixed combination. The situation is even clearer in the case of a fourgram such as *emphasis was placed on*: this MWE receives an absolute frequency of 18 in the 200 million word corpus, but its association is very high as well. If we check *emphasis was put on* in the same corpus, we get an absolute frequency of only 6, but also a very high degree of association.

As stated above, the main research question of this experiment was whether there is more overlap between the native corpus of essays (2.6 million words) and the benchmark than between the non-native corpus of essays (2.6 million words as well) and the benchmark. The results are shown in table 1. They represent, for each of the two corpora, the percentage of n-grams (from bigrams to sixgrams) that were also present in the benchmark. This overlap was computed programmatically.[3]

	BI-GRAMS	TRI-GRAMS	FOUR-GRAMS	FIVE-GRAMS	SIX-GRAMS	TOTAL
Benchmark	5 616	2 770	1 541	1 178	**348**	**11 453**
Overlap ICLE / Benchmark	**72%**	**62%**	**52%**	**38%**	**29%**	**51%**
Overlap BAWE / Benchmark	**89%**	**79%**	**67%**	**49%**	**35%**	**64%**

Table 1: Overlap between the two corpora of essays and the benchmark.

The results of this experiment are consistent with the findings of previous studies on the use of phraseology by advanced non-native speakers of English (De Cock 2007, Paquot 2010): the overall percentage of overlap is much higher between the native *Bawe corpus* and the benchmark (64 percent) than between the non-native *ICLE corpus* and the benchmark (51 percent). It is also striking

[3] A *Perl script* (JP Colson) was used, in which the different lists of grams were passed to an array and a hash algorithm.

that the difference decreases proportionally with the size of the n-grams: the difference between the percentage for natives on the one hand and non-natives on the other, is precisely 17 percent at the level of both bigrams and trigrams, but it goes down to 15 percent for fourgrams, 9 percent for fivegrams, and only 6 percent for sixgrams. This is of course partly due to the statistical decrease in absolute frequency, as the number of bigrams is much higher than that of sixgrams at the other end of the spectrum. From a theoretical point of view, however, the matter may be more complex and related to the phraseological density that each language displays at the respective n-gram levels.

Obviously, these are only partial results and should be confirmed by further studies, but they contain a strong indication that even advanced learners use far less MWEs in the broad sense than natives do. This experiment also suggests that a bottom-up and corpus-driven approach might be used in the future for the automated assessment of phraseology in learner essays or translations.

References

BAEZA-YATES, Ricardo / RIBEIRO-NETO, Berthier, 1999: *Modern Information Retrieval*. New York: ACM Press, Addison Wesley.

BARONI, Marco et al., 2009: The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation* 43, 209–226.

BARFIELD, Andy / GYLLSTAD, Henrik (eds.), 2009: *Collocating in another language: Multiple interpretations*. Basingstoke: Palgrave Macmillan.

BURGER, Harald, 1998: *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt Verlag.

ČERMÁK, František, 2001: Substance of idioms: perennial problems, lack of data or theory? *International Journal of Lexicography* 23, 1–20.

COLSON, Jean-Pierre, 2007: The World Wide Web as a corpus for set phrases. Burger, Harald et al. (eds.): *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research. Volume 2*. Berlin/New York: de Gruyter. 1071–1077.

COLSON, Jean-Pierre, 2008: Cross-linguistic phraseological studies: An overview. Granger, Sylviane et al. (eds.): *Phraseology. An interdisciplinary perspective*. Amsterdam/Philadelphia: Benjamins. 191–206.

DE COCK, Sylvie, 2004: Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures* 2, 225–246.

DE COCK, Sylvie, 2007: Routinized Building Blocks in Native Speaker and Learner Speech: Clausal Sequences in the Spotlight. Campoy, Mari Carmen / Luzón, María José (eds): *Spoken Corpora in Applied Linguistics*. Bern: Peter Lang. 217–233.

ĎURČO, Peter, 2010: Einsatz von Sketch Engine im Korpus – Vorteile und Mängel. Ptashnyk, Stefaniya et al. (Hrsg.): *Korpora, Web und Datenbanken / Corpora, Web and Databases*. Baltmannsweiler: Schneider Verlag Hohengehren. 119–131.

EVERT, Stefan, 2004: *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Stuttgart: IMS, University of Stuttgart.

EVERT, Stefan / KRENN, Brigitte, 2001: Methods for the Qualitative Evaluation of Lexical Association Measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse. 188–195.

GRANGER, Sylviane, 1998: Prefabricated patterns in advanced EFL writing: collocations and formulae. Cowie, Anthony (ed.): *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press. 145–160.

MOON, Rosamund, 1998: *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.

NESSELHAUF, Nadja, 2005: *Collocations in a learner corpus*. Amsterdam: Benjamins.

PAQUOT, Magali, 2010: *Academic vocabulary in learner writing. From extraction to analysis*. London/New York: Continuum.

La synonymie collocationnelle. Entre corpus et dictionnaire bilingue

Cosimo De Giovanni (Cagliari)

Abstract

The aim of this article is to revise the relation both between synonymy and collocation and between corpus and bilingual dictionary. Examples of synonymic collocations, at the level of base (A) and at the level of collocate (B), with a particular attention to two types of syntactic patterns VERB + NOUN and NOUN + ADJ, will be necessary to demonstrate the difference between corpus evidence and lexicographic treatment (French/Italian dictionaries). Therefore, the final aim of this article is to give new tracks for the creation of a new typology of bilingual dictionary with a holistic vision of language.

1 Introduction

La notion de collocation est une des questions épineuses propre au domaine de la lexicographie, surtout de celle bilingue. La collocation, comme le remarque Williams (2003: 33), représente un personnage invisible, voire un fantôme, mentionné souvent, mais qui n'apparaît qu'une seule fois en chair et os. Elle est au centre d'une intrigue lexicale: tout le monde l'utilise, mais personne ne la connaît véritablement. C'est donc un phénomène aux contours flous se cachant sous différentes dénominations imposées par les diverses théories dans la tentative d'en fournir une description formelle.

Ce qui fait la différence entre une collocation textuelle et une collocation lexicographique, c'est surtout son rôle dans la construction du sens. La première typologie est repérable au sein des textes, tandis que la seconde typologie ne concerne que des structures aseptiques nécessaires aux lexicographes pour la construction des entrées dictionnairiques. De plus, l'existence d'une myriade de théories autour de la collocation rend moins évident le fait que le phénomène collocationnel demeure néanmoins quelque chose de plus complexe qu'un simple mécanisme de cooccurrence de lexies. En effet, les collocations rentrent dans le mécanisme du choix conditionné d'une langue qui n'exclut pas l'autre principe, celui du choix ouvert: dans toute langue les deux principes coexistent. Une vision holistique de la langue devient possible même dans le domaine de toute lexicographie.

Le présent article veut faire le point sur la nature et le rôle des collocations au sein des dictionnaires bilingues (dorénavant DB) en présence des possibles structures synonymiques. Pour cette raison, il est nécessaire de vérifier qu'il y ait même dans le discours, une manifestation de la synonymie collocationnelle (dorénavant SC) en faisant appel aux analyses proposées par la linguistique de corpus.

Notre but est celui de revisiter la relation existant entre collocation et synonymie, ainsi que leur présence dans le corpus et dans le DB. Des exemples de SC seront proposés afin de démontrer la différence entre les données repérées du corpus et celles contenues dans les DB (domaines français et italien), surtout pour les structures du type VERBE + NOM et NOM + ADJECTIF.

2 Bases théoriques

2.1 Définition et nature de la collocation

La définition de collocation ne fait pas l'unanimité des linguistes et des lexicographes. L'existence d'une multiplicité de définitions prouve, en effet, qu'il n'y a pas d'entente entre les spécialistes. En plus, les deux approches, lexicographique et linguistique, donnent une définition différente à partir de la nature formelle de la collocation: une structure binaire pour la première approche et une structure plus ouverte, textuelle, pour la deuxième approche. Ainsi, le vrai problème réside-t-il surtout dans la terminologie employée pour définir le phénomène. Dans la majeure partie des cas, on est obligés d'associer la notion de collocation à celles de combinaison, cooccurrence, association, rapprochement, placement, ensemble d'éléments et également à celles de cooccurrence privilégiée, association habituelle, relation potentielle, rapprochement fréquent, rapprochement arbitraire. Tous les termes que nous venons d'énumérer constituent des candidats potentiels pour décrire la notion de collocation, mais pour en avoir une définition satisfaisante il faut déceler sa nature et son fonctionnement à l'intérieur de la langue et du discours. Or, si d'un côté on conçoit la collocation comme une mise en relation entre ses deux éléments, la *base* et le *collocatif*, manifestant sa force syntaxico-sémantique interne, de l'autre il faut l'imaginer à l'intérieur d'un texte et chercher à déceler ses relations avec son environnement lexical (force cohésive externe). Cela permet finalement de mettre les deux approches l'une à côté de l'autre afin de fournir une description plus détaillée de la structure collocationnelle.

Partons du principe, déjà dicté par Sinclair (1996: 81), qu'en parlant une langue, l'usager potentiel a la possibilité de faire des choix au niveau syntaxico-sémantique, mais aussi au niveau rhétorique et informationnel (Muller 2008: 27), à savoir un choix libre (qui est presque rare étant donné qu'il est toujours assujetti à des règles grammaticales bien précises), un choix ouvert (qui est toujours soumis à des pseudo-règles à caractère psychologique) et un choix conditionné (qui règle l'utilisation de structures préfabriquées telles que les collocations, les locutions etc.). Or, il est évident qu'il est presque impossible de concevoir la production de structures langagières complexes sans tenir compte du contexte, dans son sens large, et du rôle des acteurs dans l'acte communicatif. Il est évident de même que la collocation contribue, plus que le mot, à produire du sens et à faire du texte une texture. La phrase de Firth (1957: 11) «you shall know a word from the company it keeps» résume l'essentiel de la collocation, c'est-à-dire le fait qu'elle ne peut pas être considérée un mot, au sens strict du terme, si elle ne s'accompagne pas à d'autres mots de sorte que le sens est partagé par tous les éléments de la structure.

Or, ce qu'il faut faire c'est justement d'interposer entre une organisation syntagmatique de la langue, sur la base d'un principe d'intégration, et une organisation syntaxique, sur la base d'un principe de dépendance (Muller 2008: 29), une organisation phraséologique, comprenant les collocations et toutes les autres structures idiomatiques, régie par un principe d'agencement qui n'est pas seulement de type linguistique et qui ne suit pas nécessairement un ordre bien précis. Les trois organisations s'accommodent de façon différente d'une langue à l'autre. Prenons l'exemple des structures comme les colligations constituées d'un élément plein ou lexical et d'un élément vide ou grammatical, qui ont une organisation différente selon la langue de référence: en français, elles se placent entre une organisation syntagmatique et une organisation syntaxique; en italien, au contraire, elles rentrent dans une organisation de type syntagmatique. Les trois organisations ne demeurent pas seulement régies par des choix de type linguistique, mais elles concernent aussi, comme on l'a déjà vu, la façon d'organiser l'information et le rôle des acteurs de la communication. Le côté de la langue se montre toujours indissociable du côté du discours.

L'agencement constitue, donc, le principe réglant l'organisation phraséologique de la langue. Il n'est qu'un mécanisme d'arrangement résultant d'une combinaison. Dans le cas d'une collocation, l'agencement concerne plutôt le mécanisme qui préside à sa formation en tant que simple structure semi-figée, binaire, à caractère arbitraire, dont les composants sont dissymétriques. L'agencement est une manifestation mutuelle entre les éléments d'une collocation: les termes de matrice firthienne de «collocabilité» (Firth 1957: 194) et d'«attente mutuelle» (Firth 1968: 181) expliquent, en effet, la notion que nous avons adoptée. Mais ce mécanisme ne suffit pas à décrire la collocation. C'est pour cette raison que nous optons pour un autre terme, celui de contexture, ou bien l'ensemble de relations organisées entre des éléments significatifs formant un tout complexe et organique. Dans ce cas, ce qu'il faut prendre en compte c'est l'ensemble des relations que tous les éléments concourent à établir – réseau collocationnel (Williams 1998) – en identifiant leurs fonctions à l'intérieur du texte – résonance collocationnelle (Williams 2006), préférence sémantique, prosodie sémantique (Louw 1993) – pour identifier le phénomène collocationnel. Cette dernière explication éclaire encore mieux les théories de Firth sur l'analyse de la langue à travers les structures collocationnelles.

Donc, la présence du texte devient fondamentale pour l'analyse et la description du phénomène collocationnel. De même, la linguistique de corpus, s'appuyant sur l'analyse du texte, est l'observation directe du monde, elle est une «approche aux problèmes du monde» (Williams 2010: 403) qui s'accompagne du jugement de l'homme. Par conséquent, la collocation est l'intérêt principal de la linguistique de corpus. C'est pourquoi il est inadmissible d'opérer une division nette entre la tendance lexicographique et celle linguistique dans l'étude de la collocation: elles ne s'opposent pas, mais elles analysent le phénomène sous deux angles différents. Il est fondamental, pour les deux approches, de travailler sur les points communs pour atteindre le même but.

2.2 Collocation et synonymie: la synonymie collocationnelle

La synonymie est un phénomène qui a toujours intéressé les linguistes et les sémanticiens en particulier. Elle reste en partie un mystère, surtout si l'on considère le fait que la relation synonymique comporte la présence de mots ou ensemble de mots qui sont sémantiquement semblables tout en ayant des significations similaires. Ce qu'il est évident c'est le fait que l'existence d'une relation synonymique parfaite entre deux mots est presque improbable. Autrement dit, ce qui fait la différence entre deux mots qualifiés comme des synonymes c'est la connotation véhiculée par les mêmes mots et le contexte dans lequel ils sont employés. Il existe des dictionnaires qui tiennent compte de cette possible divergence entre deux mots synonymes en utilisant des éléments de discrimination (marque de registre en particulier), par exemple *enfant* fait partie du français standard et *gosse* sera marqué comme mot familier.

Essayons ici de créer une relation entre la notion de collocation et celle de synonymie en utilisant le terme de synonymie collocationnelle (SC). Pour ce faire, nous partons du constat lexicographique de collocation conçue comme une structure binaire où l'élément dominant, la *base*, garde son autonomie sémantique et l'élément dominé, le *collocatif*, change son sens en fonction du mot-*base*. Supposons, en même temps, que chaque élément de la structure puisse être remplacé par un autre élément sur le plan paradigmatique, nous sommes alors face à un cas de synonymie collocationnelle interne (ou SCI). L'autre cas est celui d'une synonymie collocationnelle externe (ou SCE) où les deux éléments de la structure peuvent être remplacés par une autre structure similaire. Les deux concepts que nous venons d'évoquer ne sont pas bien clairs et leur notion apparaît assez floue. Est-ce que les deux combinaisons de *pluie très violente* et *pluie battante* relèvent du cas d'une SCI {*pluie*} {*très violente*} ≈ {*pluie*} {*battante*} ou d'une SCE, {*pluie très violente*} ≈ {*pluie battante*}? Une réponse à cette question pourrait aider le lexicographe à un traitement pertinent des combinaisons qui présentent des cas similaires. Pour l'instant, ce qui est nécessaire, à notre avis, c'est de donner une première définition de SC pour poursuivre notre analyse.

Une combinaison AB est synonyme d'une autre combinaison A'B' *si et seulement si* se réalisent les conditions suivantes: 1. au niveau du collocatif (B) dans un cas de SCI; 2. au niveau de la base (A) dans un cas de SCI; au niveau de la structure AB dans un cas de SCE.

Nous laissons de côté le cas de SCE et nous nous concentrons sur les deux cas de SCI.

Dans la majeure partie des cas, c'est le *collocatif* qui dispose d'un ou plusieurs synonymes: des cas comme *éperdument amoureux* et *follement amoureux*, ou *hiver rigoureux* et *hiver rude* ne sont pas rares. Il arrive aussi qu'un seul collocatif compte trois ou plusieurs synonymes: *chaleur torride*, *chaleur suffocante* et *chaleur accablante.*

Il existe aussi des cas où les mots-bases d'un couple de collocations, faisant partie du même champ sémantique, peuvent être considérés comme des structures synonymiques: par exemple en italien *svolgere un ruolo/una missione* et *s'acquitter d'une charge/remplir une mission*. Un autre cas à signaler est la relation d'hypérononymie/hyponymie entre les mots-bases: par exemple en français *hiver/climat rude* et *hiver/climat rigoureux* et en italien *inverno/clima rigido* et *inverno/clima crudo*. Il y a, en outre, des mots-bases synonymiques ou quasi-synonymiques qui n'intéressent pas les mêmes mots-collocatifs en vertu d'un rapport contraignant et exclusif existant entre les deux éléments. Dans la collocation française *déclarer la guerre*, le collocatif *déclarer* ne se combine pas avec *lutte* (**déclarer la lutte*). Ainsi, en italien *collera* et *furore* se combinent avec différents collocatifs: le premier avec *montare/entrare/andare in*, tandis que le second se combine avec *essere/montare in*. De même en français, *colère* et *fureur* se combinent avec des collocatifs différents: *colère* avec *être/se mettre en* et *fureur* avec *entrer/être/mettre en.*[1]

La question de la synonymie collocationnelle n'est pas purement théorique, mais elle concerne deux domaines d'études: la lexicographie que nous verrons plus tard et l'apprentissage d'une langue étrangère ou seconde (Partington 1998: 33).

Dans la production d'une langue il faut qu'un locuteur connaisse les circonstances précises dans lesquelles utiliser une série d'items et les remplacer par d'autres sans changer le sens de la phrase (Partington 1998: 33): dans le cas de structures collocationnelles la faisabilité sémantique ne suffit pas, il faut absolument assurer une pertinence collocationnelle[2] afin de connaître les «habitudes collocationnelles» des items.

3 Études de cas

3.1 La synonymie collocationnelle à travers les dictionnaires bilingues

Le but fondamental d'un dictionnaire bilingue est celui de faire interagir deux langues en mettant en évidence non seulement leurs différences, mais aussi leurs similitudes (De Giovanni 2011: 19). Son unité fondamentale est constituée par l'équivalence qui rend évident non seulement un recours au principe de contrastivité entre deux systèmes linguistiques, mais surtout qui rend compte de la vision que les deux langues ont du monde (De Giovanni 2011: 24).

Le traitement des structures collocationnelles à l'intérieur du dictionnaire est une problématique qui préoccupe le lexicographe. Un autre aspect important est la notion que le DB donne à la collocation et qui rentre, à plein titre, dans sa fonction pragmatique (De Giovanni 2010: 32–33). Les études et les analyses des lexicologues et des lexicographes se sont toujours concentrées sur le problème du traitement des collocations dans la structure du dictionnaire, oubliant l'importance de la description dans les éléments pré-textuels (tout

[1] Un autre exemple est celui discours et conférence qui se combinent avec différents collocatifs: le premier avec faire et prononcer, tandis que le second se combine avec faire et donner. Comme nous font remarquer Bossé-Andrieu et Maréchal, dans une collocation du type N+Adj, la base ne se combine jamais avec le même collocatif adjectival. C'est le cas de lutte serrée et guerre larvée contre *guerre serrée et *lutte larvée.

[2] Nous avons emprunté à Partintgton (1998: 33) les termes de *semantic feasibility* et de *collocationnel appropriacy.*

ce qui précède la nomenclature) qui constituent un point de départ pour la consultation (De Giovanni 2011). L'attitude du lexicographe bilingue envers les collocations s'entrevoit à travers ces éléments. C'est ici que chaque lexicographe justifie ses choix lexicographiques. C'est à partir de ces éléments qu'un premier contact s'établit entre le dictionnaire (et par conséquent le lexicographe) et l'usager potentiel.[3] À un différent traitement de la notion de collocation s'ajoute l'emploi de différentes terminologies dans le domaine de la phraséologie. Les présentations, par le biais de symboles ou de caractères typographiques, empêchent les usagers d'avoir une idée de la collocation et de son utilisation: la présentation de la forme et du contenu ne coïncident pas. Comment les usagers pourraient-ils utiliser des structures complexes s'ils n'ont pas la possibilité d'en prendre connaissance? La notion de collocation, en lexicographie, nécessite d'être réglementée et uniformisée surtout en fonction du public auquel le DB s'adresse (pertinence lexicographique). Sans une juste réglementation, les dictionnaires bilingues auront du mal à appliquer le critère pragmatique qui veut qu'une relation s'instaure entre les signes et les usagers.

De même, la problématique concerne le traitement d'éventuelles structures synonymiques à l'intérieur des DB. Il reste, néanmoins, difficile d'identifier le véritable critère de traitement de la SC dans les DB et surtout le degré de sa réception de la part de l'usager. Pour cette raison, nous avons choisi d'analyser les cas des structures synonymiques à patrons V+N et N+Adj., dans les deux langues française et italienne en choisissant les collocations les plus significatives pour notre démonstration. Nous analyserons les trois dictionnaires bilingues présents sur le marché: le *Garzanti 2006* (dorénavant *G06*), le *Boch 5ème édition* (dorénavant *B5*) et le *Hachette-Paravia* (dorénavant *HP*).

3.2 Collocation du type V+N

Pour cette typologie de collocation, à patron syntaxique V+N, nous avons choisi le couple *essuyer un échec* et *subir un échec* dont l'un de des deux éléments, dans ce cas le collocatif, est remplacé par un autre verbe assurant le même sens à toute la structure collocationnelle. Le *Petit Robert*, par exemple, enregistre les deux collocations sous la même entrée *échec* en les traitant comme des structures synonymiques.

Le *G06*, à l'entrée *échec*, enregistre les deux collocations, séparées d'une virgule, suivies d'un seul équivalent italien. La seule collocation *essuyer un*

[3] Nous renvoyons à De Giovanni (2010) pour une description détaillée du traitement de la notion de collocation à l'intérieur des éléments para-textuels des DB.

échec est enregistrée sous l'entrée *essuyer*. On a enregistré une absence des deux collocations sous l'entrée *subir*.

Le *B5* n'enregistre la collocation, *essuyer un échec*, qu'à l'entrée *échec*.

Ainsi, le *HP* enregistre les deux collocations sous l'entrée *échec*. En revanche, à l'entrée *essuyer* le mot-base est indiqué entre crochets (indicateur de collocations). La seule présence des mot-bases *sopportare*, *patire* ne donne aucune possibilité à l'usager de savoir quel pourrait être l'éventuel collocatif à utiliser en italien. Il faut aussi remarquer la présence de *subir* en tant qu'indicateur sémantique utile à la désambigüisation.

Du côté italien, nous avons trois collocations qui peuvent être vues comme des synonymes: *subire uno scacco*, *ricevere uno scacco* et *subire uno smacco*. Il s'agit là d'un exemple de collocation restreinte où *smacco* a déjà le sens *insuccès humiliant* et qui ne se combine qu'avec *subire*. Au contraire, *scacco*, en qualité de mot-collocatif, change son sens en fonction de son mot-base.

Dans la section italien-français, la situation se présente différemment par rapport à ce que l'on a vu plus haut. En effet, le *G06* à l'entrée *smacco*, introduit la collocation correspondante, *subire uno smacco*, suivie de l'équivalent *essuyer un affront*. Une autre collocation qu'il enregistre est celle de *subire uno scacco*, à l'entrée *scacco*, suivie de son équivalent *essuyer un échec* précédée de la marque de registre *fig.* Aucune présence de l'autre collocation, mais à l'entrée *subir* le dictionnaire enregistre *subire un oltraggio*, *un affronto* suivies de l'équivalent *essuyer un affront*, *un outrage*.

Le *B5*, à l'entrée *smacco*, introduit *subire uno smacco* suvie de deux équivalents *essuyer un échec* et *ramasser une gamelle*, mais nous nous limiterons à l'analyse du premier. La collocation *ricevere uno scacco* est présente à l'entrée *scacco*. La collocation *subire uno scacco* est enregistrée à l'entrée *subire*. *Sopportare uno scacco* est la quatrième collocation que le dictionnaire enregistre à l'entrée *sopportare*.

Le *HP*, à l'instar des autres dictionnaires, enregistre la collocation *subire uno smacco* sous l'entrée *smacco* et *subire uno scacco* à l'entrée *scacco*. En outre, il est intéressant de remarquer la présence de la collocation *subire uno smacco* sous forme d'indicateur sémantique à l'entrée *scornare* correspondant au français *prendre*, *ramasser une gamelle*.

3.3 Collocation du type N + Adj.

Pour la deuxième typologie de collocation, à patron syntaxique N + Adj., nous avons choisi le couple *colère blanche* et *colère froide* et leurs équivalents italiens.

Le *Petit Robert* traite les deux collocations comme deux structures synonymes à travers l'enregistrement sous la même entrée, *colère*, avec la signification *«qui n'éclate pas».*

Dans la section français-italien du *G06* les deux collocations ne sont pas présentes, alors que dans le *B5 colère blanche* est présente sous *colère* et *colère froide* tant sous *colère* que sous *froid*. Le *HP* enregistre la seule collocation *colère froide* sous *froid*.

Du côté italien, le monolingue qu'on a consulté, le *Devoto-Oli 2009* ne contient aucune des deux collocations. Au contraire, c'est le *HP* qui atteste la présence de *collera fredda* à l'entrée *collera* suivie de l'équivalent *collera repressa*. Le *B5*, par contre, enregistre la collocation *collera repressa* à l'entrée *collera*, suivie des deux collocations *colère froide* et *colère blanche*.

Les différents choix de traitement des trois DB sont bien évidents. De toute façon, on se demande pourquoi un dictionnaire bilingue comme le *G06* n'enregistre aucune de nos collocations.

4 La SC dans le corpus

Une analyse sur corpus permet, en effet, de tenir compte de plusieurs aspects du mot, au sens large du terme, en indiquant comme mot tant les lexies simples que toute lexie plus complexe qui vont du mot composé à la structure figée, et de ses manifestations au sein de l'environnement lexical. Une analyse sur corpus, comme on l'a vu plus haut, sert à déceler non seulement la faisabilité sémantique du mot, mais aussi sa pertinence collocationnelle liée énormément à son sens global. Cela signifie que la manifestation à caractère formel, donc plan de l'organisation syntaxique, syntagmatique et phraséologique, coïncide fortement avec la production de sens dans une situation donnée, d'une langue générale à une langue de spécialité, d'un registre formel à un registre informel etc., où l'usage devient le facteur régulateur.

Or, ce que les dictionnaires disent n'est ni clair ni satisfaisant. Entre l'information contenue à l'intérieur d'un dictionnaire et celle repérée à partir d'un corpus, à savoir l'ensemble de textes représentatif d'une partie d'une langue en usage (le discours), il y a un abîme. C'est pour cette raison que le lexicographe doit avoir confiance en l'utilisation du corpus pour compléter les informations contenues dans les dictionnaires.

Nous avons tenté d'effectuer une analyse sur corpus des couples de collocations, à différents patrons syntaxiques, qui sont considérées comme des structures synonymiques.

4.1 Collocation du type V + N

Dans une analyse sur corpus, il serait important d'opposer non seulement les couples de collocations, mais aussi les différents éléments qui les composent pour attester leur degré de fréquence et pour établir un réseau collocationnel, tenant compte de leur résonance sur la base de leur cooccurrent. Nous nous limitons à opposer les couples dans leur structure complète sur la base de ce que nous avons déclaré ci-dessus.

Dans notre corpus, *essuyer un échec* est présent 1 154 fois où son domaine fort concerne la prise de conscience d'une faillite après une tentative ou en supposant une tentative et où, dans la majeure partie des cas, le verbe est conjugué au présent ou au passé. Sa prosodie sémantique, due à la présence d'autres éléments à l'intérieur d'une fenêtre restreinte, est négative. Du point de vue de la structure, la collocation n'est pas complètement cristallisée. Au contraire, elle admet la présence des autres éléments entre un élément et un autre et elle est assujettie au phénomène de la *passivisation*.

Dans le cas de *subir un échec*, présent 14 881 fois, sa structure, à l'instar d'*essuyer un échec*, n'est pas cristallisée étant donné qu'elle admet l'interposition d'adjectifs entre un élément et un autre. Mais, en même temps, toute la structure se caractérise par des éléments placés à droite comme *assez humiliant*, *complet*, *total*, *cuisant*, *définitif*, *lamentable*, *mémorable* qui contribuent, sur le plan sémantique, avec d'autres éléments de l'environnement lexical, à la manifestation d'une prosodie sémantique négative.

Du côté italien, la situation dans notre corpus est bien différente. En effet, si la structure *ricevere uno scacco* est absente de notre corpus, la collocation *subire uno scacco*, au contraire, apparaît une seule fois accompagnée à gauche de la fenêtre des items *senza mai* à prosodie sémantique neutre. La deuxième collocation *subire uno smacco* n'apparaît que trois fois en présence d'autres items, tant à gauche qu'à droite, en donnant lieu à une structure plus complexe, *poter subire uno smacco del genere*, *dover subire uno smacco in…*, à prosodie sémantique neutre.

4.2 Collocation du type N + Adj

Le thème dominant des collocations *colère froide/colère blanche* est bien sûr celui d'un sentiment réprimé qui tarde à se manifester à cause de nombreuses raisons liées au caractère de la personne, aux influences culturelles, historiques et sociales. Mais si les deux collocations restent, au niveau de l'expression, cristallisées n'admettant que très peu de changements internes à la structure, ce qui les différencie c'est le contexte dans lequel elles sont utilisées. En effet, la *colère froide* (945 occurrences) est un sentiment qui implique plutôt la partie psychologique de l'être humain tandis que la *colère blanche* (64 occurrences) intéresse en premier lieu la partie physique en générant de forts malaises. Des combinaisons telles que *brûler de colère froide*, *accès de colère froide*, *se mettre dans une colère froide* confirment le caractère abstrait de la *colère froide* qui la rend similaire à la notion de *colère réprimée* (53 occurrences). Par contre, la *colère blanche* non seulement est fortement liée à des changements physiques de l'individu, palpitations du cœur, agressivité, mais à la différence de l'autre collocation qui a un caractère individuel, elle est souvent sociale et partagée par tous ceux qui éprouvent un sentiment de contrariété envers certains comportements.

La situation en italien est différente surtout pour ce qui concerne la fréquence: 2 occurrences pour *collera bianca* et 3 occurrences pour *collera fredda* auxquelles il faut ajouter 14 occurrences pour *collera repressa*. Pour cette dernière le thème dominant reste toujours celui d'un sentiment réprimé en impliquant surtout la partie physique de l'être humain.

La toile, au contraire, contient 793 occurrences de *collera fredda*. Dans la majeure partie des cas, il s'agit de sites, de forums ou de blogs de littérature où la collocation en question a une utilisation plutôt littéraire. Parmi les informations repérées, il est curieux de signaler la présence d'un *Traité d'agriculture* publié en 1805, où l'auteur, Piero de Crescenzi, conseille l'utilisation de la hièble contre la *colère blanche*. La collocation *collera fredda*, par contre, a une fréquence majeure, 1 370 occurrences, étant fortement liée au sens de grande irritation qui n'implique que très rarement la partie physique de l'être humain. Mais la collocation la plus fréquente sur la toile est bien sûr celle de *collera repressa* avec 3 060 occurrences. La *collera repressa* est l'expression d'un sentiment caché qui peut se manifester à la fois psychologiquement et physiquement. En effet, très souvent la *collera repressa* peut se transformer en un accès de folie, de vengeance, d'intolérance jusqu'à donner la mort à ses proches.

5 Conclusions

Définir la collocation et ses relations synonymiques avec d'autres structures similaires est encore une entreprise difficile à réaliser. La collocation rentre, avec les autres structures idiomatiques, dans une organisation de la langue de type phraséologique, comprise entre l'organisation syntagmatique et celle syntaxique, régie par un principe d'agencement. L'agencement n'est qu'un mécanisme d'arrangement résultant d'une combinaison. Dans le cas d'une collocation, l'agencement concerne plutôt le mécanisme qui préside à sa formation en tant que simple structure semi-figée, binaire, à caractère arbitraire où les composants sont dissymétriques. Ce seul mécanisme ne suffit pas à décrire la collocation. C'est pour cette raison, que nous faisons appel au terme de contexture, à savoir l'ensemble de relations organisées entre des éléments significatifs formant un tout complexe et organique. Dans ce cas, ce qu'il faut prendre en compte c'est l'ensemble des relations que tous les éléments concourent à établir (réseau collocationnel) en identifiant leurs fonctions à l'intérieur du texte (résonance collocationnelle, préférence sémantique, prosodie sémantique) pour identifier le phénomène collocationnel. Partir de ces postulats nous permet de déceler les différences, et les ressemblances, entre collocations synonymiques en faisant appel aux méthodologies d'analyse de la linguistique de corpus pour une description pertinente au sein des DB.

Ce qui est nécessaire c'est une fusion entre l'approche lexicographique et celle linguistique pour en établir une troisième qui soit en mesure de donner une vision globale du phénomène collocationnel dans le discours.

Références bibliographiques

DE GIOVANNI, Cosimo, 2010: Pragmatique et didactique du dictionnaire. Quelques réflexions terminologiques. *Verbum* 1, 27–36.

DE GIOVANNI, Cosimo, 2011: L'équivalence lexicographique dans la différence. Des réflexions pour l'avenir. *Verbum* 2, 18–26.

FIRTH, John Rupert, 1957: *Papers in linguistics*. London: Oxford University Press.

FIRTH, John Rupert, 1968: A Synopsis of Linguistic Theory, 1930–1955. Palmer, Franck Robert (ed.): *Selected Papers of J. R. Firth 1952–1959*. Londres: Longman. 168–205.

LOUW, Bill, 1993: Irony in the Text or Insincerity in the Writer ? The Diagnostic Potential of Semantic Prosody. Baker, Mona et al. (eds.): *Text and Technology: In Honour of John Sinclair*, Amsterdam: John Benjamins. 157–192.

MULLER, Claude, 2008: *Les bases de la syntaxe. Syntaxe contrastive français-langues voisines*. Pessac: Presses Universitaires de Bordeaux.

PARTINGTON, Alain, 1998: *Patterns and Meanings: Using Corpora for English Language Research and Teaching.* Amsterdam: John Benjamins.

SINCLAIR, John, 1996: The Search for Units of Meaning. *Textus. English Studies in Italy* 9, 75–102.

WILLIAMS, Geoffrey, 1998: Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics* 3, 151–171.

WILLIAMS, Geoffrey, 2003: Les collocations et l'école contextualiste britannique. Tutin, Agnes, Francis Grossmann (ed.): *Les collocations. Analyses et traitement*. Amsterdam: De Werelt. 33–44.

WILLIAMS, Geoffrey, 2006: Biblical resonance: A corpus-driven analysis of collocational resonance in French and English Texts. Hédiard, Marie (ed.): *Lezioni di Dottorato 2005*. Cassino: Edizioni Spartaco. 1–26.

WILLIAMS, Geoffrey, 2010: Many rooms with corpora. *International Journal of Corpus Linguistics* 15, 400–408.

Dictionnaires

Le Nouveau Petit Robert de la langue française, 2012. Paris: Le Robert.

Il Nuovo Hachette-Paravia dizionario francese italiano italiano francese, 2007. Torino: Paravia.

Il Boch. Dizionario francese italiano italiano francese, 2007. Bologna: Zanichelli.

Grande Dizionario Francese Garzanti, 2006. Milano: Garzanti.

Dizionario della lingua italiana Devoto-Oli, 2009. Firenze: Le Monnier.

Kollokationenlexikographie. Ein Bericht aus der Praxis

Marcel Dräger, René Frauchiger, Marlène Linsmayer, Alessandra Widmer (Basel)

Abstract

Kookkurrenzanalysen liefern das sprachliche Rohmaterial, aus welchem Kollokationenwörterbücher entstehen. Die Qualität der Kookkurrenzenlisten ist direkt vom Korpus und der verwendeten Analysemethode. Unstrittig ist auch, dass von der Kookkurrenzenliste bis zum fertigen Wörterbuch abhängig noch einiges an intellektuellem und manuellem Aufwand nötig ist. Unseres Wissens gibt es aber bislang keine Berichte darüber, welche Entscheidungen in den Wörterbuchredaktionen anstehen und auf welcher theoretischen Basis sie getroffen werden. Mit diesem Aufsatz möchten wir einen ersten Schritt in diese Richtung gehen und mit drei Fallbeispielen aus der Praxis des *Wörterbuchs der typischen und gebräuchlichen Wortverbindungen im Deutschen* die Notwendigkeit und die Herausforderungen einer manuellen Nachbearbeitung von Kookkurrenzdaten unterstreichen. Die Wortverbindungen *jemanden zum Mann nehmen*, *korrupte Politikerin* und *Würfelzucker* sind syntaktisch wie semantisch unterschiedlicher Art, doch es gibt etwas, das sie verbindet. Um sie zu lexikalisieren war – trotz einer schlüssigen theoretischen Ausgangslage – “Handarbeit” nötig. Anhand von drei Beispielen wird der Stellenwert und die Herausforderung von manuell-qualitativer Nachbearbeitung bei der Produktion eines Kollokationenwörterbuchs aufgezeigt: Der Aspekt der Relevanz von Kollokationen zeigt auf, wie im Fall der Basislemmata *Mann* und *Meinung* unterschiedliche indikatorengestützte Auswahlverfahren zum Tragen kommen (müssen). Ebenso erweist sich die quantitative und qualitative Genderasymmetrie in Korpora und Kookkurrenzenlisten als Herausforderung bei der adäquaten Versprachlichung der Geschlechter. Zuletzt werden redaktionelle Herangehensweisen und Stolpersteine bei der Aufnahme von Komposita in Kollokationenwörterbüchern besprochen.

1 Für die lexikographische Praxis sind Kollokationen nicht ausreichend definiert

Ein Wörterbuch zu Kollokationen lässt sich nicht per Knopfdruck aus korpuslinguistisch ermittelten Kookkurrenzlisten zusammenstellen. Dass – je nach anvisiertem Endprodukt – manuell und intellektuell nachbearbeitet werden muss, ist hinlänglich bekannt (Schütze/Manning 1999: 151f.). Selbst tendenziell computerlinguistisch geprägte Projekte wie das *Wörterbuch der Kollokationen* *im Deutschen* kommen/kamen nicht ohne redaktionelles Zutun aus (Quasthoff 2011).[1] Es stellt sich also nicht die Frage, ob in der Kollokationslexikographie eine qualitative Nachbearbeitung vonnöten ist, sondern in welchem Ausmaß, mit welcher Tiefe und auf welchem theoretischen Fundament sie vollzogen wird. Unseres Wissens fehlen bislang grundlegende Darstellungen und Überlegungen, die sich qualitativ mit der Nachbearbeitung von Kookkurrenzen befassen. Anhand von Beispielen aus der lexikographischen Praxis des *Wörterbuchs der typischen und gebräuchlichen Wortverbindungen im Deutschen*[2] zeigen wir im Folgenden den Stellenwert und einige Herausforderungen einer qualitativen Überarbeitung quantitativ-statistisch ermittelter Kookkurrenzvorschläge. Dies geschieht stets aus einer auf das Projekt zentrierten Perspektive, die speziell Sprachlernende als Wörterbuchbenutzende im Blick hat.

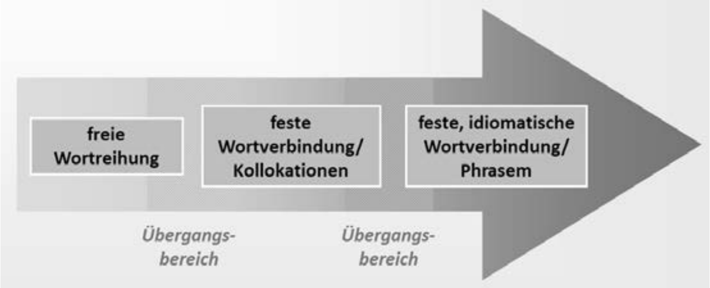

freie Wortreihung
feste Wortverbindung/ Kollokationen
feste, idiomatische Wortverbindung/ Phrasem
Übergangsbereich
Übergangsbereich

Abbildung 1: Kollokationen im Kontinuum der Festigkeit von Wortverbindungen.

Sehr schnell sollte sich im Verlauf des Projektes herausstellen, dass schon eine einigermaßen homogene Identifikation von Kollokationen innerhalb eines Redaktionsteams problematisch ist. Der Grund hierfür ist in der Definition von Kollokationen zu suchen: Kollokationen definieren sich als Abschnitt innerhalb eines Kontinuums. Dieses reicht von freien, ad hoc gebildeten Wortreihungen (*einen Satz wegwischen*) über etwas gefestigtere Wortverbindungen (*die Tafel wischen*) bis zu festen, hoch idiomatischen Phrasemen (*Tabula rasa*

[1] Quasthoff (2011: XII) schildert in der Einleitung des *Wörterbuchs der Kollokationen im Deutschen* die anfängliche Hoffnung, dass statistisch errechnete Nachbarschaftskookkurrenzen auch Kollokationen seien. Letztlich konnte die Entscheidung ob wörterbuchrelevant oder nicht aber nur von einer Redaktion getroffen werden.

[2] Das *Wörterbuch der typischen und gebräuchlichen Wortverbindungen des Deutschen* wird vom Schweizerischen Nationalfonds unterstützt und an der Universität Basel unter der Leitung von Prof. Dr. Annelies Häcki Buhofer erarbeitet. Mitarbeitende sind außer der Autorinnen und Autoren Markus Gasser, Jana Göke, Lorenz Hofer, Stefanie Meier, Eva Rösch, Tobias Roth und Caroline Runte.

machen) (vgl. Abbildung 1). Das definitorische Spektrum der Kollokationen hat allerdings keine scharfen, eindeutig zieh- und nachvollziehbaren Grenzen, sondern einen variablen Übergangsbereich in beide Richtungen.

Als wichtiges Kriterium für die Abgrenzung von Kollokationen gegenüber den freien Wortreihungen gilt in den meisten Definitionen die Festigkeit (Burger 2010: 15f.) oder Usualität (Steyer 2001). In der empirischen Praxis wird die Festigkeit/Usualität über Korpusanalysen bestimmt, weshalb dieses Definitionsmerkmal von Kollokationen zu einer direkten Abhängigkeit von den verwendeten Korpora und Algorithmen führt.[3] Die statistisch-quantitative Festigkeit stimmt allerdings nicht immer mit der psycholinguistischen Festigkeit – also dem, was man intuitiv als feste Wortverbindung erachtet – überein (Burger 2010: 15f.), wodurch sich im Wörterbuch eine Diskrepanz zwischen den empirisch ermittelten Einträgen und dem Gefühl der Nutzenden ergeben kann. Es ist eine zentrale Aufgabe einer Wörterbuchredaktion, diese Diskrepanz möglichst klein zu halten.

Die Abgrenzung der Kollokationen in die andere Richtung, also gegenüber den Idiomen, fällt noch schwerer, da sich Idiomatizität als semantisches Kriterium einem korpusanalytischen Zugang weitestgehend verschließt. Es zeigt sich, dass Kategorien wie „nicht oder nur schwach idiomatisch“, mit welchen beispielsweise Burger (2010: 52) Kollokationen definiert, griffig und handlich in der Theorie aber schwammig und biegsam in der lexikographischen Praxis sind.

Für Lexikographinnen und Lexikographen eines Kollokationenwörterbuchs entsteht damit ein großer Handlungsspielraum bei der Auswahl der für ein Wörterbuch relevanten Kollokationen. In diesem Aufsatz möchten wir aufzeigen, wie viel Handlungsspielraum und Entscheidungsnotwendigkeit in der lexikographischen Praxis bestehen, obwohl wir uns an der theoretisch absolut schlüssigen Definition Burgers (2010: 52) orientieren. Unsere Erfahrung basiert auf dem redaktionellen Prozess des an der Universität Basel unter der Leitung von Prof. Dr. Annelies Häcki-Buhofer erarbeiteten Kollokationenwörterbuchs.

In der Theorie lässt sich der redaktionelle Prozess eines Kollokationenwörterbuchs folgendermaßen darstellen: Eine Definition wird quasi als Filter auf die *langue* angewendet, wodurch eine hypothetische Menge *X* an Kollokationen für ein Basiswort ermittelt werden kann. Im praktischen redaktionellen Prozess hingegen ist die tatsächliche Menge *Y* an gefundenen Kollokationen kleiner als die Menge *X* (vgl. Abbildung 2). Das Ziel bei der Erarbeitung eines

[3] Dementsprechend sind die in einem korpusbasierten Kollokationenwörterbuch dargestellten Wortverbindungen primär vom verwendeten algorithmischen Verfahren, den angewandten Schwellenwerte und dem zugrundeliegenden Korpus abhängig.

Kollokationenwörterbuchs liegt unter anderem darin, die Differenz zwischen theoretischem und praktischem Ergebnis, also zwischen *X* und *Y*, möglichst gering zu halten. Abbildung 2 zeigt eine Gegenüberstellung des lexikographischen Verfahrens aus theoretischer und praktischer Perspektive:

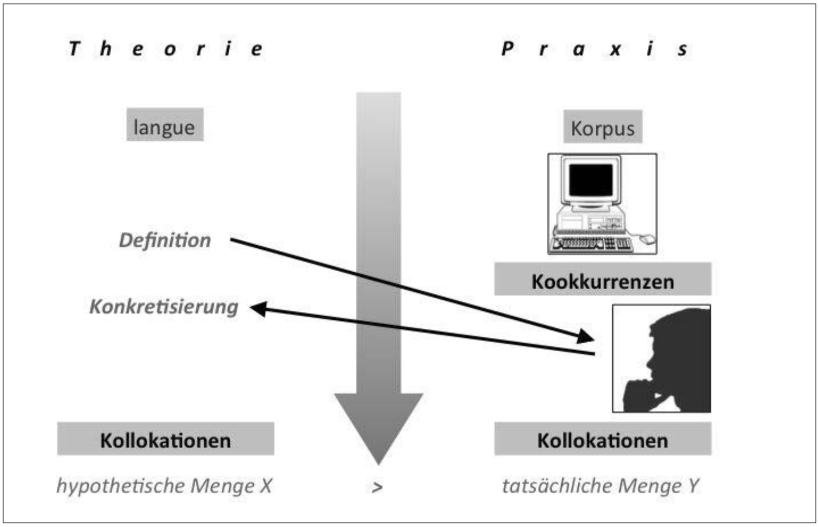


Abbildung 2: Theorie und Praxis in der lexikographischen Kollokationserhebung.

Ein großer Teil der Differenz zwischen *X* und *Y* ist der Tatsache geschuldet, dass Korpora nur einen (meist schriftsprachlichen) Teilausschnitt der *langue* abbilden. Dementsprechend tritt die Diskrepanz zwischen statistischer Festigkeit und psycholinguistischer Festigkeit beispielsweise häufig bei solchen Wortverbindungen auf, die in schriftlichen Korpora unterrepräsentiert sind. Weitere Verluste bei den tatsächlich gefundenen Kollokationen entstehen im ersten Arbeitsschritt, in welchem mit korpusanalytischen Verfahren maschinell die Kookkurrenzpartner zu einem Basiswort ermittelt werden.[4] Diese Kookkurrenzdaten enthalten im Idealfall alle relevanten Kollokationen. Dem Idealfall stehen allerdings einige korpusanalytisch bedingte Störfaktoren

[4] Wir möchten an dieser Stelle nicht auf die teilweise erheblichen Unterschiede der existierenden Methoden zur Kookkurrenzberechnung eingehen. Logischerweise fallen bei unterschiedlichen Methoden auch die Verluste unterschiedlich aus. Die Tatsache aber, dass durch die Anwendung korpusanalytisch-statistischer Verfahren die Differenz zwischen theoretischer Menge *X* und tatsächlich gefundener Menge *Y* größer wird, dürfte niemand anzweifeln.

entgegen: bspw. Lemmatisierungsfehler, Fehler im Part-of-speech-tagging oder im Parsing, Homographie. Somit kommt dem nächsten, manuellen Auswahlschritt dahingehend eine besondere Bedeutung zu, dass nicht nur die vorhandenen Kookkurrenzen reduziert, sondern idealerweise auch fehlende Kollokationen ergänzt werden sollen. Ein solcher manueller Auswahlprozess folgt in der Praxis Handlungsanweisungen, die aus der zugrundeliegenden Definition von Kollokationen abgeleitet sind. Diese Handlungsanweisungen bedürfen allerdings einer beständigen Konkretisierung während des Redaktionsprozesses in dem Ausmaß, in welchem sich bei der Kollokationsauswahl neue Entscheidungsspielräume eröffnen.

Im Folgenden haben wir drei Fälle herausgegriffen, in welchen bestehende Kollokationsdefinitionen an ihre Grenzen stoßen und lexikographischer Entscheidungsbedarf entsteht.

Im ersten Beispiel fragen wir, welche Rolle die Relevanz einer Wortverbindung bei der Kollokationsauswahl spielen soll. Am beispielhaften Vergleich der Kollokationen zu den Basislemmata *Mann* und *Meinung* wird sehr schnell deutlich werden, dass nicht alle ermittelten Kookkurrenzen auch Kandidaten für ein Kollokationenwörterbuch sein können.

Im zweiten Beispiel besprechen wir das generische Maskulinum als Herausforderung der gendersymmetrischen Erfassung von Kollokationen. Dabei wird sich zeigen, dass die korpusbedingte genderasymmetrische Ausgangslage nur durch qualitative (Nach-) Bearbeitung zu brauchbaren Nennformen und Einträgen im Wörterbuch führt.

Mit dem dritten Beispiel, in welchem wir das Augenmerk auf Komposita richten, stellen wir ein Definitionskriterium von Kollokationen, nämlich die Polylexikalität, auf den Prüfstand. Es wird sich zeigen, dass der Unterschied zwischen Kollokation und Kompositum oft verschwindend gering ist. Häufig stehen Komposita und Kollokationen nämlich aus semantisch-pragmatischer Perspektive in direkter Konkurrenz zueinander und setzen sich aus dem gleichen Wortmaterial zusammen.

2 Zur Relevanz von Kollokationen

Vergleicht man die binomisch ermittelten Kookkurrenzen zu den beiden Lexemen *Meinung* und *Mann*, wie sie die beispielsweise das DWDS-Korpus liefert, dann fällt folgendes auf: *Mann* weist deutlich mehr Einträge auf und *Meinung* hat die statistisch signifikanteren Kollokationen. Dieses Bild spiegelt sich auch in Wörterbüchern wider. Beispielsweise führt das *Oxford Collocations*

Dictionary für das Englische unter *opinion* 15 Nomen-Verb-Kollokationen auf (McIntosh 2009: 566):

> **verb + opinion**: have, hold, air, express, give (sb), offer (sb), state, venture, voice, share, ask (sb), seek, want, get, form, change
> **opinion + verb**: change, differ, vary

Unter dem Lemma *man* (male person) sind keine Nomen-Verb-Kollokationen im *Oxford Collocations Dictionary* eingetragen (McIntosh 2009: 500). Dahinter ist nicht nur das Ergebnis einer Korpusanalyse zu vermuten, sondern auch die bewusste Entscheidung einer Lexikographin oder eines Lexikographen, für *man* keine Kollokationen aufzunehmen.

2.1 Viele aber wenig signifikante Verb-Verbindungen mit *Mann*

Basiswort *Meinung*		Basiswort *Mann*	
Kookkurrenz	**häufigste Verbindung**	**Kookkurrenz**	**häufigste Verbindung**
vertreten	eine Meinung vertreten	*sagen*	ein Mann sagt etw.
sollen	nach meiner/seiner/… Meinung sollte …	*sehen*	ein Mann sieht etw.
teilen	eine Meinung teilen	*kommen*	ein Mann kommt (herein)
bilden	sich eine (eigene) Meinung bilden	*stehen*	ein Mann steht … **aber auch:** seinen Mann stehen
sagen	jemandem die Meinung sagen	*gehen*	ein Mann geht
müssen	nach meiner/seiner Meinung muss …	*sitzen*	ein Mann sitzt
(wieder-) geben	eine Meinung wiedergeben	*sprechen*	ein Mann spricht
ändern	ihre/seine Meinung ändern	*kennen*	einen Mann kennen
dürfen	nach meiner/seiner Meinung darf …	*wissen*	ein Mann weiss etwas
hören	jemandes Meinung hören	*nehmen*	ein Mann nimmt … **aber auch:** sich einen Mann nehmen

Table 1: Die häufigsten Kookkurrenzen zu *Meinung* und *Mann*

Das *Wörterbuch der typischen und gebräuchlichen Wortverbindungen* wird auf der Basis von binomischen Kookkurrenzenlisten erarbeitet, die nach die nach dem Log-Likelihood-Maß sortiert sind.[5] Aus diesen Listen werden anhand von statistischen und manuellen Kriterien jene Kollokationen ausgewählt, die ins Wörterbuch kommen sollen. Dabei ist die Kookkurrenzenliste für *Meinung* mit rund 600 Einträgen deutlich kürzer als jene für *Mann* mit rund 1600 Einträgen. Ungeachtet der Anzahl der Kookkurrenzen sind die Wortverbindungen der beiden Lexeme von sehr unterschiedlicher Qualität. Das wird stellvertretend an den ersten zehn Einträgen deutlich, die in Tabelle 1 abgebildet sind:

Die ermittelten Kookkurrenzen zu *Meinung* verweisen fast alle auf valide Kollokationen. Außer *sollen*, *müssen* und *dürfen* handelt es sich hier um Verbindungen, welche in der Redaktion des *Kollokationenwörterbuchs* als relevant für das Wörterbuch eingestuft wurden.

Die *Mann*-Liste ist jedoch problematischer. *Ein Mann sagt etwas* ist die häufigste Wortverbindung in den Korpusbelegen, die sich hinter dem Binom *Mann + sagen* verbirgt. Doch scheint sie uns aus dem Sprachgefühl heraus uninteressant, ja geradezu banal zu sein. Ein Großteil der Kookkurrenzen zu *Mann* besteht aus solchen Verbindungen. Auf der Basis des Log-Likelihood-Maßes und einem Suchabstand von 5 Wörtern würde sich ein Wörterbuchartikel zum Wort *Mann* ergeben, der sich über mehrere Seiten erstreckt. Er bestünde aus eher losen Verbindungen wie *ein Mann geht*, *ein Mann sitzt*, *ein Mann spricht*. Für zukünftige Benutzerinnen und Benutzer erscheint eine solche Liste durchaus unbefriedigend. Sie sagt ihnen lediglich, was ein Mann als Person alles tun kann und in deutschen Texten offensichtlich häufig tut. Sie berührt jedoch kaum Probleme, die beim Lernen der deutschen Sprache entstehen können oder beim Formulieren von Texten auftreten. Die Anzahl der Kollokationen lässt sich nun möglicherweise durch weitere statistische Berechnung, durch den Vergleich der Signifikanzwerte oder andere Verfahren der Kookkurrenzermittlung verringern.

Dass dies wünschenswert ist, zeigt ein Blick auf weitere Wortverbindungen aus der Kookkurrenzenliste zu *Mann*: *einen Mann finden*, *jemanden zum Mann nehmen*. Hierbei handelt es sich aus unserer Sicht um Kandidaten für ein Kollokationenwörterbuch, da sie beispielsweise für Sprachlernende Schwierigkeiten bereiten können. Dementsprechend müssen solche Verbindungen in einem Wörterbuch aus der Masse der eher losen Wortverbindungen herausragen, wenn sie nicht gar die einzigen aufnahmewürdigen sind. Das Verhältnis der wörterbuchrelevanten Kollokationen zu den weniger relevanten ist allerdings

[5] Zur Methode vgl. Roth 2012b: Kap. 3.2: Kollokationsextraktion.

gerade beim Lexem *Mann* sehr ungünstig, und es ist – zumindest bei Log-Likelihood – keineswegs so, dass die relevanten Treffer vorrangig gelistet sind.

2.2 Konzentration auf relevante Kollokationen

In der Wörterbuchredaktion muss nun auf Basis der korpusanalytisch ermittelten Wortverbindungen eine Auswahl der Kollokationen getroffen werden, die im Wörterbuch aufgeführt werden sollen. Keinesfalls ist es möglich und sinnvoll die ganze Anzahl an statistisch signifikanten Wortverbindungen für das Wort *Mann* aufzunehmen. Es kommt damit die Relevanz ins Spiel. Das heißt, die Auswahl der Kollokationen basiert nicht mehr nur auf statistischen Werten und Frequenzdaten, sondern sie wird auch im Hinblick auf die Wörterbuchbenutzenden getroffen. Indikatoren für die Relevanz einer Kollokation liefern uns die Korpusbelege. Mögliche Entscheidungshilfen sind beispielsweise:

- die Festigkeit
  Ein Mann sagt … weist eine geringere Festigkeit auf als *jemanden zum Mann nehmen*. Das spiegelt sich vor allem im Kontext wider. So sind Variationen möglich wie *ein kluger Mann sagte einmal …*, *ein richtiger Mann sagt so etwas nicht* usw. *Jemanden zum Mann nehmen* lässt sich hingegen kaum abwandeln.

- Übersetzungsschwierigkeiten
  Ein Mann sagt … verursacht kaum Übersetzungsprobleme, während man *jemanden zum Mann nehmen* wortwörtlich weder ins Französische noch ins Englische übertragen kann.

- die Generalisierbarkeit der Kollokation
  Viele Verbindungen mit *Mann* lassen sich auf alle beliebigen Personenbezeichnungen übertragen: *ein Schreiner sagt/sieht/kommt/steht…*, *meine Chefin sagt/sieht kommt/steht …* Bei *jemanden zum Mann nehmen* hingegen ist außer *jemanden zur Frau nehmen* keine Generalisierung möglich.

- äquivalente Ausdrücke
  Zu ein Mann sagt … sind äquivalente Ausdrücke gleicher Konstruktionsweise möglich: *ein Mann spricht/erzählt/meint/findet …* Bei *jemanden zum Mann nehmen* gibt es keine konstruktiv gleichen Ausdrücke. Lediglich *jemanden heiraten* ist möglich, oder die Bedeutung der Kollokation muss umständlich paraphrasiert werden.

Kriterien wie Festigkeit, Äquivalenz oder Generalisierbarkeit helfen der Redaktion über die Relevanz einer Wortverbindung für das Wörterbuch zu ur-teilen. Sie sind Konkretisierungen der Definition von Kollokationen, die im Sinne der Abbildung 1 während des redaktionellen Prozesses entstanden sind. Diese Indikatoren haben aber keinen definitorischen und schon gar keinen ausschließenden Charakter, sondern dienen dazu, lexikographische Entscheidungen ausgewogener zu gestalten. Schlussendlich wird somit der Artikel zu *Mann* erheblich kürzer werden, als es die lange binomische Kookkurrenzenliste zuerst vermuten ließ: Den 1600 korpusanalytisch vorgeschlagenen Nomen-Verb-Verbindungen stehen zehn gegenüber, die Einzug in das Wörterbuch finden. Beim Lexem *Meinung* hingegen ist die Diskrepanz zwischen rund 600 ermittelten Kookkurrenzen und letztendlich 60 aufgenommenen Kollokationen deutlich geringer. Daran zeigt sich die unterschiedliche Bindungsqualität der einzelnen Basiswörter und ihrer Kollokatoren.

3 Gender(a)symmetrie als Herausforderung in der Kollokationslexikographie

Zeit online berichtete am 28. 10. 12 zur Publikation der sogenannten Lagarde-Liste durch die Zeitschrift *Hot Spot*:

> Am Vortag hatte die Zeitschrift eine Liste mit insgesamt 2.059 angeblichen Steuersündern veröffentlicht, die unversteuerte Gelder aus Griechenland in die Schweiz überwiesen hätten. Darunter waren auch einige Politiker, Journalisten, aber auch Hausfrauen und Studenten.[6]

Die Zeit-Redaktion spricht hier die gesellschaftliche Durchmischung der 2.059 „Steuersünder“ an: nicht nur Politiker und Journalisten, sondern auch Hausfrauen und Studenten werden genannt. Man könnte also meinen, es sei in dieser Auflistung das volle Personenensemble der Lagarde-Liste präzise benannt – dies ist jedoch nicht der Fall. Es handelt sich hier nämlich um die klassische ambige Bezeichnungssituation, die durch die Verwendung von generischen Maskulina hervorgerufen wird. Im Gegensatz zur femininen Form, die im Deutschen nur auf eine Frau resp. eine Gruppe von Frauen verweisen kann, kann die maskuline Form alle möglichen Personen(-gruppen) bezeichnen. So können die Pluralformen *Politiker*, *Journalisten* und *Studenten* im obigen Beispiel sowohl ausschließlich auf Männer als auch auf eine gemischtgeschlechtliche Gruppe referieren. Das Maskulinum in seiner geschlechtsunspezifischen Zusatzfunktion vermag Personenbezeichnungen nur bedingt zu

[6] <www.zeit.de/politik/ausland/2012-10/griechenland-journalist-steuersuender>. Zugriff: 9. 12. 2012.

neutralisieren. Nach Pusch (1983: 53f.) handelt es sich dabei folglich nur um eine Pseudoneutralisierung. In der obigen Auflistung führt zudem die Nennung der Hausfrauen zu weiteren Unsicherheiten. Als einzige geschlechtsspezifische Personenbezeichnung stärkt *Hausfrauen* die Annahme, dass es sich auch bei den anderen Personenbezeichnungen um geschlechtsspezifische, also männliche, Personengruppen handelt (Bussmann u.a. 2003: 159).

Solch ambige Bezeichnungssituationen schlagen sich auch als weitaus häufigster Fall von Personenreferenz in den Korpora nieder. Der Umgang mit generischen Maskulina stellt deshalb eine Herausforderung für den lexikografischen Arbeitsprozess dar. Dies soll im Folgenden an den drei Hauptarbeitsschritten der lexikographischen Kollokationserhebung – der Korpusarbeit, der Kookkurrenzenanalyse und der Kollokationsauswahl – dargestellt werden.

3.1 Korpora belegen vor allem das (generische) Maskulinum

Das generische Maskulinum als die vorherrschende Form der pseudoneutralen Personenreferenz bildet sich in den Korpora signifikant ab. Das DWDS-Kernkorpus belegt bspw. das Lemma *Arzt* und alle seine gebeugten Formen 14.093-mal, die jeweiligen Wortformen zu *Ärztin* sind lediglich 462-mal vorhanden. Eine ähnlich große Diskrepanz zeigt sich auch in weniger historischen Korpora.[7] Die extreme Überzahl an maskulinen Formen erklärt sich nicht nur durch die historisch und kulturell bedingte häufigere Versprachlichung von männlichen Akteuren sondern auch durch den erwähnten Zusammenfall der generischen und der maskulinen Form in den Korpora. Ob ein Beleg generischer oder geschlechtsspezifischer Art ist, kann meist weder computerlinguistisch noch intellektuell erschlossen werden. Die Bedingungen für eine kollokationslexikographische Unterscheidung männlicher und weiblicher Personenbezeichnungen sind daher denkbar schlecht. Ein ausgewogeneres Korpus wäre hier wünschenswert, wenngleich seine Erstellung ebenfalls recht problembehaftet sein dürfte. In unseren Korpora stellen wir eine Unterrepräsentation weiblicher Personenbezeichnungen und gleichzeitig eine unauflösbare Polysemie männlicher Personenbezeichnungen fest. Diese ambige und nicht auflösbare Verzerrung wirkt sich entsprechend auf eine häufigkeitsbasierte Kookkurrenzanalyse aus.

[7] Das von Roth (2012a) 2011 erstelltes Web-Korpus mit 775 Millionen Tokens belegt zu *Arzt/Ärztin* 70.180 maskuline und 6.226 feminine Formen. Zugriff 30. 11. 12.

3.2 Genderasymmetrische Beleglage in den Kookkurrenzenlisten

Die Substantiv-Verb-Kookkurrenzenlisten[8] zu den Wortformen *Politikerin* und *Politiker* gestalten sich in der Bearbeitungsmaske unseres Projektes wie folgt (vgl. Tabelle 2):

Politikerin : Politiker	
Verben ♀ (3 von 3)	Verben ♂ (12 von 430)
sein	müssen
erfahren	sein
verschwinden	sollen
	wählen
	sagen
	fordern
	haben
	wissen
	reden
	glauben
	vertreten
	versuchen
	/…/

Tabelle 2: Substantiv-Verb-Kookkurrenzen zu *Politikerin* und *Politiker.*

Für *Politikerin* wurden lediglich drei – wenig relevante – Verbkookkurrenzen ermittelt, die in enormem Gegensatz zu den rund 430 Kookkurrenzen mit maskulinen Formen stehen. Weiter sind die Belegzahlen für *Politikerin* generell deutlich niedriger. Die quantitative Asymmetrie und das Zuordnungsproblem von generischen respektive maskulin-geschlechtsspezifischen Formen die für das Korpus gilt, überträgt sich in dem Maße auf die Auswertung, dass keinerlei Aussagen über die Geschlechtsspezifik von Kollokationen getroffen werden können.

Aus einer rein quantitativ-deskriptiven Sicht müsste aufgrund der Korpusbeleglage davon abgesehen werden, Substantiv-Verb-Kollokationen zum Basissubstantiv *Politikerin* im Wörterbuch aufzuführen. Erst durch eine qualitative Analyse der Belege wird dieses Ungleichgewicht aufgefangen.

[8] Zur Erstellung der Substantiv-Verb-Kookkurrenzen vgl. Fußnote 6.

3.3 Korrupte Politiker und verantwortliche Politikerinnen: Herausforderungen in der statistikbasierten Kollokationsanalyse

Die Tatsache, dass heute sowohl Frauen als auch Männer politisch aktiv sind, spricht intuitiv für einen gleichwertigen Artikel *Politikerin* neben *Politiker* oder einen gemeinsamen Artikel für beide Formen. Die aus der Realität abgeleitete Einschätzung deckt sich jedoch weder quantitativ noch qualitativ mit den statistikbasierten Ergebnissen der Kookkurrenzanalyse.

Politikerin : Politiker	
Adjektive ♀ (23 von 23)	Adjektive ♂ (23 von 320)
konservativ	korrupt
prominent	verantwortlich
engagiert	unfähig
grün	deutsch
freiheitlich	verantwortungslos
verantwortlich	serbisch
ostdeutsch	fähig
bekannt	gewählt
beliebt	etabliert
mutig	führend
populär	bürgerlich
jung	regierend
führend	beliebt
palästinensisch	verlogen
aktiv	konservativ
bürgerlich	inkompetent
link	amtierend
einzig	prominent
gewiss	maßgeblich
österreichisch	mächtig
gut	bekannt
erst	ahnungslos
ander	heutig
	/…/

Tabelle 3: Adjektiv-Substantiv-Kookkurrenzen zu *Politikerin* und *Politiker*

Nehmen wir die Adjektiv-Substantiv-Kookkurrenzen[9] zu den Basiswörtern *Politikerin* und *Politiker* als Beispiel. Hier stellt die deutliche qualitative Gen-

[9] Die Adjektiv-Substantiv-Kookkurrenzenlisten im Wortabstand 1 wurden anhand des Webkorpus von Tobias Roth generiert (Roth 2012a: 32–33).

derasymmetrie[10] innerhalb der Kookkurrenzenlisten ein großes Problem dar. Während die soziolinguistische Forschung von der Kookkurrenzanalyse tiefe Einblicke in gesellschaftliche Geschlechterdiskurse erwartet (Baker 2008: 77), muss die Kollokationslexikographie sich fragen, bis zu welchem Grad sie diesen Diskurs abbilden soll oder darf. Die Adjektiv-Substantiv-Kookkurrenzen von *Politikerin* und *Politiker* vermitteln folgendes Bild (vgl. Tabelle 3).

Alle für *Politikerin* aufgeführten Adjektive erscheinen auch in der weitaus längeren Kookkurrenzenliste zu *Politiker*. Allerdings scheint sich nur die maskuline Form mit pejorativen Adjektiven wie *korrupt*, *verantwortungslos* und *verlogen* zu verbinden. Dabei kann allerdings wie weiter oben ausgeführt häufig nicht erschlossen werden, ob es sich um generische oder geschlechtsspezifische Formen handelt. Schauen wir auf die Notation in einem Wörterbuch: Die Nennform von Kollokationen im Wörterbuch steht in der Regel im Singular, sodass folglich *verantwortungsloser Politiker* angelegt würde. Dadurch wird jedoch die pseudoneutralisierende Wirkung des generischen Maskulinums (wie evtl. in *verantwortungslose Politiker*) aufgehoben: die Nennform wird nun ausschließlich auf männliche Personen bezogen. Die Belege, aus welchen diese lexikographische Information abgeleitet wurde, bezogen sich allerdings zu unklaren Anteilen eindeutig auf männliche Personen und auf geschlechtsunspezifische Gruppen. Politikerinnen würden in einem beträchtlich kürzeren Artikel als *konservativ*, aber eben vor allem als *prominent*, *engagiert*, *freiheitlich*, *verantwortlich* und *beliebt* beschrieben. Ohne redaktionelle Überarbeitung dieser Ergebnisse der Kookkurrenzanalyse ergäben sich schlussendlich sowohl quantitativ als auch inhaltlich stark unterschiedliche Artikel für *Politikerin* und *Politiker*. Für die Benutzenden wäre diese Diskrepanz jedoch weder auflösbar noch produktiv.

Wichtige Voraussetzung für eine symmetrische Beschreibung männlicher und weiblicher Personenbezeichnungen in Kollokationenwörterbüchern sind geeignete Korpora und eine erhöhte Sensibilität für die Tücken des generischen Maskulinums. Bis dato – und speziell in unserem Projekt – bleibt nur der Weg, redaktionell die gravierendsten Verzerrungen zu entzerren und die größten sprachlichen Lücken durch eine belegbasierte Nachforschung zu schließen.

[10] Pober (2007: 42) definiert die Genderasymmetrie einerseits als die „ungleiche Versprachlichung der Geschlechter, die durch morphologische, semantische und syntaktische Strukturen gestützt wird", sieht sie andererseits aber auch in der „lexikologische[n] und -grafische[n] Bevorzugung des Maskulinums in der Wörterbuchgestaltung" selbst realisiert. Das generische Maskulinum als ein Phänomen von Genderasymmetrie deckt beide dieser Aspekte ab.

Dabei müssen genderlinguistische Argumente und Forderungen[11] ebenso berücksichtigt werden wie die historisch-kulturelle Bedingtheit der Korpora. Die größte Schwierigkeit bereitet hierbei die Ambiguität des generischen Maskulinums, die sich im Falle der Kollokationen kaum quantitativ hinsichtlich der geschlechtsspezifischen und der generischen Verwendung auflösen lässt. Da gleichzeitig definitorische Kriterien wie bspw. die Festigkeit der Wortverbindung zu beachten sind, ist das Auffüllen gendersprachlicher Lücken (Pober 2007: 169–175) für die Kollokationslexikographie ein anspruchsvolles Unterfangen. Nicht minder anspruchsvoll gestaltet es sich, eine geeignete, übersichtliche Darstellung für komplexe Nennformen zu finden.[12]

Im Falle von *Politikerin* und *Politiker* bietet sich die konsequente Symmetrierung der Wortverbindungen an, also eine Gleichstellung der männlichen und weiblichen Form hinsichtlich ihres Kollokationsspektrums. Damit können nicht nur die gendersprachlichen Lücken der femininen Basislemmata geschlossen werden, sondern es wird auch verhindert, dass – hier pejorative – pseudogenerische Bezeichnungen das Ansetzen einer Nennform im Singular zu geschlechtsspezifischen männlichen Personeneigenschaften werden. Damit trifft man im Kollokationenwörterbuch sowohl auf die Nennform *verantwortungslose Politikerin* als auch auf den Eintrag *verantwortungsloser Politiker*, und man erfährt, dass die Stimmberechtigten *eine_n Politiker_in wählen*.

Es hat sich gezeigt, dass die nach wie vor dominante Verwendung des generischen Maskulinums und die kulturell-historische Bedingtheit der Korpora in der Kookkurrenzanalyse Folgen haben, die eine manuelle Nachbearbeitung unumgänglich machen. Eine Genderasymmetrie aufgrund der korpusbasiert-quantitativen Altlasten aufrechtzuerhalten erweist sich als nicht zeitgemäß und entspricht auch nicht der angestrebten Nutzungsfreundlichkeit eines Kollokationenwörterbuchs. Nur so erfüllt ein (Kollokationen-)Wörterbuch sein Ziel, „alle Phänomene sagbar und beschreibbar“ (Pober 2007: 36) zu machen – auch *korrupte Politikerinnen*.

[11] So hat bspw. die Dudenredaktion mittlerweile auf die Forderung, sowohl männliche als auch weibliche Personenbezeichnungen zu versprachlichen, reagiert (Kunkel-Razum: 2004). Die nach wie vor bestehenden Genderasymmetrien im Duden hat Pober (2007: 86–199) einer genaueren Untersuchung unterzogen. Pober (2007: 119) kritisiert dort u. a. genau, dass in den Großen Duden von 1993–95 und 1999 Gendersymmetrie „nur quantitativ nicht jedoch inhaltlich“ vollzogen wird.

[12] Im Fall des Kollokationenwörterbuchs wurde für die Darstellung aller Geschlechter in den Nennformen der gender gap von Herrmann (2003: 22) gewählt. Im Gegensatz zum Binnen-I ist er auch auf Artikel (eine_n) und Adjektive (korrupte_r Politiker_in) anwendbar und versprachlicht konzeptionell nicht nur Männer und Frauen, sondern auch queere Geschlechtsidentitäten.

4 Komposita im Kollokationenwörterbuch – Widerspruch oder Notwendigkeit?

Im Laufe des Projekts hat sich gezeigt, dass sich bei Sprachwissenschaftlerinnen und Sprachwissenschaftlern oftmals eine gewisse Verwunderung einstellt, wenn sie hören, dass Kollokationen und Komposita in ein Wörterbuch gepackt werden sollen. Abgesehen vom definitorischen Kriterium der Polylexikalität für Kollokationen sehen wir allerdings keine derart großen Differenzen, die eine solche Erwägung von vornherein ausschließen würde.[13] Auch Hausmann (2004: 318) schreibt dem Kompositum Kollokationscharakter zu.[14] Aus semantischer und pragmatischer Perspektive sind Kollokationen und Komposita häufig gegeneinander austauschbar und sie bestehen nicht selten sogar aus dem gleichen Wortmaterial.

Zudem haben Komposita und Kollokationen nicht nur stilistische, sondern auch strukturelle Gemeinsamkeiten. Abgesehen vom fehlenden Spatium bilden auch die meisten Komposita eine feste Einheit von mindestens zwei Worten bzw. Lexemen, wie es die gängigen Definitionen von Kollokationen verlangen (Hausmann 1985: 118ff. und Burger 2010: 5ff.). Genauso wie Kollokationen werden auch Komposita mental als Einheit abgerufen. Sie besitzen eine psycholinguistische Festigkeit und eine falsche Komposition zweier Lexeme kann zu markiertem Sprachgebrauch führen. Beim Übersetzen in eine Fremdsprache bereiten Komposita oft die gleichen Probleme wie Kollokationen, da beide Formen häufig nicht wortwörtlich übersetzt werden können. Zudem ist in L1 nicht selten ein Kompositum die gängigste Formulierung, während in L2 eine andere Formulierung gewählt werden muss: Zucker in Form eines Würfels heißt im Deutschen nicht „Zucker in Stücken“ (frz. *sucre en morceaux*) sondern *Würfelzucker*.

Die folgende Erzählung mag als Einstieg in die Problematik dienen, weil sie stilistische Aspekte der Textproduktion in Bezug auf Kollokationen und Komposita verdeutlicht.

[13] Vgl. Roth (2012b: Kap. 4.2: Komposita im Kollokationenwörterbuch) der im Sinne des Wörterbuchs der gebräuchlichen und typischen Wortverbindungen im Deutschen dafür plädiert, Komposita in Kollokationenwörterbücher aufzunehmen, da sie insbesondere im Deutschen „quantitativ wie qualitativ einen wichtigen Teil der kombinatorischen Begriffsbildung ausmachen.“

[14] Hausmanns Basis-Kollokator-Konzept lässt sich genauso wie auf Kollokationen auch auf Komposita anwenden – mit denselben Vor- und Nachteilen (Hausmann 1985 und 2008).

> Franziska möchte Kaffee kochen. Sie schüttet die fermentierten Bohnen der entsprechenden Pflanze in die Maschine, die sie zur Herstellung des Kaffees benutzt. Sie drückt den Knopf und der Mechanismus zum Mahlen der Bohnen beginnt zu rattern. Der frisch gebrühte Kaffee fließt aus der Maschine in eine Tasse, die blau ist wie der Himmel. Jetzt noch ein Stück Zucker, das in die Form eines Würfels gepresst wurde, und einen Schuss haltbare Sahne, die üblicherweise für Kaffee verwendet wird. Franziska setzt sich an den Tisch in ihrer Küche und geniesst ihre Pause, in der sie immer einen Kaffee trinkt.

Die Beschreibung eines alltäglichen Vorganges, hier des Kaffeekochens, ist ein plausibler Anlass für eine Textproduktion. Die kleine Erzählung vom Kaffeekochen ist verständlich, orthographisch und syntaktisch fehlerfrei und enthält sogar verschiedene Kollokationen. *frisch gebrühter Kaffee* ist beispielsweise eine sehr feste Kollokation, für die kaum ein stilistisch unauffälligerer Ausdruck gefunden werden könnte. Dennoch wirkt der Text aufgrund seiner umständlichen Formulierungen unnatürlich, was daher rührt, dass keine Komposita verwendet wurden. Mit Komposita könnte der Bericht neu wie folgt lauten:

> Franziska möchte Kaffee kochen. Sie schüttet Kaffeebohnen in die Kaffeemaschine. Auf Knopfdruck beginnt das Mahlwerk zu rattern. Der frisch gebrühte Kaffee fließt aus der Maschine in eine himmelblaue Tasse. Jetzt noch ein Stück Würfelzucker und einen Schuss Kaffeesahne. Franziska setzt sich an den Küchentisch und geniesst ihre Kaffeepause.

Die zweite Beschreibung exakt desselben Vorganges wirkt mit den Komposita viel klarer und verständlicher. Denn genauso wie der adäquate Umgang mit Kollokationen gehört auch der angemessene Gebrauch von Komposita zur stilistisch ansprechenden Textproduktion. Da wir in der Textproduktion einen – möglicherweise sogar den – typischen Anwendungsfall von Kollokationenwörterbüchern sehen, erscheint es uns konsequent auch Komposita in ein Kollokationenwörterbuch aufzunehmen. Um die Benutzenden in einem unmarkierten Sprachgebrauch zu unterstützen, ist ein Hinweis auf ein adäquat gebildetes Kompositum ebenso hilfreich wie der Verweis auf eine geläufige Kollokation. Zudem stehen Kollokationen und Komposita häufig in einer stilistischen Konkurrenz: Die Tasse in der obigen Geschichte beispielsweise kann als *himmelblau* oder als *blau wie der Himmel* beschrieben werden. Solche Konkurrenzen von Kollokationen und Komposita mit gleichem Wortmaterial treffen wir häufig an, wobei sie teilweise semantisch sehr ähnlich sind (wie im Fall der Tasse) aber auch sehr verschieden sein können (bspw. *Individualreise* vs. *individuelle Reise*). Diese Gleich- und Verschiedenheit kann ein Kollokationenwörterbuch nur aufzeigen, wenn es Kollokationen und Komposita nebeneinander aufführt.

4.1 Redaktionelle Herangehensweise

Durch die Aufnahme der Komposita in das Kollokationenwörterbuch ergeben sich einige Entscheidungsanforderungen, die redaktionell zu bearbeiten sind. Die größte liegt wohl darin, dass verschiedene Formulierungen für ein Denotat konkurrieren können: Komposita, Genitivkonstruktionen, Präpositionalphrasen oder Konstruktionen mit einer Konjunktion. *Milchkaffee* konkurriert mit *Kaffee mit Milch*, *blau wie der Himmel* konkurriert mit *himmelblau* und *Szenario des Schreckens* konkurriert mit *Schreckensszenario.* Dabei ist von redaktioneller Seite stets abzuwägen, ob eine semantische Äquivalenz oder Bedeutungsverschiedenheit vorliegt. Gegebenenfalls muss eine Entscheidung für die gebräuchlichere Formulierungsalternative getroffen werden oder Unterschiede müssen mit Beispielsätzen illustriert werden. Mithilfe der Korpusbelege und den Angaben zur Frequenz kann redaktionell entschieden werden, ob eine Verbindung als Kollokation, als Kompositum oder in beiden Varianten aufgenommen werden soll. Diese Entscheidungen reichen weit über die Kollokations- und Kompositadefinitionen hinaus und obliegen der Erfahrung und Sachkenntnis der Lexikographinnen und Lexikographen.

5 Schluss: Das Ausmass der redaktionellen Bearbeitung beeinflusst den Informationsgehalt eines Kollokationenwörterbuchs

Auch wenn heute niemand die Notwendigkeit einer manuellen Nachbearbeitung der Korpusanalyseergebnisse in Frage stellt, ist es uns mit diesem Aufsatz ein Anliegen, Aspekte dieser Nachbearbeitung aus der praktischen Perspektive einer Wörterbuchredaktion zu beleuchten. Neben den erwähnten Fallbeispielen gibt es zahlreiche weitere lexikographische Arbeitsschritte, die zumindest einer fachlichen Revision bedürfen. Dazu gehören speziell die Formulierung von anwendbaren Nennformen, die Konkretisierung von Leerstellen und Platzhaltern innerhalb der Kollokationen sowie die Auswahl geeigneter Beispielsätze. Im Falle der drei geschilderten Beispiele aus der Praxis zeigt sich, dass immer dann viel Arbeit auf die Redaktion zukommt, wenn die Korpusbasis keine geeignete Ausgangslage für die Erhebung von Kollokationen bietet (bspw. im Fall von generischen Maskulina), wenn die statistikbasierten Kookkurrenzenlisten nicht aussagekräftig genug sind (bspw. bei der Relevanz der Kollokationen) oder wenn mehrere Ausdrucksvarianten in Konkurrenz treten (bspw. bei Komposita). Als Korrektiv für sprachlichen Phänomenen die unausgewogen in den Korpora repräsentiert sind und für Bereiche der Sprache, die korpusanalytisch kaum zugänglich sind (bspw. die Semantik), bleibt der Redaktion des *Wörterbuchs der typischen und gebräuchlichen Wortverbindungen im Deut-*

schen nur die sprachliche Erfahrung und Intuition, die entsprechend durch Korpusbelege abzusichern ist. Letztlich ist für ein Wörterbuch, das nicht den Bestand an Kollokationen katalogisieren, sondern ein benutzbares Nachschlagewerk für die Sprachproduktion sein soll, nicht die Frage entscheidend, ob eine Kookkurrenz mittels des einen oder des anderen Korpusanalyseverfahrens aufgedeckt wird. Vielmehr geht es darum, die richtigen Kandidaten für das Wörterbuch auszuwählen und zwar im Hinblick auf die Relevanz für die Nutzenden einerseits und auf die Repräsentativität für den Sprachusus andererseits. Dieses Verfahren wird dadurch erschwert, dass gängige Definitionen von Kollokation sehr theoretisch formuliert und in der Praxis oftmals nicht eindeutig anzuwenden sind. Alles in allem entfällt ein erheblicher Teil der Produktionszeit des *Wörterbuchs der typischen und gebräuchlichen Wortverbindungen im Deutschen* auf die manuelle und intellektuelle Weiterverarbeitung der empirischen Sprachrohdaten. Wir sind allerdings der Überzeugung, dass sich dieser arbeitszeitintensive Aufwand lohnt, da sich der informative Gehalt des Kollokationenwörterbuchs durch die redaktionelle (Nach-)Bearbeitung erheblich erhöht. Ohne die korpuslinguistischen Vorarbeiten aber, wäre ein solches Wörterbuch nicht denkbar. Man muss sich vergegenwärtigen, dass korpusbasierte Wörterbücher wie das beschriebene sich nicht mehr aus ihren Vorgängern (bspw. *Duden. Stilwörterbuch*) speisen, sondern empirische Daten des aktuellen Sprachgebrauchs repräsentieren.

Literatur

BAKER, Paul, 2008: „Eligible“ Bachelors and „Frustrated“ Spinsters: Corpus Linguistics, Gender and Language. Harrington, Kate u. a. (Hrsg.): *Gender and Language Research Methodologies.* Hampshire: Palgrave Macmillan. 73–84.

BURGER, Harald, 2004: Phraseologie – Kräuter und Rüben? Traditionen und Perspektiven der Forschung. Steyer, Kathrin (Hrsg.): *Wortverbindungen – mehr oder weniger fest.* Berlin u. a.: de Gruyter. 19–40.

BURGER, Harald, 2007: *Phraseologie. Eine Einführung am Beispiel des Deutschen.* 3. Aufl. Berlin: Erich Schmidt.

BURGER, Harald, 2010: *Phraseologie. Eine Einführung am Beispiel des Deutschen.* 4. Aufl. Berlin: Erich Schmidt.

BUSSMANN, Hadumod u. a., 2003: Engendering female visibility in German. Hellinger, Marlis u. a. (Hrsg.): *Gender Across Languages. Vol. 3.* Amsterdam: Benjamins. 151–174.

DWDS Kernkorpus: <www.dwds.de/>. Zugriff 10. 12. 2012.

HAUSMANN, Franz J., 1985: Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. Bergenholtz, Henning u. a. (Hrsg.): *Lexikogra-*

phie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.–30.6.1984. Tübingen: Niemeyer. 118–129.

HAUSMANN, Franz J., 2004: Was sind eigentlich Kollokationen? Steyer, Kathrin (Hrsg.): *Wortverbindungen – mehr oder weniger fest*. Berlin u. a.: de Gruyter. 309–334.

HAUSMANN, Franz J., 2008: Kollokationen und darüber hinaus. *Lexicographica. Internationales Jahrbuch für Lexikographie* 24, 1–8.

HERRMAN, Steffen Kitty (aka S_he), 2003: Performing the Gap – Queere Gestalten und geschlechtliche Aneignung. *Arranca!* 28, 22–26. <http://arranca.org/ausgabe/28/performing-the-gap>. Zugriff 30. 11. 2012.

KUNKEL-RAZUM, Kathrin, 2004: Die Frauen und der Duden – der Duden und die Frauen. Eichhoff-Cyrus, Karin M. (Hrsg.): *Adam, Eva und die Sprache*. Mannheim etc.: Dudenverlag. 308–315.

MANNING, Christopher D. / SCHÜTZE, Hinrich, 1999: *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

McINTOSH, Colin (Hrsg.), 2009: *Oxford Collocations Dictionary for students of English*. Oxford: Oxford University Press.

POBER, Maria, 2007: *Gendersymmetrie. Überlegungen zur geschlechtersymmetrischen Struktur eines Genderwörterbuches im Deutschen*. Würzburg: Königshausen und Neumann.

PUSCH, Luise, 1983: Das Deutsche als Männersprache. Diagnose und Therapievorschläge. Pusch, Luise (Hrsg.): *Das Deutsche als Männersprache*. Frankfurt am Main: Suhrkamp. 46–68.

ROTH, Tobias, 2012a: Using Web Corpora for the Recognition of Regional Variation in Standard German Collocations. Kilgarriff, Adam u. a. (Hrsg.): *Proceedings of the Seventh Web as Corpus Workshop (WAC7)*. 31–38. <https://sigwac.org.uk/raw-attachment/wiki/WAC7/wac7-proc.pdf >. Zugriff: 9. 12. 2012.

ROTH, Tobias, 2012b: *Wortverbindungen und Verbindungen von Wörtern. Lexikografische und distributionelle Aspekte kombinatorischer Begriffsbildung zwischen Syntax und Morphologie*. Dissertation. Basel: Universität Basel. Unveröffentlichtes Manuskript.

Schweizer Textkorpus: <http://chtk.unibas.ch/search>. Zugriff 10. 12. 2012.

STEYER, Kathrin, 2001: Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. *Deutsche Sprache* 2/28, 101–125.

Étude du figement dans les *Curiositez françoises* (1640) d'Antoine Oudin

Claire Ducarme (Liège)

Abstract

For old languages, it is not possible to study the frozenness using the competence of modern speakers nor the intuition about the value of frozen structures (FSs) of modern languages. Finding a way to analyse FSs in an old language is the main goal of my research. I tackle this issue from a particular corpus: the dictionary *Curiositez françoises* (1640) of Antoine Oudin. This dictionary collects different polylexical units from the first part of the 17th century. This lexical diversity arises a reflection about the differentiation and the identification of the lexical unit types. In this paper, I will focus on internal (implicitly or explicitly given by the lexicographer) and external (obtained through the consultation of other lexicographic and literary sources) indications allowing to identify FSs and to determine their status. My approach will go back and forth between concrete case study and a general theory.

1 Introduction

Notre recherche repose sur l'étude lexicographique et lexicologique du dictionnaire intitulé *Curiositez françoises pour supplement aux dictionnaires* (1640) d'Antoine Oudin. Antoine Oudin, polyglotte, est, à la suite de son père, traducteur et interprète du roi. En même temps que ses charges officielles, il poursuit une carrière de grammairien et de lexicographe. Son dictionnaire possède plusieurs spécificités. Destiné aux étrangers, il peut être qualifié de spécialisé, en ce sens qu'il recueille un vocabulaire bas, vulgaire. Antoine Oudin montre, à travers cet ouvrage, qu'à côté de la langue normée, il existe un vaste matériel linguistique.[1] Une autre grande spécificité de cet ouvrage relève de la diversité des unités formant la nomenclature. En effet, celle-ci comporte aussi bien des unités simples que des unités polylexicales, phraséologiques. Le sous-titre du dictionnaire, *Recueil de plusieurs belles proprietez, avec une infinité de Proverbes & Quolibets, pour l'explication de toutes sortes de Livres*, témoigne de l'importante place qu'y occupe ce second type d'unité.

[1] Ce vaste matériel linguistique est traité avec un certain souci d'exhaustivité: les *Curiositez françoises* comportent quelque 8000 entrées.

L'étude lexicologique des unités polylexicales nécessite au préalable de les identifier précisément au sein de la nomenclature. Même si la question de l'identification des unités phraséologiques semble, de prime abord, poser moins problème dans le cadre d'un dictionnaire que dans celui d'un texte littéraire, étant donné que l'unité est présente en mention dans l'entrée, elle ne va cependant pas de soi. Ainsi, nos diverses lectures et recherches nous ont montré qu'il n'est pas rare, au 17^{e} siècle, qu'un lexicographe rende compte d'une unité phraséologique en la plaçant dans un contexte plus large, ce qu'Alain Rey (1973) a montré à propos des phraséologismes dans le dictionnaire de Furetière (1690). Pour rendre compte de la séquence polylexicale verbale *avoir le feu au cul*, Fur 1690 propose ceci: «On dit d'un homme qui s'enfuit fort vîte, qu'il court comme s'il avoit le feu au cul», énoncé où la séquence phraséologique est contextualisée au moyen de la séquence libre *il court comme si*. Certaines unités sont dès lors parfois mal dégagées d'un contexte libre et facultatif. La question de l'identification des unités phraséologiques au sein d'un dictionnaire ancien est donc cruciale. Elle se ramène à la question suivante: parmi les formes qui font l'objet d'un traitement lexicographique, lesquelles répondent à la définition que l'on donne actuellement d'une unité phraséologique? C'est à cette question que nous tenterons de répondre en exposant notre cadre théorique et notre méthodologie de travail (1) et en mettant cette dernière à l'épreuve de l'étude de quelques cas concrets (2).

2 Cadre théorique et méthodologie de travail

2.1 Avant de présenter notre méthodologie d'analyse proprement dite, il est important de préciser le cadre théorique sur lequel elle repose. Que considérons-nous comme unité phraséologique? La terminologie concernant le figement est foisonnante (*locution*, *séquence*, *idiotisme*, *expression idiomatique*, *lexie complexe*, *proverbe*, *quolibet*, *dicton*…) et chaque terme ne recouvre pas nécessairement la même réalité chez les différents linguistes. Il est donc nécessaire de préciser la terminologie que nous employons dans notre étude. À la suite de Salah Mejri (1997), nous considérons le figement comme un processus de solidification en langue de plusieurs unités qui, au départ, sont libres dans le discours. Apparaît dès lors la notion de séquence figée (SF) définie comme un groupement d'unités solidifié en langue. Le figement est appréhendé comme un processus qui s'inscrit dans le temps (point de vue diachronique), faisant d'un groupement d'unités lexicales libres une séquence appartenant à la langue (cf. notion de polylexicalité chez Gross 1996 et Mejri 1997, 2003). Identifier une SF, c'est donc identifier un signe linguistique possédant une forme complexe à laquelle est associée un sens conventionnel, global, inscrit en langue.

Le processus sémantique à l'origine est appelé *globalisation* par Mejri (2003: 28) et peut être défini comme «l'opération par laquelle la pluralité est ramenée à l'unicité: la SF dont le signifiant est polylexical ne peut avoir de signification qui lui correspond en tant qu'unité que lorsque la globalisation intervient pour opérer la synthèse sémantique nécessaire à l'unicité sémantique exigée par la SF». Elle est sous-tendue par la valeur intentionnelle, c'est-à-dire la non actualisation des composants. Mejri (2003: 28) parle dans ce cas du phénomène de *conceptualisation*, qui prend appui sur des unités linguistiques autonomes référant, à l'origine, à leurs propres concepts. Comme ceux-ci sont suspendus par la non actualisation, ils peuvent servir à l'émergence du nouveau concept. Globalisation et conceptualisation concourent à la rupture entre le sens des composants et le sens global, car il y a toujours un enrichissement sémantique. Suivant le rapport qu'entretiennent le sens compositionnel, c'est-à-dire produit par la somme des composants, et le sens global, la SF sera plus ou moins transparente. Le tout sémantique est figuré iconiquement, dans le cas des lexies simples, par l'unité formelle et, dans le cas des polylexèmes, par une certaine rigidité de la forme. En ce qui concerne la forme, deux précisions doivent être apportées. Identifier la forme d'une SF, c'est, d'une part, déterminer précisément la portée du figement, c'est-à-dire le nombre réel d'unités touchées par ce processus et, d'autre part, établir un cadre variationnel. Selon nous, trop souvent, les études sur le figement ont laissé de côté la variation. Or, comme toute unité de langue, et encore plus par le fait de sa polylexicalité, toute SF possède potentiellement des variations. Celles-ci sont plus ou moins limitées, d'où notre terme de 'cadre', tant sur le plan paradigmatique (le nombre de variantes est restreint) que syntagmatique (les variations ne s'opèrent souvent qu'en un point de la séquence). La variation doit être également envisagée au niveau sémantique. La non-compositionalité du sens est, selon nous, un facteur important favorisant la polysémie et l'évolution du sens.

2.2 Si nous avons précisé ce que nous entendons par l'identification d'une SF, nous n'avons pas encore abordé la question essentielle du comment réaliser cette identification dans un état de langue ancien. La plupart des études sur le figement (cf. entre autres les travaux de Kleiber, Mejri), du moins pour le français,[2] concernent un état actuel de langue et peuvent prendre appui sur des tests transformationnels et commutatoires, lesquels sont soumis à un jugement de recevabilité par des locuteurs compétents. Pour un état de langue ancien, l'absence de locuteurs et de situations de contrôle/d'expérimentation empêche de tels jugements. Les outils développés en vue de l'étude des SF

[2] Il existe quelques études sur les SF en ancien français (Buridant 1984, Zumthor 1976), ainsi qu'en français des 16^{e} et 17^{e} siècles (Biason 1992, Kramer 2000, Rey 1973).

d'un état de langue actuel sont insuffisants voire insatisfaisants pour notre recherche. Néanmoins, nous avons d'autres outils à notre disposition. En effet, nous n'avons plus de locuteurs compétents, mais nous avons des témoins contemporains à travers les textes et les sources lexicographiques; certes, nous ne pouvons opérer des tests transformationnels et commutatoires, mais nous avons des indices formels à notre disposition. Notre démarche d'analyse consiste dès lors à prendre appui sur les diverses données à notre disposition, à les croiser et à établir ainsi un faisceau d'indices permettant d'assurer l'identification d'une SF.

Les indices peuvent être répartis en deux catégories. La première regroupe ceux, ici qualifiés d'*internes*, qui sont présents au sein du dictionnaire et la seconde ceux, ici qualifiés d'*externes*, qui proviennent d'autres ouvrages. Au sein des indices externes, nous distinguons ceux qui proviennent de sources secondaires, c'est-à-dire des données textuelles attestant la vie de l'expression dans la langue, et de sources tertiaires, c'est-à-dire de données lexicographiques montrant une attention à l'expression portée par les descripteurs. Comme nous l'avons déjà dit, les sources primaires, c'est-à-dire la langue orale qui véhicule majoritairement les SF, font défaut. Au sein des indices internes, nous distinguons les indices formels et sémantiques, et les indices (méta)lexicographiques. Les premiers relèvent des caractéristiques propres à la séquence en tant qu'elle est figée. D'un point de vue sémantique, la non compositionalité du sens constitue un indice fort. En effet, l'opacité sémantique provoquée par celle-ci est le signe visible du processus de globalisation propre au figement. Il faut néanmoins préciser que, selon nous, le sens d'une SF, bien que global, peut se révéler transparent et donc quasi-identique au sens compositionnel. D'un point de vue formel, la structure particulière de la séquence constitue, elle aussi, un indice dans le processus d'identification. Par exemple, la séquence qui ne suit pas la syntaxe propre à une période donnée peut se révéler être le résultat du figement d'une structure appartenant à un état antérieur de la langue.[3] De plus, en ce qui concerne particulièrement les séquences figées proverbiales, quelques études (Mejri 1997, 2001, Anscombre 2000) ont mis au jour des schémas rythmiques en nombre restreint. Plus précisément, selon Mejri, tout proverbe possède un binarisme structurel caractéristique rendu par une diversité syntaxique. Les indices (méta)lexicographiques relèvent, quant à eux, du commentaire lexicographique, explicite ou implicite. Selon nous,

[3] Mirella Conenna (1988), dans un article consacré à un lexique-grammaire comparé, traite des proverbes qui, en italien comme en français, ont comme sujet une proposition relative sans antécédent: *qui*, *chi*. Elle remarque qu' «en italien, il s'agit d'une construction habituelle de la langue, tandis qu'en français, l'absence de l'antécédent en fait plutôt une forme archaïque» (1988: 106).

l'analyse du discours lexicographique, si celui-ci est cohérent, nous indique le statut des formes en entrée.

Dans l'approche de cas concrets, notre première étape consiste à prendre appui sur les données présentes dans le dictionnaire en tenant compte des pratiques lexicographiques avant de confronter ces données à des données externes, qui doivent permettre d'attester l'existence d'une séquence (forme et sens) en langue. Dans l'état actuel de notre recherche, nous avons consulté comme sources textuelles des ouvrages littéraires appartenant au style burlesque: *L'histoire comique de Francion*, 1623 (éd. Giraud 1979) et *Le berger extravagant*, 1627 (éd. Originale) de Charles Sorel; *Les ramonneurs*, 1624 (éd. Gill 1957); *La comédie des proverbes*, 1633 (éd. Kramer 2003); *Le Jodelet ou le Maistre valet*, 1645 (éd. Dickson 1986) et *Le Virgile travesti*, 1650 (éd. Fournel 1858) de Paul Scarron. Afin d'identifier les SF dans ces sources, nous avons opéré une lecture continue des ouvrages, lecture éclairée par notre connaissance des *Curiositez.* Nous avons également consulté les glossaires d'édition ainsi que l'article de M. Kramer (2002) portant sur les sources littéraires des *Curiositez*. Nous avons aussi eu recours aux bases de données Frantext et Google books. En ce qui concerne les sources tertiaires, nous avons consulté plusieurs dictionnaires du 17^{e} siècle: Nicot 1606, Cotgrave 1611, Richelet 1980, Furetière 1690 et Académie 1694. Nous avons également utilisé le Littré (1873) ainsi que le *Französisches Etymologisches Wörterbuch* (FEW).

3 Étude de cas

Les divers types d'indices que nous venons de présenter peuvent être illustrés par de nombreux exemples. Nous reprenons ici, pour notre étude de cas, les plus représentatifs, grâce auxquels nous pouvons illustrer l'idée de faisceau d'indices. Nous distinguons les cas avérés, c'est-à-dire ceux pour lesquels les données externes nous permettent de prouver l'inscription d'une séquence dans la langue, et les cas où cette appartenance est moins évidente.

3.1 Dans la plupart des cas, il est assez facile de trouver des attestations de la séquence dans des sources secondaires et tertiaires.

Cas 1

- *il mord à la Grappe. i. *il est ravy, il prend un extreme plaisir.*
- Mordre à la grappe, *voyez à* grappe.

Dans cet exemple, plusieurs indices internes de différents niveaux nourrissent l'analyse. Le premier relève du sens. La glose définitionnelle montre bien la non-compositionalité du sens. L'opacité sémantique est un des critères essentiels du figement en langue. Le second tient à la pratique lexicographique qui consiste à traiter deux fois,[4] sous deux mots clés différents, une même séquence, apparaissant parfois sous des formes légèrement distinctes. Dans ce cas 1, la première forme semble être une forme actualisée de la SF verbale *mordre à la grappe*. Le *il* apparaît comme une forme neutralisée du sujet. Les sources secondaires et tertiaires[5] viennent confirmer qu'il s'agit d'une SF verbale pouvant être actualisée de différentes manières.

- Celuy qui la lisoit /…/ proferoit les mots avec un ton de Comedien, et il sembloit qu'il mordist à la grappe. (Sorel 1623: 174)
- Pour vous faire mordre à la grappe, écoutez ce que de bon cœur je prétends donner au vainqueur. (Scarron 1650: 185)
- Cotgr 1611, s. v. *mordre*: Il sembloit mordre à la grappe. *He spoke so heartily that he seemed to doe what he delivered.*
- Acad 1694, s. v. *mordre*: On dit aussi, qu'*Un homme mord à la grape*, Quand il escoute avec plaisir un recit, une proposition qu'on luy fait. *Je ne luy ay pas eu plustost fait cette proposition, qu'il a mordu à la grape.*

Cette pratique lexicographique qui consiste à actualiser une SF verbale au moyen d'un sujet postiche se rencontre souvent dans les *Curiositez* et peut s'effectuer au moyen de divers pronoms: *vous estes bien loin de vostre compte* (*être bien loin de son compte*), *j'y perds mon latin* (*y perdre son latin*). La problématique liée à ce type de cas tient à la détermination précise de la portée du figement. C'est grâce à un faisceau d'indices que nous pouvons statuer: le pronom personnel sujet appartient-il ou non à la SF? Il nous semble qu'il faut traiter de la même façon les séquences en *comme*, afin de déterminer si l'élément précédent la comparaison appartient pleinement ou non à la SF[6]. De plus, les données externes viennent préciser le sens de la séquence. Dans ce cas précis, la variation sémantique tient à la détermination de celui qui *mord à la grappe*. Pour Sorel 1623 et Cotgr 1611, c'est le locuteur qui *mord à la*

[4] Il se peut que le traitement lexicographique se limite à un renvoi.

[5] L'expression est attestée dans Cotgr 1611, s. v. *grappe*, s. v. *mordre*; Rich 1680, s. v. *grape*; Acad 1694, s. v. *grappe*, s. v. *mordre*; Littré, s. v. *mordre*, s. v. *grappe*; FEW16, 359b, **krappa*; FEW 6/3, 127a, *mordere*.

[6] Par exemple, on peut s'interroger sur le degré de liberté de l'élément précédant le *comme* dans *il est menteur comme un arracheur de dents*. En effet, dans le discours, on peut avoir *il ment comme…*, *il a menti comme…* (cf. Gross 1996: 119–120 et Mejri 1997: 439–450).

grappe, pour Scarron 1650 et l'Acad 1694,[7] il semble que ce soit le récepteur. Dans les *Curiositez*, l'absence de contexte maintient l'ambigüité.

Cas 2

- *Qui bon l'Achepte bon le boit; vulg. *c'est pour dire qu'il est mieux d'achepter une bonne marchandise cherement*; *qu'une mauvaise à bon marché. Le reste du proverbe est,* ou le respand en chemin.

Si certaines structures phrastiques doivent être considérées comme des actualisations de SF sous-phrastiques, d'autres répondent aux schémas canoniques de SF phrastiques. Ainsi, outre l'indice relevant de l'opacité sémantique, la structure formelle en *qui* avec rythme binaire marqué par la répétition de *bon* constitue un indice fort indiquant que la séquence phrastique dans son ensemble a de grandes chances d'être figée (cf. Mejri 1997: 523). De plus, notons que le lexicographe utilise le terme *proverbe* pour qualifier l'unité en entrée, même s'il faut éviter tout anachronisme: la terminologie de l'époque est flottante et ne correspond pas nécessairement à notre conception actuelle[8]. La séquence est attestée par de nombreuses sources secondaires et tertiaires[9] qui viennent confirmer les données internes. Dans l'extrait suivant datant de 1692, nous voyons que la séquence est à nouveau qualifiée de proverbe.

– /…/ Car selon l'art de la massonnerie, le moins de bois que l'on peut mettre dedans les murs, c'est le meilleur, le fer y estant d'un meilleur usage, quoiqu'il coûte un peu plus, mais qui bon l'achete, bon le boit, dit le proverbe. (de Ferrière[10] 1692: 432)

Notons que certaines sources tertiaires donnent un sens plus précis, considérant que le champ d'application du proverbe concerne le vin.

– NicProv 1606: Qui bon l'achepte, bon le boit. *Optima qui redimunt, optima vina bibunt.*

– Cotgr 1611, s.v. *boire*: He that buyes good wine drinks good wine.

– Fur 1690, s.v. *acheter*: On dit proverbialement en parlant du vin, qui bon l'achete, bon le boit.

– Littré, s.v. *acheter*: Proverbe. Qui bon l'achète, bon le boit, se dit du vin et en général de toute marchandise.

[7] On retrouve également ce sens dans Rich 1680, s.v. *grape*.

[8] Suivant l'état actuel de nos recherches, nous avons rencontré le terme *proverbe* à cinq reprises dans les *Curiositez*. Il est appliqué à chaque fois à des séquences phrastiques.

[9] Les sources tertiaires attestant cette séquence sont: NicProv 1606; Cotgr 1611, s.v. *achepter*, s.v. *bon*, s.v. *boire*; Fur 1690, s.v. *bon*; Acad 1694, s.v. *acheter*; Littré, s.v. *acheter*; FEW 24, 66b, **accaptare*.

[10] Tiré de Google books.

Le sens transmis par Oudin apparaît comme une extension du sens premier, plus transparent, c'est-à-dire plus proche du sens compositionnel.

Cas 3

- à Belles dents, à belles ongles. i. *à force de dents, à force d'ongles.*

La problématique liée à cet exemple et au suivant tient particulièrement à la mise au jour d'un cadre variationnel. Le lexicographe donne en entrée deux séquences adverbiales identiques à l'exception des éléments *dents* et *ongles* qui apparaissent dès lors comme des variantes potentielles. À chaque forme présentée en entrée correspond une périphrase synonymique. Les deux périphrases sont elles aussi identiques à l'exception de *dents* et *ongles* qui gardent leur sens propre. Il y a donc une décompositionalité du sens. La partie la plus obscure de la séquence est *à belles X*, glosée par 'à force de X'. Le sens global de la séquence change suivant ce *X* qui garde son sens plein. Les données textuelles et lexicographiques[11] recensent souvent la forme avec *dent.* Mais nous trouvons également de nombreuses séquences avec des substantifs féminins pluriels comme *griffes*, *mains*, *pierres*, *épées*, *etc.*

- S'il faut qu'entre mes mains ce détestable tombe, le moindre de ses maux est celuy de la tombe: je le deschirerois, le traistre, à belles dents, je l'irois affronter entre cent faux ardens. (Scarron 1645: 31)
- De cela, les autres espouvantées se leverent; et toutes ensemble, comme ceste-là, à belles pierres, se mirent à lapider ceste bouteille. (Beroalde de Verville[12] 1610: 121)
- Afin donc de s'en despestrer, et d'en tirer encore quelque bon office en les laissant, il s'efforca de les oster. Les courroyes de son casque estoient si usees qu'elles furent aysement rompuës, tellement qu'il le prit à belles mains, et le jetta contre ses enemys. (Sorel 1627: 377)

Plus rarement, nous trouvons un substantif masculin pluriel:

- Acad 1694, s.v. *beau*: il me mangeroit volontiers à belles dents, Laniaret me lubens dentibus, Liu. lib. 22. comme on dit, ils s'entrechoquent à belles injures, à beaux coups de poings.

Les données externes confirment et complètent notre analyse en montrant que les variantes peuvent être nombreuses. Elles invitent, de plus, à nous défaire d'un sens imposé par notre conception actuelle de l'expression *à belles dents* dans *manger à belles dents* et signifiant 'manger avec appétit'. Au moyen des indices internes et externes, nous pouvons mettre au jour le cadre variationnel

[11] La séquence est attestée dans Nic 1606, s.v. *dent*; Rich 1680, s.v. *dent* et Acad 1694, s.v. *beau.*

[12] Tiré de Frantext.

de la SF: la variation se fait en un point de la séquence et, bien que limité dans une certaine mesure, le paradigme est assez ouvert. De plus, le sens global de la séquence varie en fonction de la variante choisie. À la suite de Montoro 2011 (avant lui Zuluaga 1980), nous qualifierons ce type de SF, possédant un tel cadre variationnel, de 'SF à case vide'. Dans le cas des locutions à case vide, les unités pouvant remplir cette case appartiennent à une catégorie lexicale pleine (nom, verbe). Les alternatives sont nombreuses même s'il y a certaines limites sémantiques.

Cas 4

- *Qui a de l'Argent a des piroüettes,[13] *ou* des coquilles. i. *qui a de l'argent peut avoir ce qu'il désire.* vulg.

Tout comme le cas 2, nous retrouvons ici des indices internes formels et sémantiques (la séquence possède une structure formelle typique et une certaine opacité sémantique) qui nous invitent à considérer l'unité phrastique comme entièrement figée. Le lexicographe introduit une variante lexicale (*coquilles* peut commuter avec *pirouettes*) au moyen de la conjonction de coordination *ou*. Contrairement à l'exemple précédent, il n'y a pas changement de sens suivant que l'on ait l'une ou l'autre variante: la périphrase définitoire est unique. Les sources tertiaires (nous n'avons pas trouvé de sources secondaires) renseignent principalement la variante *piroüettes.*[14] Nous trouvons chez Cotgr 1611 la variante *chapeaux*.

– Cotgr 1611, s. v. *argent*: *qui a argent a des chapeaux*: Prov. *Most men salute the monyed man*; *or he that hath money hath most things.*

Nous pouvons donc mettre au jour, grâce aux indices internes et externes, le cadre variationnel de la SF: la variation se fait en un point de la séquence, elle est limitée à deux voire trois variantes: *pirouettes*, *coquilles*, *chapeaux*. La forme avec *pirouettes*, citée en premier par le lexicographe, est la plus fréquente. Le sens global de la SF ne varie pas en fonction de la variante choisie. À la suite de Montoro 2011 (avant lui Zuluaga 1980), nous qualifierons ce type de SF de 'SF à variante lexicale'. Contrairement aux séquences à case vide, celles à variantes lexicales possèdent des variantes qui, même si elles peuvent

[13] *Pirouette* est à prendre ici dans le sens de 'jouet composé d'un disque de bois que l'on fait tourner autour d'un pivot passant en son centre'.

[14] La séquence avec la variante *pirouettes* est attestée dans Acad 1694, s. v. *argent*, s. v. *pirouette*; Littré, s. v. *argent*, s. v. *pirouette*. Le FEW mentionne que la première attestation vient d'Oudin: FEW 8, 564b, **pir-*: frm. *qui a de l'argent a des pirouettes* «avec de l'argent on a toute sorte de choses» (OudC 1640 –Ac 1798). La variante *coquilles* est attestée dans Rich 1680, s. v. *coquille*.

être simples (seulement deux options) ou multiples (plus de deux options), forment un paradigme fermé. De plus, contrairement aux SF à case vide, peu importe la variante choisie, il n'y a pas de changement de sens de la séquence. Ce dernier critère est déterminant.

Cas 5

- *Bon jour bon œuvre, *cela se dit quand on fait une mauvaise action un jour de feste remarquable*, vulg.

Dans ce dernier exemple, outre la structure formelle rythmée (rythme binaire marqué par la répétition de *bon*), la pratique lexicographique qui consiste à atteindre le 'contenu' de la séquence en entrée via la description de son emploi constitue un indice fort dans le processus d'identification. De nombreux articles, tout comme l'exemple pris ici, ne possèdent pas de champ définitionnel à proprement parler mais uniquement un champ concernant l'emploi, introduit par une copule du type *cela se dit quand*, *cela se dit à*, *on se sert de cela pour*, … Plus précisément, le lexicographe précise soit le contexte d'emploi, comme dans le cas exposé ici, soit la force illocutoire de l'expression (*on se sert de ce quolibet pour nous mocquer*). Nos recherches montrent que la description de l'emploi en discours concerne principalement les séquences de type phrastique. Ceci nous invite à une analyse lexicologique plus générale: les unités polylexicales phrastiques, tout comme les interjections, par leur autonomie discursive (elles ne peuvent jouer le rôle d'intégrant, cf. Benveniste 1966: 119–131), possèdent un noyau sémantique particulier. Les données externes, qu'elles soient textuelles ou lexicographiques,[15] attestent la SF et viennent confirmer nos analyses. Il faut noter néanmoins que quelques dictionnaires[16] mentionnent deux 'sens', le second étant présenté comme dérivé par ironie du premier.[17]

– Littré, s.v. *jour*: Bon jour, bonne œuvre, se dit en parlant d'une bonne action faite en un jour solennel. Ils se sont réconciliés le jour de Pâques: bon jour, bonne œuvre. Ironiquement et par antiphrase. Bon jour, bonne œuvre, les voleurs font les meilleurs coups les jours de fête.

Dans les textes littéraires, certains auteurs jouent avec cette polysémie. Ainsi, dans *La comédie des proverbes*, alors qu'on soupçonne l'enlèvement d'une jeune fille, un personnage s'écrie:

[15] La séquence (forme et sens) est attestée chez Nic 1606, s.v. *estrenne*; Cotgr 1611, s.v. *oeuvre*; Fur 1690, s.v. *oeuvre*, Acad 1694, s.v. *bon,* s.v. *jour*, s.v. *oeuvre*; Littré, s.v. *jour*, s.v. *oeuvre*; FEW 7, 359b–360a, *opera*.

[16] Cf. Nic 1606, s.v. *estrenes*, Littré, s.v. *jour* et FEW 7, 359b–360a, *opera*.

[17] Fur 1690 et Acad 1694 ne mentionnent que le sens par ironie.

> – Helas mon voisin, plus l'on va en avant, et pis c'est. Il y a d'aussi meschantes gens dans ce monde qu'en lieu où on puisse aller. On dit bien vray qu'une fille est de mauvaise garde, et à bon jour bonne œuvre, aux bonnes festes se font les bons coups. (*La comédie des proverbes*, 1633: 178)

Les données externes viennent compléter les données internes: elles permettent de mettre au jour la variation sémantique de la séquence.

Les différents cas étudiés ici nous ont permis de mettre au jour des indices internes formels (la structure en *qui*, le rythme binaire) et sémantiques (l'opacité sémantique, la décompositionalité ou non du sens), ainsi que des indices (méta)lexicographiques (le double traitement lexicographique, la manière d'atteindre le sens, la manière d'introduire les variantes). Nous avons montré comment les données externes, qu'elles soient textuelles ou lexicographiques, confirment les analyses internes: la forme et le sens de la SF sont attestés et donc assurent l'inscription de la séquence dans la langue. Certaines fois, elles complètent les données internes.

3.2 Si, dans la plupart des cas, il est possible de trouver des attestations des SF dans les sources secondaires et tertiaires, prouvant ainsi que la séquence est bien figée, dans d'autres, l'absence ou le peu d'attestations pose problème. Comment faut-il dès lors analyser ces séquences? Doit-on conclure qu'elles ne sont pas figées? Nous étudions ici trois cas particuliers.

Cas 6

- Rouge au soir & blanc au matin, c'est la journée du pelerin, *le commun applique ce proverbe au temps, & je croy qu'il est mieux de l'entendre du vin.*

D'un point de vue formel, nous avons en entrée une forme phrastique avec un rythme binaire (*Rouge au soir & blanc au matin/c'est la journée du pelerin*) marqué par une assonance ([ɛ̃]). D'un point de vue sémantique, le lexicographe présente deux champs d'application de la séquence. Le premier, relatif au temps, est le fait du *commun*, alors que le second, relatif au vin, apparaît comme le fait du lexicographe s'exprimant en *je*. La forme se trouve attestée dans de nombreuses sources et existe depuis longtemps.[18] Cependant, le dépouillement des sources tertiaires, nous montre peu d'attestations avec le

[18] Il est fait mention de ce constat sur le temps dans l'évangile de Saint Matthieu, chap 16: 2.3: «Jésus leur répondit: 'Quand vient le soir, vous dites: 'Voici le beau temps, car le ciel est rouge.' Et le matin, vous dites: 'Aujourd'hui, il fera mauvais, car le ciel est d'un rouge menaçant.' Ainsi l'aspect du ciel, vous savez l'interpréter; mais pour les signes des temps, vous n'en êtes pas capables.»

sens relatif au vin. Seuls Fur 1690[19] et Littré en rendent compte. Il ne faut pas oublier que ces deux lexicographes ont eu accès aux *Curiositez.* Littré mentionne, par ailleurs, qu'il s'agit d'un sens par plaisanterie.

– Littré, s. v. *pèlerin*: Rouge au soir, blanc au matin, c'est la journée du pèlerin, c'est-à-dire ces deux couleurs du ciel montrent qu'il doit faire beau temps durant le jour. Cela signifie aussi par plaisanterie: il faut boire du vin rouge le soir, et du vin blanc à déjeuner.

Oudin introduirait donc un sens 'personnel', par plaisanterie, repris par la suite par certains lexicographes. Cette hypothèse est mise à mal par la seule attestation avec le sens relatif au vin trouvée dans l'œuvre de Charles Sorel, *le Berger Extravagant*, paru en 1627, quelques années avant les *Curiositez.*

– Aportez moy du clairet, belle Deesse potagere, ou nous ne serons pas bons amys. Et bien encore pour ce coup cy à t'il raison, dit Adrian, rouge au soir blanc matin, c'est la journée du pelerin. L'on entend cela pour le temps, mais je l'entens pour le vin, moy. (Sorel 1627: 86)

Dans le cadre de cette scène littéraire, où l'on parle de vin et plus précisément de la couleur de celui-ci, l'on comprend le jeu que l'auteur opère sur le proverbe en en détournant le sens. Ceci nous invite à faire plutôt l'hypothèse qu'il existe un lien étroit entre les *Curiositez françoises* et la littérature burlesque de l'époque: le sens particulier relatif au vin aurait été introduit, par plaisanterie, d'abord dans la littérature et recueilli ensuite par Oudin.

Cas 7

- Bains de Valentin, *voyez le sujet de cecy dans Francion, c'estoit un vieillard qui s'alla baigner de nuit dans le fossé d'un Chasteau pour se rendre habille à coucher avec sa femme, qui fut pendant cela desbauchée par un autre.*

Dans cet article, le lexicographe ne donne pas la définition à la séquence nominale en entrée. À la place, il opère un renvoi à une scène de l'œuvre de Charles Sorel parue en 1623. Le lien avec la littérature, contrairement au cas 6, est ici explicite. Nous avons retrouvé la séquence dans le *Francion*:

– Le Curé voulut sçavoir de luy par quel moyen il avoit esté mis là. Il fut contrainct de raconter les enchantements que luy avoit appris Francion, et de dire aussi pour quel subjet il les avoit voulu entreprendre. Quelques mauvais garçons en ayant entendu

[19] Fur 1690, s. v. *pelerin*: «On dit proverbialement, Rouge au soir, blanc au matin, c'est la journée du pelerin. Le proverbe s'explique en deux façons: l'une qu'il faut boire du vin rouge au soir, & le matin du vin blanc à desjeûner: l'autre, que ces deux couleurs de l'air monstrent qu'il doit faire beau temps durant le jour.»

> l'histoire s'en allerent la publier par tout à son infamie, si bien qu'encore aujourd'huy l'on s'en souvient, et lorsqu'il y a quelqu'un qui a froide queuë, l'on luy dit par mocquerie qu'il s'en aille aux bains de Valentin. (Sorel 1623: 74)

L'extrait met en place l'expression *bains de Valentin*, ou peut-être *s'en aller aux bains de Valentin*, en la présentant comme appartenant à la langue grâce à l'emploi d'un *on* générique («on luy dit…») ainsi qu'en inscrivant une distance temporelle et un décrochage actanciel entre la scène littéraire y faisant référence et l'emploi 'actuel' de la séquence s'appliquant à quiconque: «si bien qu'encore aujourd'huy l'on s'en souvient et lorsqu'il y a quelqu'un...». Excepté cette attestation, nous n'en avons trouvée aucune autre, ni antérieure ni postérieure. Les indices internes et externes, tout comme pour le cas précédent, nous orientent vers une allusion littéraire.

Cas 8

- Sept P sous un P. i. *poüils, pulces, punaises, pauvreté, patience, petite portion que les escoliers endurent sous un Pedant.* vulg.

Tout comme le cas 7, le lexicographe ne donne aucune définition de la forme en entrée, simplement explique-t-il le jeu de mot: les sept *P* sont les sept mots commençant par *p* (*poüils*, *pulces*, *punaises*, *pauvreté*, *patience*, *petite portion*), qui traduisent les maux qu'endurent les écoliers sous un pédant, terme désignant au 17^{e} siècle un maître d'école. Cette forme n'est attestée dans aucune source secondaire ni tertiaire. Nous avons néanmoins trouvé, dans le *Francion* (1623), non la séquence à proprement parler mais un extrait présentant un jeu de mots fort semblable:

> – J'apris alors à mon grand regret, que toutes les paroles qui expriment les malheurs qui arrivent aux escoliers, se commencent par un P, avec une fatalité très remarquable: car il y a Pedant, peine, peur, punition, prison, pauvreté, petite portion, poux, puce et punaise. (Sorel 1623: 169)

Le *Francion* mentionne onze mots commençant par P, les sept derniers sont identiques à ceux repris par Oudin. Malgré l'absence d'attestation précise, nos analyses nous invitent à voir, pour ce cas-ci également, un lien étroit entre la séquence recueillie par Oudin et l'œuvre de Charles Sorel.

Grâce aux données internes, à leur analyse et à la confrontation de celles-ci avec les données externes, certes peu nombreuses, nous avons pu, pour l'étude particulière réalisée ici, mettre en évidence un lien étroit entre les *Curiositez* et la littérature burlesque de l'époque, plus précisément avec les œuvres de Charles Sorel. Nous faisons l'hypothèse que les séquences peu voire pas attestées, puisque le lexicographe les a recueillies, ont eu une certaine vitalité dans le langage oral auquel nous n'avons pas accès. Si l'on devait parler de ces séquences en termes de degré[20] de figement, nous dirions qu'elles en ont un moindre car, d'une part, elles ne s'inscrivent pas durablement dans la langue et, d'autre part, elles gardent un lien étroit avec la littérature qui leur a donné vie. Elles se situent entre l'allusion citative et la séquence figée.

4 Conclusion générale

Notre étude sur le figement dans les *Curiositez françoises* (1640) d'Antoine Oudin, et plus précisément sur la question de l'identification des SF articule analyse de cas concrets, particuliers et méthodologie plus générale.

Sur le plan méthodologique, ce présent article a montré que toute étude lexicologique (ou si l'on préfère phraséologique) menée à partir d'un dictionnaire doit reposer sur une description métalexicographique rigoureuse. Chaque champ informationnel d'un article doit être identifié et analysé en profondeur. C'est grâce à cela que nous avons pu mettre au jour des indices internes, implicites ou explicites, récurrents se situant à différents niveaux d'analyse: soit relevant de la forme et du sens de la séquence en tant que SF, soit relevant d'une pratique lexicographique particulière. Ces indices internes doivent nécessairement être mis en relation avec des données externes, secondaires et tertiaires, afin de déterminer l'inscription de la séquence dans la langue. Ces données viennent également compléter les données internes en précisant la forme de la séquence (la portée et le cadre variationnel) ou encore son sens. La variation s'observe tant sur le plan formel (*polymorphisme*) que sur le plan sémantique (*polysémie*). Ce sont les attestations de la séquence dans la langue qui assurent son identification en tant que SF. Cependant, l'absence d'attestation ne signifie pas que la séquence n'est pas figée. En effet, comme nous l'avons montré au début de cet article, d'une part, les sources secondaires et tertiaires consultées sont factuellement en nombre limité et, d'autre part, les sources primaires, à savoir les échanges oraux, que nous considérons comme le vecteur de transmission privilégié des SF, nous font défaut. Dans les cas où nous avons peu voire pas d'attestations, nous avons montré l'importance des indices internes

[20] De nombreuses études actuelles font remarquer que le figement ne se manifeste pas de façon binaire, mais qu'il s'agit plutôt d'une question de degré. Le figement, comme nous l'avons dit, est un processus qui s'inscrit dans le temps et qui permet l'intégration dans la langue de séquences qui étaient à l'origine libres dans le discours. Il y a donc continuum entre séquences libres (SL) et séquences figées (SF). Nous comprenons dès lors le degré de figement comme la gradation qu'occupe une séquence sur cette échelle.

implicites (l'absence de définition, l'opposition entre un *je* particulier et un *on* générique) et explicites (le renvoi direct à une œuvre littéraire).

Sur le plan des *Curiositez*, nos recherches nous invitent à souligner la qualité du travail d'Antoine Oudin. Interprète-traducteur du roi, il connaît la langue et se révèle un descripteur fiable de celle-ci. La plupart des séquences répertoriées par le lexicographe sont attestées et donc bien inscrites dans la langue de l'époque. Les *Curiositez* constituent dès lors un véritable trésor pour la connaissance des SF de la première moitié du 17^{e} siècle. Prenant appui sur la qualité du travail, l'étude de cas peu voire pas attestés s'est révélée intéressante. En effet, à travers l'étude de trois cas particuliers, nous avons montré l'existence d'un lien étroit entre les *Curiositez* et la littérature burlesque de l'époque. Ce constat nous a conduit à mener une analyse lexicologique plus fine sur le degré du figement.

Le problème de l'identification des séquences figées d'un état de langue ancien auquel nous avons tenté d'apporter quelques solutions mérite encore qu'on s'y attarde largement. C'est grâce à une méthodologie rigoureuse que nous pourrons continuer à mettre au jour la richesse phraséologique de ce véritable trésor que constituent les *Curiositez*.

Références bibliographiques

Textes littéraires

ANONYME, 1624: *Les ramonneurs*. Éd. Austin Gill 1957. Paris: Marcel Didier.

ANONYME, 1633: *La comédie des proverbes*. Éd. Michael Kramer 2003. Genève: Droz.

BEROALDE DE VERVILLE, François, 1610: *Le moyen de parvenir*. Éd. Bernard de la Monnoye 1879. Paris: Garnier Frères.

DE FERRIÈRE, Claude, 1692: *Corps et compilation de tous les commentateurs anciens et modernes sur la coutume de Paris*. Paris: Denys Thierry.

SCARRON, Paul, 1645: *Le Jodelet ou le Maistre valet*. Éd. William Dickson 1986. Exeter: University of Exeter.

SCARRON, Paul, 1650: *Le Virgile travesti*. Éd. Victor Fournel 1858. Paris: Adolphe Delahays.

SOREL, Charles, 1623: *Histoire comique de Francion*. Éd. Yves Giraud 1979. Paris: Garnier-Flammarion.

SOREL, Charles, 1627: *Le berger extravagant*. Paris: Toussainct de Bray.

Base de données

Frantext: < http://www.frantext.fr/>. Août 2012.

Google books: < http://books.google.fr/>. Août 2012.

Dictionnaires

ACADÉMIE FRANÇAISE, 1694: *Le Dictionnaire de l'Académie françoise*. Paris: Coignard.

COTGRAVE, Randle, 1611: *A Dictionarie of the French and English Tongues*. London: Adam Islip.

FURETIÈRE, Antoine, 1690: *Dictionnaire universel, contenant généralement tous les mots françois, tant vieux que modernes, et les termes de toutes les sciences et des arts, sçavoir la philosophie /…/*. La Haye et Rotterdam: A. et R. Leers.

LITTRÉ, Émile, 1873: *Dictionnaire de la langue française*. Supplément 1877. Paris: Hachette.

NICOT, Jean, 1606: *Thresor de la langue francoyse, tant ancienne que moderne*. Paris: David Douceur.

OUDIN, Antoine, 1640: *Curiositez françoises, pour supplément aux dictionnaires ou recueil de plusieurs belles propriétez, avec une infinité de proverbes et quolibets, pour l'explication de toutes sortes de livres*. Paris: Antoine de Sommaville [id. Genève: Slatkine reprints 1993].

RICHELET, Pierre, 1680: *Dictionnaire françois, contenant les mots et les choses, plusieurs nouvelles remarques sur la Langue française*. Genève: Widerhold.

WARTBURG, Walther von, 1928–: *Französisches etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes* (FEW). Bonn: Klopp; Basel: Zbinden.

Littérature

ANSCOMBRE, Jean-Claude, 2000: Parole proverbiale et structures métriques. *Langages* 139, 6–26.

BENVENISTE, Emile, 1966: *Problèmes de linguistique générale.* Paris: Gallimard.

BIASON, Maria Teresa, 1992: Les formes brèves dans le contexte culturel du XVII^e^ siècle en France. *Rivista di letterature moderne e comparate* 45, 263–286.

BURIDANT, Claude et SUARD, François, 1984: *Richesse du proverbe. Le proverbe au Moyen Age.* Lille: Université de Lille III.

CONENNA, Mirella, 1988: Sur un lexique-grammaire comparé de proverbes. *Langages* 90, 99–116.

GROSS, Gaston, 1996: *Les expressions figées en français*. Paris: Orphis.

KRAMER, Michael, 2000: Les armes de Caïn: une expression sous enquête diachronique. *Neophilologus* 84/2, 165–187.

KRAMER, Michael, 2002: Sources littéraires des 'Curiositez françoises'. *Revue de linguistique romane* 261–262, 131–157.

MEJRI, Salah, 1997: *Le figement lexical*. Tunis: Publications de la Faculté des Lettres de Manouba.

MEJRI, Salah, 2001: La structuration sémantique des énoncés proverbiaux. *L'information grammaticale* 88, 10–15.

MEJRI, Salah, 2003: Le figement lexical. *Cahiers de lexicologie* 82, 23–39.

MONTORO DEL ARCO, Esteban Tomás, 2011: Locutions à cases vides, locutions à cases libres et phénomènes apparentés. Anscombre, Jean-Claude et Mejri, Salah (éd.): *Le figement linguistique: la parole entravée.* Paris: Honoré Champion. 249–265.

REY, Alain, 1973: La phraséologie et son image dans les dictionnaires de l'âge classique.

Mélanges à Paul Imbs. Travaux de linguistique et de littérature 11/1, 97–107.

ZULUAGA, Alberto, 1980: *Introducción al estudio de las expresiones fijas*. Frankfurt am Main: Peter Lang.

ZUMTHOR, Paul, 1976: L'épiphonème proverbial. *Revue des sciences humaines* 163, 313–328.

General Extenders: From Interaction to Model

Peter Grzybek (Graz)
Darinka Verdonik (Maribor)

Abstract

Based on the assumption that general extenders represent a separate category in the linguistic and phraseological system of a given language, the present study attempts to show that their frequency occurrence is regular and law-like, as the result of a diversification process. Based on an empirical analysis of the Slovene GOS corpus, a specific discrete probability distribution and its continuous counterpart, the Zipf-Alekseev model, is presented as an adequate and interpretable model.

1 Introduction

In this contribution, we will represent our research and some new ideas with regard to a group of pragmatic markers in spoken interaction. These linguistic units have recently been summarized under the term *general extenders* (GEs); other terms for these set of expressions are, among others: *set marking tags*, *utterance final tags*, *clause terminal tags*, *extension particles*, *generalized list completers*, *generalizers*, *final coordination tags.* In English, this group of expressions includes examples such as *and stuff*, *and everything*, *or something*, *or anything*, and others; in German, we have corresponding expressions such as *und so*, *und so weiter*, *und so weiter und so fort*, *und solche Sachen*, *und alles Mögliche*, *oder so*, *oder so was*, and many others. In speech, such expressions are used in verbal contexts such as:

(1) *Ich hab' nun jetzt erstmal meine ganzen Pflanzen da in die Erde gebracht und – und – sehr viel Tulpen und Krokusse und so was gesteckt.*

(2) *I've just got all my plants there in the ground and – and – lots of tulips and crocuses and so was put in.*

(Overstreet 2005: 1849)

Thus far, GEs have not been dealt with systematically in the field of phraseology, and they have been dealt with much more broadly in various branches of linguistics (see below). Yet, from a phraseological point of view, GEs can be considered to be a special kind of pragmatic phrases[1] (Burger et al. 1982: 110ff., Fleischer 1982: 133f., Jakop 2005) since they are multi-word expressions functioning as one unit being stored and retrieved from memory at the time of use, on the one hand, and because they have no "ideational" meaning on their own, thus being of a pragmatic rather than semantic kind, not contributing to the "propositional content" of the utterance, on the other. As a consequence, the term *pragmatic phrase* seems to be preferable as compared to *pragmatic idiom*, an argument fully in line with Fleischer's (1982: 132) reasoning that such units lack (full) idiomaticity, and also in accordance with more recent distinctions between phrasemes in a narrow sense (including idiomaticity) and phrasemes in a broader sense (Burger 1998: 32). In fact, the attribution of GEs to pragmatic phrasems can, from a phraseological perspective, be traced back, among others, to Burger et al. (1982), where the authors list examples such as *und so weiter* (ibd., 126) as speech-specific pragmatic phrases. Likewise, Prodromou (2008: 117) makes a comparable categorization for two-word pragmatic markers such as *you know* and *I mean*, which he calls *pragmatic phrases*, in order to capture the formulaic nature of these discourse markers.

Given the relatively sparse treatment of GEs in the field of phraseology, this contribution will take the following course: Subsequent to a discussion of general extenders from a linguistic point of view (2.1), we will focus on hitherto almost ignored material from Slovene (2.2.) and argue in favor of the notion that GEs represent a specific linguistic and phraseological category in its own right (2.3.) which, as a consequence, follows well-known regularities as to their frequency behavior; in this respect, we will analyze corpus material from the Slovene GOS corpus, attempting to show that the frequencies of GEs are not arbitrarily or chaotically organized, but follow a well-defined frequency distribution model, as a result of what we consider to be a diversification process.

2 General extenders

2.1 General extenders in linguistics

In the field of linguistics, Dines (1980) has been one of the first to pay special attention to GEs, using a variationist framework to study their use in Australian English; in her terminology, expressions of this kind are called *clause terminal tags* and *set-marking tags*. Later, Dubois (1992) used term *extension particles*: she analyzed the socio-demographic factors of their use, carried out

[1] German: *pragmatische Phraseologismen*, Slovene: *pragmatični frazemi*.

a distributional analysis of certain expressions which are component parts of extension particles, and conducted an analysis of the conditioning of distinct categories of extension particles by linguistic and social factors. Finally, the term *general extender* was suggested by Overstreet (1999); it is the most-widely term used still today (cf., e. g.: Cheshire 2007; Fernandez/Yuldashev 2011; Martinez 2011; Tagliamonte/Denis 2010).

General extenders (GEs) have been discussed in different theoretical frameworks, and from different perspectives, and it seems worth to initially mention some major studies, in order to demonstrate the broad spectrum of approaches within the field of linguistics:

(a) Overstreet has discussed them within the functional grammar framework of ideational and interpersonal function; for the analytic approach she follows discourse analysis, emphasizing the socio-cultural perspective, and claiming that "within their actual context of occurrence, GEs appear to have function that is primarily interpersonal and tied to the nature of the social relationship of the participants" (1999: 145).

(b) Subsequent to Overstreet's work, Aijmer (2002) has treated GEs as particles with vague reference; her study is empirical, using corpus methods, in combination with a mixed approach in its analytic parts, including a description with regard to form and function (claiming they have textual and interpersonal functions), prosody and situations in which they are used.

(c) Yucker et al. (2003) have included GEs in their study as one of the elements of vagueness in conversation.

(d) Overstreet (2005) has compared the use of GEs in English and German conversations between adults.

(e) Cheshire (2007) has analyzed GEs in the speech of adolescents from three English towns, comparing the use of different forms, grammaticalization processes and the multifunctionality of GEs.

(f) Cucchi (2007) has compared the use of the GE *and so on* by native and non-native speakers of English, based on the corpus of EU parliamentary debates.

(g) Tagliamonte and Denis (2010) have examined GEs in the English spoken in Toronto, using quantitative techniques and investigating their phonetic reduction, de-categorization, semantic change and pragmatic shift.

(h) Palacios Martinez (2011) has analyzed the use of GEs in the speech of British teenagers and adults, comparing the frequency of use, grammaticalization and pragmatic functions.

(i) Terraschke (2010) has explored the use of the English GE *or so* by native speakers in New Zealand and by German non-native speakers of English, comparing frequency and functions of use.

(j) Pichler and Levey (2010) have presented a quantitative analysis of the co-occurrence of GEs (e. g., *and stuff*, *or something like that*) with other discourse features (e. g., *like*, *you know*) in a corpus collected in North-East England, concentrating on three age groups.

(k) Fernandez and Yuldashev (2011) have analyzed the use of GEs in instant messages, comparing their use (forms, frequency, set of reference, functions) by native and non-native English speakers.

As can be seen from the list above, most research on GEs has been done in English language; yet, this category is of course present in many other (if not most) languages. However, for languages other than English, GEs have been dealt with to a lesser degree, and their representation and discussion is far more limited for languages other than English; nevertheless, we find discussions concerning Japanese (Wiezbicka 1991, Honda, 1996), Montreal French (Dubois 1992), Swedish (Winter/Norby 2000), German (Overstreet 2005), Spanish (Cortés/Rodríguez 2006), Persian (Parvaresh et al. 2010), or Lithuanian (Ruzaitė 2010), to mention but a few, and there seems reason to assume that we are concerned with a broadly (if not or even universally) distributed category.

As to the status of this category in the field of linguistics and phraseology, it seems worthwhile noting that, although GEs, on the one hand, seem to represent a specific class of linguistic elements in its own right, with their own specific communicative and pragmatic functions in discourse, and that this class is represented by different kinds of items, or even subcategories, on the other.

Concentrating on GEs as a specific subcategory of the linguistic or phraseological system, we will not deal with details of such further distinctions or sub-categorizations here; let it therefore suffice to mention merely the most important and most general sub-categorization of GEs, with regard to their structure and function, i. e., the distinction of two major groups, which are termed (a) adjunctive and (b) disjunctive: those GEs beginning with *and* in English, or with *und* in German, fall into one category which is called adjunctive, and those beginning with *or* (English) or *oder* (German), belong to the second class of expressions, called disjunctive.[2] Both groups differ not only in structure, but also in (textual) function: according to Aijmer (2002), adjunctive GEs have expanding and illustrative function, whereas disjunctive GEs have

[2] In Slovene – which will be discussed in more detail further below in Section 3 – we have analogical expressions, such as, for example, *in* or *pa*, on the one hand, and *ali*.

an approximation function; similar differences are reported by Overstreet (2005: 1855), for example, who summarizes that "the primary function of adjunctive general extenders is to indicate 'there is more'", while the primary function of disjunctive GEs is "tied to indicating potential alternatives, and hence hedging on what has been said". Besides the textual functions mentioned above, there is common acknowledgment that all GEs (adjunctive as well as disjunctive) perfom interpersonal functions as well:[3] based on the assumption that GEs are used to indicate that there is, in fact, more to be said than is said explicitly, GEs are seen to express the assumption of shared knowledge and experience between speaker and listener(s), to represent an appeal for solidarity and understanding, to indicate a lack of certainty, etc.

As has been said above, in this contribution we do not intend to go into further details of describing the GEs discourse functions: accepting that differences in usage and subdivisions exist, the category of GEs thus consisting of heterogeneous elements, we focus instead on GEs as a distinct class of linguistic elements as a whole, the members of which we consider to be the result of some diversification process (see below).[4] In order to attempt to provide plausible arguments in favor of this assumption, we will base our analyses on Slovenian material, which shall be presented in the next subsection.

3 General extenders in Slovene: Data

As has been mentioned above, GEs have received only sporadic attention in Slovene (Verdonik/Kačič 2012). As a consequence, it seems reasonable, in a first step, to provide some relevant information about this kind of expressions

[3] Overstreet (1999) even claims that such interpersonal functions are primary to general extenders; as compared to this, Cheshire (2007) postulates one should consider their functions within the local context in which they are used, that they are multifunctional, and that we cannot define a principal function – even more, it would be counterproductive to do so.

[4] The idea to see GEs as a distinct class of linguistic and/or phraseological elements is of course not new; Overstreet (1999: 6), for example, starts her monograph on GEs with the words: "They represent a distinct set of linguistic elements /…/". However, in the further course of her book, she mentions the possibility that "it might be possible to describe general extenders as types of 'discourse markers'" (ibd.: 12). In fact, GEs and discourse markers have some common characteristics, but there are important differences, too; as a consequence, it has been rather a matter of perspective to either group GEs into one common class with discourse markers, or to see both as members of some larger class of forms, e. g., 'pragmatic operators'. For details on this discussion see, among others: Dubois (1992), Aijmer (2002), or Martinez (2011).

which, in Slovene, is represented by items such as, e.g., *in tako naprej* 'and so on', *pa to* 'and so', *in podobno* 'and similar', *pa tako* 'and so', *ali pa kaj takega* 'or something like that' – they all fall into the GE category, cf. the following examples:

(1) *se boste o otrocih pogovarjale* ***pa tko***
'you will talk about kids **and so**'

(2) *mama je umrla in mislim z bratom sva drgač zmenena* ***in tko naprej*** *a ne*
'my mother died and I mean me and my brother have different arrangements ***and so on*** y'know'

(3) *recimo ko gremo v savno* ***al pa kaj takšnega***
'for example when we go to sauna ***or something like that'***

The data used for our analysis are taken from the GOS (*GOvorjena Slovenščina*) corpus, the reference corpus of Slovene speech (Verdonik/Zwitter Vitez 2011; available also at <www.korpus-gos.net>). This corpus consists of 1,032,775 words, or 120 hours of recordings. The GOS corpus contains speech events from five different discourse types with different channels, as shown in Table 1. An important characteristic of the corpus, which also ensures a good comparability of different discourse types in the corpus, is that the majority of the recordings include spontaneous speech (as opposed to read speech).

Discourse type	Channel	Number of tokens	Totals	Percentage
Classes		162,750	162,750	15.76 %
Media – informative	Radio	94,536	196,799	19.06 %
	TV	102,263		
Media – entertainment	Radio	123,152	228,765	22.15 %
	TV	105,613		
Official	Phone	33,484	153,471	14.86 %
	Personal communication	119,987		
Private	Phone	68,083	290,990	28.18 %
	Personal communication	222,907		

Table 1: Discourse types in the GOS corpus.

The data in the GOS corpus are available in two transcription formats: a pronunciation-based transcription and a standardized transcription. Pronunciation-based transcription is an orthographic transcription which represents a more or less faithful account of acoustic forms of words; the standardized transcription follows the Slovene written standard and offers a common form for different pronunciation realizations of the same lexeme. The web-interface enables also listening the audio recording for each concordance.

In the GOS corpus, we found more than 50 different Slovene expressions functioning as general extenders; at closer sight it becomes evident, however, that many of them are merely different variations of one basic expression. On the whole, we identified four basic variants:

(1) Pronunciation variation: Since the GOS corpus includes pronunciation-based transcriptions, many vocal reductions and other phoneme-based variations were evident, for example *tako* ‘so’ (occurring in the general extenders *in tako naprej* ‘and so on’, *in tako* ‘and so’, *ali pa tako* ‘or so’, etc.) has variations like: *tko*, *tk*, *tak*, *teku*, etc.

(2) Grammatical variation: Slovene is a highly inflectional language. A common variation in general extenders includes the opposition of nominative form (e. g., *in podobno* ‘and similar’, *in vse te stvari* ‘and all that things’) vs. genitive form (e. g., *in podobnega* ‘and similar’, *in vseh teh stvari* ‘and all that things’).

(3) Synonym variation: Slovene general extenders vary considerably by including different synonyms; common variations are: synonym conjunctions *in* ‘and’, which is standard (*in to* ‘and that’), vs. *pa* ‘and’, which is colloquial (*pa to* ‘and that’); synonym or semantically close pronouns *ta* ‘this’ (*in te stvari* ‘and these things’) vs. *tak* ‘such’ (*in take stvari* ‘and such things’) vs. *takšen* ‘such’ (*in takšne stvari* ‘and such things’); synonym nouns *stvar* ‘thing’ (*in te stvari* ‘and these things’) vs. *zadeva* ‘thing’ (*in take zadeve* ‘and such things’) vs. *reč* ‘thing’ (*pa take reči* ‘and such things’); synonym adverbs *naprej* ‘on’ (*in tako naprej* ‘and so on’) vs. *dalje* ‘forth’ (*in tako dalje* ‘and so forth’).

(4) Optional item addition: Most disjunctive general extenders in Slovene have the optional particle *pa* after the beginning conjunction *ali* ‘or’, e. g., *ali karkoli* ‘or whatever’ vs. *ali pa karkoli* ‘or whatever’ (there is no equivalent in English). Some general extenders also vary in length, i. e., they have a basic form (e. g., *pa vse* ‘and all’) which can be prolonged by one item (e. g., *pa vse to* ‘and all that’) or even more items (e. g., *pa vse to skupaj* ‘and all that together’); other examples would be *pa vse* ‘and all’ vs. *pa vse skupaj* ‘and all together’ vs. *pa vse to skupaj* ‘and all that together’; *ali karkoli* ‘or whatever’ vs. *ali karkoli drugega* ‘or whatever else’; *ali pa kaj* ‘or something’ vs. *ali pa kaj takega* ‘or something like that’.

We considered that keeping all different expressions – which sum up to more than 50 – would blur the common picture of general extenders frequencies, as there would be a substantially bigger gap among some expressions (e. g.. between *in tako naprej* 'and so on' vs. *pa to* 'and that') than among others (e. g., *in tako naprej* 'and so on' vs. *pa tako naprej* 'and so on'). Therefore we decided to group expressions into basic groups of general extenders, disregarding the variation described above, as a result obtaining 14 groups of general extenders. These are represented in the Table 2, along with their frequencies[5] of usage in the GOS corpus.

GE group	English translation	Variations	Frequency
in tako naprej	and so on	in/pa tako naprej/dalje	286
in podobno	and similar	in podobne zadeve/reči/dalje	27
pa to	and that	pa/in to/tega (vse)	342
in te stvari	and such things	in/pa te/teh/take/takšne stvari/zadeve/reči	35
in vse stvari	and all things	in/pa vse/vseh (te) stvari/reči	4
pa vse (to skupaj)	and all (that together)	pa/in vse/vsega (to/tega) (skupaj)	96
pa tako	and so	pa/in tako	155
ali pa kaj (takega)	or something like that	ali (pa) kaj (takega/takšnega) (v tem smislu)	94
ali pa kaj jaz vem	or I don't know	–	4
ali nekaj takega	or something like that	ali (pa) nekaj takega/takšnega	39
ali kaj podobnega	or similar	ali (pa) kaj/česa podobnega	19
ali karkoli (že/takega)	or whatever (similar)	ali (pa) karkoli/česarkoli (že/ drugega/takega/pač)	32
ali pa to	or that	–	5
ali pa tako	or so	–	10
TOTAL			**1148**

Table 2: Slovene general extenders in the GOS corpus.

[5] It should be noted that most of the phrases had to be disambiguated because they can function either as a general extender (e. g., *vrtna opravila* <u>*pa to*</u> 'garden work <u>and so</u>') or not (e. g., ordinary conjunction + pronoun: *potem bo* <u>*pa to*</u> *pomenilo* 'than <u>this</u> will mean').

Yet, in this contribution, we intend to go one step further, not restricting our objective to the mere description of Slovene GEs, their functions, forms and usage. Rather, we will extend our interest to theoretical issues, regarding GEs to represent a specific linguistic sub-system in its own right, characterized by a common (pragmatic) function. In this context, locating them on the phraseological level of language, we assume GEs to underlie a process of diversification; as a result, the frequency with which each member of this class (i. e., each individual GE) occurs, should not be arbitrary of chaotically organized, but rather in a law-like manner, each individual GE as well as its frequency of occurrence thus being the result of a specific diversification process. If this is true, we will thus, with our contribution, not only enrich hitherto research by adding new Slovene material, but substantially expend their theoretical treatment from a methodological point of view, seeing GEs from a synergetic perspective.

4 Modeling general extenders in terms of a diversification process

4.1 Diversification

Diversification is a well-known process characterizing all living and dynamic systems: in biology, for example, there would be no variation in organic nature without diversification. In linguistics, diversification processes are well-known, too, mainly in the field of grammar and semantics. Here they are assumed to take place when the attribute space of an entity expands in one or more dimensions, e. g., when a morpheme gets enriched by new allomorphs, when a word gets enriched by new meanings, etc. Since over-diversification in language would not be economic for the communicative system as a whole, satisfying equally well producers' and receivers' interests, diversification processes must necessarily be counter-balanced by processes of unification, the whole system thus turning out to be in a state of dynamic equilibrium.

4.2 Diversification in linguistics

In trying to model the observed frequencies, one of our basic assumptions is, as has been said above, that general extenders "represent a distinct class of linguistic elements" (Overstreet 1999: 3). Along with this assumption goes the hypothesis, well-known from the field of quantitative linguistics, that frequencies with which different forms of a linguistic category occur, are regularly distributed. In other words: not only the lexical inventory (of a given text or corpus), but also specific sub-categories are expected to follow regularities as to their frequency behavior. This can be seen in context of, or as a result of, diversification processes: just like in biology (or bionics) the rise of new species is a result of diversification, or as the introduction of new products into the market can be seen as diversification process in economics, diversification is a crucial process in linguistics (Altmann 1991, 2005). Here, diversification can be seen as a process of enlarging the number of forms or meanings of a given linguistic entity. In this sense, diversification processes have repeatedly been described with regard to various linguistic levels:

(a) *Paradigmatic*: e.g. the rise of cases, numbers, tenses, etc.,
(b) *Phono-morphemic*: e.g. the rise of allophones, allomorphs etc.,
(c) *Geographical*: e.g. the increase in the number of different expressions of a concept,
(d) *Social*: e.g. the rise of different words or meanings of a word or different pronunciations,
(e) *Idiolectal*: within a community,
(f) *Semantic*: e.g. the increase in synonymy and polysemy,
(g) *Contextual*: e.g. the increase in the usage of a unit in different contexts.

Given these observations, and based on the assumption that every linguistic entity diversifies, that is, it generates variants and secondary forms, we set up the hypothesis that the frequency distribution of general extenders as a specific subcategory of language is the result of a diversification process, and that, as a result of this, general extenders are regularly distributed.

4.3 Linguistic law of diversification

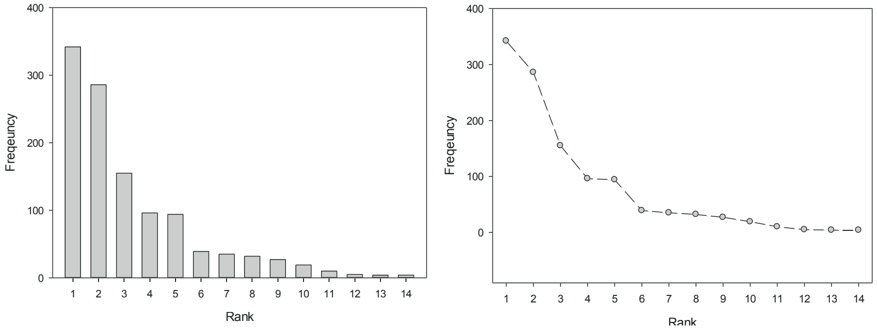
Generally speaking, a frequency distribution is based on the individual class members' frequency of occurrence. With regard to diversification processes, we are concerned with the more specific hypothesis that the diversifying linguistic entity under study abides by a ranking law, resulting in a specific rank-frequency distribution: if the members of the diversified entity are ordered according to their frequency, then the frequencies are "lawfully" connected. It goes without saying that rank-frequency distributions as functions expressing the decrease of frequencies ranked according to their magnitude, there are, eo ipso, no bell-shaped frequency distributions, which are rather of a left-skewed form.

Table 3 represents the data presented above (cf. Table 2), transformed in a rank-frequency distribution: in the first column, the rank (*x*) is given, in the second column, the frequency *f(x)* of the corresponding entities.

x	*f(x)*
1	342
2	286
3	155
4	96
5	94
6	39
7	35
8	32
9	27
10	19
11	10
12	5
13	4
14	4

Table 3: Rank frequencies of Slovene general extenders in the GOS corpus.

The rank-frequency distribution is graphically illustrated in Figures 1a and 1b: whereas Figure 1a presents them in the form of a bar chart, usual for the representation of discrete frequency distributions, Figure 1b presents them in the form of a line plot, usually preferred for the representation of continuous data and functions.



1a: Discrete bar chart

1b: Continuous line plot

Figure 1: Frequencies of general extenders in the Slovene GOS Corpus.

Given the observed frequencies, the next step will include their theoretical modeling.

4.3 Modeling the frequencies of general extenders in a diversification framework

An attempt to model the frequencies illustrated in Figures 1a and 1b implies the search for an adequate mathematical function, i. e., a discrete probability distribution and/or a continuous function. As to a theoretical derivation of such a model, it seems first important, albeit trivial, to take account of the fact that, if the above hypothesis holds, the frequencies of elements of the given linguistic class are not distributed uniformly. Because the entities are ranked, and because of the corollary, it is true that for the probabilities of classes it holds that $P_{\mathrm{x}} \leq P_{\mathrm{x-1}}$ – here, P_x is the frequency of the elements of a given class, and P_{x-1} is the frequency of the elements of its preceding class. Moreover, since P_{x} and $P_{\mathrm{x-1}}$ ($x = 2,3,\ldots$) are related in a law-like manner, we can understand P_{x} to be a function of $P_{\mathrm{x-1}}$. In mathematical terms, we can thus write

$$P_x = g(x) P_{x-1} \qquad (1)$$

with *g(x)* ≤ 1, since we have a monotonously decreasing rank order. Furthermore, attempting to find such a model, it seems reasonable to take into consideration the fact that every diversification process evokes a unification process operating on the same entity and working against the total decay of the phenomenon. This idea goes back to George K. Zipf's (1935, 1949) ideas about the antagonistic economy of producer and recipient in communication processes, and they are well-known today by the name of Zipfian forces in quantitative linguistics. As a result, assuming *g(x)* to represent the Zipfian forces of unification and diversification, *g(x)* can be set up as

$$g(x) = \frac{f(x)}{h(x)}, \qquad (2)$$

where *f(x)* represents the diversification process, and *h(x)* the unification component, i. e., the controlling, regulating effect of the communication community. Moreover, *f(x)* can be understood to be a function composed of a language constant (e. g., *a*), on the one hand, plus the diversifying force of the producer (e. g., *b·x*), on the other. With *h(x)* = *c·x*, we would thus obtain

$$g(x) = \frac{a + bx}{cx}, \qquad (3)$$

so that, according to equation (1)

$$P_x = \frac{a+bx}{cx} P_{x-1}, \qquad (4)$$

which would, after re-parametrization, result in the well-known negative binomial distribution, in its zero-truncated (positive) form. This model has repeatedly been applied to model diversification processes in linguistics; there is no need to go into further details here, the more since in our case, i.e., with regard to general extenders, it seems more adequate to set up *f(x)* differently, with *f(x)* = *a* + *b* · ln(*x*). This results in the Zipf-Alekseev distribution

$$P_x = C \cdot x^{-(a+b\ln x)} P_{x-1} \qquad x = 1, 2, 3, \ldots \qquad (5)$$

with $C^{-1} = \sum_{j=1}^{\infty} j^{-(a+b\ln x)}$ as the normalizing constant.

Analogically, one obtains the continuous Zipf-Alekseev function

$$y = f(x) = C \cdot x^{-(a+b\ln x)}, \qquad (6)$$

for which the requirement that the sum of all probabilities Σ *P* = 1 needs not be fulfilled, as is the case in the discrete model (5).

This Zipf-Alekseev model, too, has been proven to be adequate in the modeling of linguistic diversification (cf. Altmann 2005). In this context, it has also been theoretically derived, with reference to either the psychophysical Weber-Fechner law, based on the perceptibility of minimal differences between classes (cf. Hammerl 1991), or to the well-known Menzerath-Altmann law from linguistics, concerning the construct-constituent relation of linguistic units (cf. Hřebíček 1996, Altmann/Hřebíček 1996).

Applying this model to our data, in its discrete form, it seems reasonable to use the right-truncated version

$$P_x = C \cdot x^{-(a+b\ln x)} P_{x-1} \qquad x = 1, 2, 3, \ldots, n \qquad (7)$$

with $C^{-1} = \sum_{j=1}^{n} j^{-(a+b\ln x)}$ as the normalizing constant.

With parameter values $a = -0.03$ and $b = 0.61$, the fit turns out to be good[6], as indicated by the discrepancy coefficient $C = 0.019$. Figure 2 illustrates the result in graphical form.

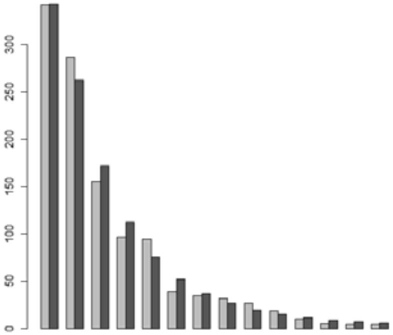


Figure 2: Observed and theoretical frequencies of general extenders in the Slovene GOS Corpus – right-truncated Zipf-Alekseev distribution.

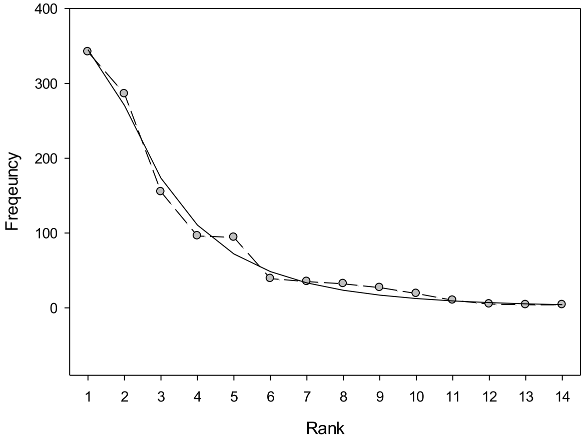


Figure 3: Observed and theoretical frequencies of general extenders in the Slovene GOS Corpus – continuous Zipf-Alekseev function (6).

[6] The usual goodness of fit test would be the well-known chi square test. Since the X^2 value increases linearly with an increase of sample size, it tends to yield significant result the sooner, the larger the sample is. Since this is the standard case in linguistics, the discrepancy coefficient is preferred for larger samples, with $C < 0.02$ being interpreted as a good, $C < 0.01$ as a very good fit.

A similar good fit is obtained with the continuous function (6): here, the goodness of fit uses to be evaluated with reference to the determination coefficient R^2, which in our case, with parameter values $C = 345$, $a = 0.12$, and $b = -0.68$ is excellent ($R^2 = 0.99$).[7] The result is graphically represented in Figure 3.

5 Conclusion

Given the theoretical discussions and empirical findings reported above, a number of conclusions seem to be justified, which shall be pointed out here:

(1) General extenders are fixed (stereotypic) pragmatic expressions.

(2) In a pragmatic framework, general extenders can be described as expressions with vague reference, used to indicate that there is more to say, to express assumption of shared knowledge, appealing for solidarity and understanding, or indicating lack of certainty.

(3) General extenders are a distinct category of linguistic elements, which refer to preceding items, and which are known for their fairly homogeneous structure and fixed clause position.

(4) As a distinct linguistic category, general extenders as a group lend themselves to processes of diversification (and, as a consequence, unification going along with it).

(5) As a result of the diversification processes, general extenders occur with varying frequencies.

(6) The frequencies of general extenders are regulated in a law-like manner, representing a process of self-regulation.

(7) For Slovenian general extenders, the well-known Zipf-Alekseev model (both as a continuous function and as a discrete probability model) turns out to be adequate to theoretically model the general extenders' frequencies and explain the generating process behind it.

(8) It will be interesting and important to do analogical analyses with linguistic material form other languages, in order to arrive at more general conclusions.

[7] Setting $C = f_1 = 342$ (cf. Table 3) results in a reduction of the number of parameters to be estimated; in this case, and with a slightly different parameter value for $a = 0.135$, the result is practically unchanged ($R^2 = 0.99$).

References

AIJMER, Karin, 2002: *English discourse particles: Evidence from a Corpus*. Amsterdam/Philadelphia: John Benjamins.

ALTMANN, Gabriel 2005: Diversification processes. Köhler, Reinhard / Altmann, Gabriel / Piotrowski, Rajmund G. (eds.): *Quantitative Linguistics. An International Handbook. Quantitative Linguistik. Ein Internationales Handbuch*. Berlin/New York: de Gruyter. 646–658.

ALTMANN, Gabriel / HŘEBÍČEK, Luděk, 1996: Levels of Order. Schmidt, Peter (ed.): *Glottometrika 15. Issues in General Linguistic Theory and the Theory of Word Length*. Trier: wvt. 38–61.

ARIEL, Mira, 1994: Pragmatic Operators. Asher, Ron E. (ed.): *Encyclopedia of Languages and Linguistics*, vol. 6. Oxford: Pergamon and Aberdeen University Press. 3250–3253.

BLAKEMORE, Diane, 2002: *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.

BURGER, Harald / BUHOFER, Annelies / SIALM, Ambros, 1982: *Handbuch der Phraseologie*. Berlin/New York: de Gruyter.

BURGER, Harald, 1998: *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt Verlag.

CHESHIRE, Jenny, 2007: Discourse variation, grammaticalisation and stuff like that. *Journal of Sociolinguistics* 11/2, 155–193.

CORTÉS RODRÍGUEZ, Luis, 2006: Los elementos de final de serie enumerativa del tipo ‘y todo eso, o cosas así, y tal etc. Perspectiva interactiva. *Boletin de Linguistica* XVIII (26), 102–129.

CUCCHI, Constanza A., 2007: *An investigation of general extenders in a corpus of EU parliamentary debates. Proceedings from the Corpus Linguistics Conference Series, University of Birmingham (UK), 27–30 July 2007.* <http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/242Paper.pdf>.

DINES, Elizabeth R., 1980: Variation in discourse – “and stuff like that”. *Language in Society* 9/1, 13–31.

DUBOIS, Sylvie, 1992: Extension particles, etc. *Language Variation and Change* 4, 179–203.

FERNANDEZ, Julieta / YULADASHEV, Aziz, 2011: Variation in the use of general extenders *and stuff* in instant messaging interactions. *Journal of Pragmatics* 43, 2610–2626.

FLEISCHER, Wolfgang, 1982: *Phraseologie der deutschen Gegenwartssprache*. Leipzig: VEB Bibliographisches Institut Leipzig.

FRASER, Bruce, 1999: What are discourse markers? *Journal of Pragmatics* 31/7, 931–952.

HAMMERL, Rolf, 1990: Untersuchungen zur Verteilung der Wortarten im Text. Hřebíček, Luděk (ed.): *Glottometrika 11*. Trier: wvt. 142–156.

HONDA, Akiko, 1996: Daigaku Koogi ni okeru kotoba no danjosa [= Gender differences in language use in university lectures]. *Kotoba* 16, 15.

HŘEBÍČEK, Luděk, 1996: Word Associations and Text. Schmidt, Peter (ed.): *Glottometrika 15. Issues in General Linguistic Theory and the Theory of Word Length*. Trier: wvt. 96–101.

JAKOP, Nataša, 2005: Opis pomena pragmatičnih frazemov s pragmatičnimi kategorijami. Kržišnik, Erika / Eismann, Wolfgang (eds.): *Frazeologija v jezikoslovju in drugih vedah / Phraseology in linguistics and other branches of science*. Ljubljana: Filozofska fakulteta. 167–182.

JUCKER, Andreas H. / SMITH, Sara W. / LÜDGE, Tanja, 2003: Interactive aspects of vagueness in conversation. *Journal of Pragmatics* 35, 1737–1769.

OVERSTREET, Maryann, 1999: *Whales, Candlelight and Stuff Like That. General Extenders in English Discourse*. Oxford: Oxford University Press.

OVERSTREET, Maryann, 2005: And stuff *und so*: Investigating pragmatic expressions in English and German. *Journal of Pragmatics* 37/11, 1845–1864.

PALACIOS MARTINEZ, Ignacio M., 2011: *"I might, I might go I mean it depends on money things and stuff"*. A preliminary analysis of general extenders in British teenagers' discourse. *Journal of Pragmatics* 43/9, 2452–2470.

PARVARESH, Vahid / TAVANGAR, Manoochehr / RASEKH, Abbas Eslami, 2010: General extenders in Persian discourse: Frequency and grammatical distribution. *Cross-cultural Communication* 6/3, 18–35.

POUTSMA, Hendrik, 1916: *A Grammar of Late Modern English*. Groningen, NL: P. Noordhoff.

PRODROMOU, Luke, 2008: *English as a Lingua Franca: A Corpus-based Analysis*. New York/London: Continuum.

RUZAITĖ, Jūratė, 2010: Translation equivalents of vague language items: A study of general extenders in parallel corpus. *Kalbų studijos / Studies about Languages* 16, 33–38.

SCHIFFRIN, Deborah, 1987: *Discourse Markers*. Cambridge: Cambridge University Press.

SCHOURUP, Lawrence, 1999: Discourse markers. *Lingua* 107, 227–265.

TAGLIAMONTE, Sali A. / DENIS, Derek, 2010: The stuff of change: General extenders in Toronto, Canada. *Journal of English Linguistics* 38/4, 335–368.

TERRASCHKE, Agnes, 2010: *Or so, oder so, and stuff like that* – general extenders in New Zealand English, German and in learner language. *Intercultural Pragmatics* 7/3, 449–469.

VERDONIK, Darinka / KAČIČ, Zdravko, 2012: Označevalci odprte propozicije v govorjenem diskurzu. *Slavistična revija* 52, 267–281.

VERDONIK, Darinka / ZWITTER VITEZ, Ana, 2011: *Slovenski govorni korpus GOS*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

WIERZBICKA, Anna, 1991: *Cross-cultural Pragmatics: The Semantics of Human Interaction*. Berlin: de Gruyter.

WINTER, Joanne / NORRBY, Catrin, 2000: *Set marking tags 'and stuff'. Proceedings of the 1999 Conference of the Australian Linguistic Society*. < http://www.als.asn.au/proceedings/als1999/winter&norrby.pdf>.

ZIPF, George K., 1935: *The psycho-biology of language. An introduction to dynamic philology*. Boston: Mifflin.

ZIPF, George K., 1949: *Human behavior and the principle of least effort. An introduction to human ecology*. Cambridge, Mass.: Addison-Wesley.

A Study of a new Phraseological Unit – *be on against* as an Example

Ai Inoue (Yokosuka)

Abstract

This research descriptively shows that *be on against*, consisting of *be* + a complex preposition (CP), functions a newly observed phraseological unit (PU) in contemporary English. Generally, a CP is defined as a word group that functions like a single preposition *e.* g. *according to*, *apart from*, and *in accordance with*. The recent trend in English is that two adverbial particles *in* and *on* co-occur with various prepositions and help establish new CPs like *in at*, *in for*, and *on against*, all of which, in most cases, are in the complement position of *be*. *Be on against* is frequently observed in corpora of present-day English. However, previous research and English dictionaries do not deal with it as far as I have seen in extensive reading of reference works. Hence, this research focuses on a PU *be on against* and describes its function in different contexts and how it comes to be established as a PU consisting of *be* + CP.

1 Introduction

This study shows that *be on against* consisting of *be* + a complex preposition (CP) functions a newly observed phraseological unit (PU), defined as repeatedly used word-strings consisting of at least two words, in contemporary English.

As an interesting phenomenon of present-day English, two prepositions are put together with a single meaning. Specifically, the prepositions *in* and *on* co-occur with various other prepositions, and each resulting prepositional phrase they form becomes established as a new PU consisting of *be* + CP with a single meaning. For example, *Macmillan English Dictionary*, 2nd edition (*MED*^{2}), describes PUs such as *be in at*, *be in for*, and *be in on* under the entry for *in*. Similarly, it includes various PUs, such as *be on about*, *be on at*, and *be on for*, under the entry for *of. Oxford Advanced Learner's Dictionary*, 8th edition (*OALD*^{8}) describes *be in at*, *be in for*, and *be in with* (s. v. *in*) and *be on about*, *be on at*, and *be on for* (s. v. *on*). However, not all PUs described in *MED*^{2} and *OALD*^{8} appear in other dictionaries. Even so, they can easily be retrieved from the British National Corpus (BNC) and other corpora. These PUs include *be in of*, *be in to*, *be on against*, and *in and of.* Especially, *be on against* shown in (1) and (2) can be frequently observed in contemporary English corpora (underlined by the author).

This study focuses on *be on against*, arguing that the phrase functions as a new PU. In addition, it shows the meaning of the phrase and the process of how it was established as a new PU and gained its new meaning.

(1) KING: Dr. Phil, his new book, "Self Matters: The Self Matters Companion," helping you create your life from the inside out. It's a companion to the number one best seller. It is now available in stores everywhere. And of course, he hosts the "Dr. Phil Show," syndicated. You would have to check newspapers in your area for time and station. One thing though, if you see Oprah, he won't be on against her. (COCA, spoken, 2002)

(2) KING: One hour?
MCGRAW: It's going to be an hour.
KING: You said 3:00, 4:00 in the afternoon. Are you going to be on against Oprah?
MCGRAW: Oh, absolutely not. My mama didn't raise a fool.
KING: Is that part of the rule, you can't be placed on a station…
MCGRAW: When we decided to do this, Oprah has created the show, of course. So there's no sense in us working at cross purposes. So if she's on at 4:00 in the afternoon, I'm generally on at 3:00. If she's on at 3:00, I'm on at 4:00. (Larry King Live, Feb. 27, 2002)

2 What is a CP?

According to previous research, prepositions can be classified into the following three types: 1) single prepositions such as *at*, *in*, and *of*; 2) compound prepositions, which consist of two or more prepositions used together, such as *into*, *onto*, *within*, and *until*; and 3) a prepositional phrase that functions as a complex preposition (CP), which consists of a preposition and another word, such as *due to* and *as far as*.

Generally, CPs are phrases such as *according to*, *apart from*, *in accordance with*, *with regard to*, *result of*, *in agreement with*, and *in case of*. They are further classified into three types according to their constituents. CPs of the first type consist of a preposition + noun + preposition (e. g. *by means of*, *in addition to*). Those of the second type are formed by an adjective, adverb, or conjunction + preposition (e. g. *ahead of*, *because of*). Lastly, CPs of the third type do not fit into any other subcategories (e. g. *as far as*, *thanks to*).

Quirk et al. (1985: 669) note that CPs are divided into two- and three-word sequences and define them as follows: "In the strict definition, a complex preposition is a sequence that is indivisible both in terms of syntax and in terms of meaning… Rather, there is a scale of 'cohesiveness' running from a sequence which behaves in every way like a simple preposition, to one which behaves in every way like a set of grammatically separate units…." (ibid.: 671f.). Further, Quirk et al. (1985: 671f.) offer the nine standards listed in (3), regarding whether or not word sequences comprising a preposition + noun + preposition function as CPs.

(3) a. Prep 2 can be varied: *on the shelf at* [but not: **in spite for*]

b. The noun can be varied as between singular and plural: *on the shelves by (the door)* [but not: **in spites of*]

c. The noun can be varied in respect of determiners: *on a/the shelf by*; *on shelves by (the door)* [but not: ** in a/the spite of*]

d. Prep 1 can be varied: *under the shelf by (the door)* [but not: **for spite of*]

e. Prep + complement can be replaced by a possessive pronoun: *on the surface of the table ~ on its surface* [but *in spite of the result ~ * in its spite*]

f. Prep 2 + complement can be omitted: *on the shelf* [but not: **in spite*]

g. Prep 2 + complement can be replaced by a demonstrative: *on that shelf* [but not: **in that spite*]

h. The noun can be placed by nouns of related meaning: *on the ledge by (the door)* [but not: **in malice of*]

i. The noun can be freely modified by adjectives: *on the low shelf by (the door)* [but not: **in evident spite of*]

3 *Be on against* – the change to a PU

This section discusses how *be on against* functions in context and describes the process of its formation.

3.1 Functions of *on* and *against*

No attempt has been made to give a functional explanation of *be on against*. As discussed in the research, *on* is the complement of the verb *be* and functions as an adverb. It is used to mean that something is happening or taking place. In case words like *actor/actress*, *TV programme*, or *movie* appear as a subject following *on*, *on* is used to show that the actor/actress, TV program, or movie is being broadcast. Please see the examples of *on* in (4), quoted from various dictionaries.

(4) a. Eastenders is <u>on</u> TV tonight. (*MED*[2])

b. What time is 'Star Trek' <u>on</u>? (*Longman Dictionary of Contemporary English,* 5th edition)

c. She'll be on soon. (*Youth Progressive English-Japanese Dictionary*)

d. You should go to Chicago while the festival is on. (*Longman English-Japanese Dictionary*)

The preposition *against* is commonly understood to mean 'in opposition to somebody or something', which means that it retains its lexical meaning. Moreover, there is not much difference in the current definition of *against* and previous studies on it.

3.2 Function of *be on against*

This section will describe the function of *be on against*. Please see examples (5) to (11); note that (5) and (6) are repetitions of (1) and (2), respectively.

(5) KING: Dr. Phil, his new book, "Self Matters: The Self Matters Companion," helping you create your life from the inside out. It's a companion to the number one best seller. It is now available in stores everywhere. And of course, he hosts the "Dr. Phil Show," syndicated. You would have to check newspapers in your area for time and station. One thing though, if you see Oprah, he won't be on against her. (COCA, spoken, 2002)

(6) KING: One hour?
MCGRAW: It's going to be an hour.
KING: You said 3:00, 4:00 in the afternoon. Are you going to be on against Oprah?
MCGRAW: Oh, absolutely not. My mama didn't raise a fool.
KING: Is that part of the rule, you can't be placed on a station…
MCGRAW: When we decided to do this, Oprah has created the show, of course. So there's no sense in us working at cross purposes. So if she's on at 4:00 in the afternoon, I'm generally on at 3:00. If she's on at 3:00, I'm on at 4:00. (Larry King Live, Feb. 2002)

(7) KING: But it was a Tuesday night show and a – under today's rules might not have made it? Right? Don-?
DON HEWITT: No, I don't – we never thought it was going to make it. You know-
KING: – would they have stayed with you?
DON HEWITT: – we've been on what, 29 years?
MIKE WALLACE: Twenty-eight.
DON HEWITT: I would have taken 29 weeks. Well, it will be 29.
KING: You were on against Marcus Welby, right?
DON HEWITT: Marcus Welby, M. D. Great doctor. (Larry King Live, May 1996)

(8) KING: She, of course, is starting her fourth season this Monday on – she's with Paramount domestic television. She's in 217 markets. And now her rival, replacing Mayor Ed Koch as the host of "The People's Court" is Judge Jerry Sheindlin, and that show is produced by Warner Brothers. And in many markets, they will be on against each other. Why, Jerry?
JUDGE JERRY SHEINDLIN, "THE PEOPLE'S COURT": Why not?

KING: Why did you do this?
JERRY SHEINDLIN: I don't know. Don't ask me why.
KING: Why did you take this? (Larry King Live, Sep. 1999)

(9) KING: Was it a hit right away?
VAN DYKE: No. We went in the toilet the first year.
KING: You are kidding!
VAN DYKE: We were on against "Perry Como," which was a very, very popular show. And, of course, my name didn't mean any – nobody had ever heard of me.
KING: Yes, you were a Broadway star then. (Larry King Live, Sep. 2000)

(10) KING: There's a shot, on a beautiful night in New York, of the Empire State Building. They have lit it up in blue and white tonight, the colors of the New York Yankees, who are participating in their 35th World Series. As you know, they're on against us. But, this program is repeated at midnight, Eastern time, 9:00 Pacific. (Larry King Live, Oct. 1998)

(11) KING: You'll have to come back.
Mr. MOYERS: Thank you very much, Larry.
KING: Bill Moyers – simply one of the best. Sports for Sale tomorrow night on PBS, a three-hour special including your phone calls. We'll be back tomorrow night with a full hour with H. Ross Perot – We'll be on against each other. Americans have a great choice tomorrow! Wolf Blitzer is on The Larry King Radio Show in one hour. Bernard Shaw and Susan Rook are next. (Larry King Live, Mar., 1991)

On the basis of the above examples, *be on against* is followed by a noun (phrase) and is used most often to mean 'compete'. Further investigation into the examples of *be on against* reveals two syntactic patterns: 'subject (e. g. somebody, TV or radio programme) + *be on against* + somebody / TV or radio programme' and 'subject (somebody or races) + *be on against* + each other'. The first pattern means that somebody or a TV/radio programme is competing with someone else or a TV/radio programme that is being broadcast simultaneously on another station. Examples (5), (6), (7), (9), and (10) can be classified as the first pattern. The second pattern means that people or races are competing with each other, as in the case of examples (8) and (11).

Let us consider each example of *be on against* in (5) to (11). In the case of (5), the subject *he*, follows *is on against*, which means that he competes with Oprah Winfrey, who hosts a popular talk show in the U. S. Hence, *be on against* is used to indicate that something competes with something else.

In (6), King and McGraw are talking about Oprah Winfrey. *Be on against* is used to ask McGraw whether his talk programme intends to compete with hers by airing simultaneously on another station.

Similar to (6), *be on against* in (7) is used to show that the programme co-hosted by Hewitt and Wallace for 28 years has been competing with that of Welby, which is counterprogrammed.

In (8), the second syntactic pattern 'subject + *be on against* + *each other*' is used. This pattern means that the programme under discussion and Judge Judy Sheindlin's strive against each other as if in a series of market competitions.

The first syntactic pattern is used in (9), when *be on against* is used to mention that the programme hosted by Dyke is counterprogrammed with 'Perry Como'.

Similarly, in (10), the first syntactic pattern is used in reference to 'Larry King Live' airing at the same time as the 35th World Series. *Be on against* in (10) means that a programme strives against another which is counterprogrammed.

Lastly, *be on against* in (11) uses the second syntactic pattern to say that a PBS programme competes with 'Larry King Live'.

These examples demonstrate that *be on against* is used to mean that something competes with another thing and has two syntactic patterns: 'subject + *be on against* + noun (phrase)' and 'subject + *be on against* + each other.' Also, the all examples above are quoted from spoken corpora and no examples of *be on against* are found in written materials. In other words, *be on against* is a brand new PU and it takes time to appear in written materials.

3.3 The reason why *be on against* is mainly observed in present tense

An examination of (5) to (11) reveals that *be on against* is used only in present and past tense as far as I have investigated from the data provided by BNC, WB, and COCA, which I could obtain. Please note that the examples (5), (6), (8), and (11) are used in future forms, but it is safe to regard future forms such as *will* and *be going to* as present tense because the time of utterance is done in present.

The reason why *be on against* is frequently observed in present tense is due to the function of *on*. As explained in the Section 3.1 (Functions of *on* and *against*), *on* is used to say that something is happening or taking place. Especially, when words like *actor/actress*, *TV programme*, or *movie* are used as a subject following *on*, *on* implies that the actor/actress, TV programme, or movie is being broadcast. Similarly, a person, TV or radio programme appears as a subject in the case of the two syntactic patterns of *be on against*, so *on* is used to show that something is happening or taking place. Consequently, it is natural to assume that *be on against* is mainly used in present tense.

However, some examples of *be on against* ((7) and (9)) are used in past tense. Similar with the examples of *be on against* in present tense, the subjects following *on* are a person, but the function of *on* seems to have been weakened in past tense. Thus, *be on against* is also observed in past tense. A few examples of *be on against* are found in present-day English, the study continues to show the relation between the phrase *be on against* and tense by collecting the examples of *be on against*.

3.4 Why does the CP *on against* co-occurs with the verb *be*?

As I have explained in the Sections 3.2 and 3.3, *be on against* is established as an independent PU and is used with mainly in present tense due to the function of *on*. This sections focuses on the reason why *be on against* collocates with *be*.

As far as I have investigated what kinds of verbs other than *be* go with *on against*, the verbs are as follows: *go*, *work*, *carry*, *focus*, *hold*, *hang*, *move*, *come*, and *keep*, all of which are repeatedly used with *on against*. From a cursory examination of the verbs, the following fact emerges: all verbs co-occurring with *on* are established as a phrasal verb such as *go on*, *work on*, *carry on*, *focus on*, *hold on*, *hang on*, *move on*, *come on*, and *keep on*. To cite a couple of examples, *carry on against* syntactically consists of [carry on] [against] and *work on against* is formed of [work on] [against]. The syntactic structure pattern holds true for other phrasal verbs. To put it differently, the expressions such as *work on against*, *hold on against*, and *focus on against* accidentally put a phrasal verb like *work on* and *come on* together with *against*. Hence, the lexical meanings of each phrasal verb survive.

On the other hand, *be* is the most appropriate verb for *on against* to show its meaning (i. e. compete) because other verbs such as *go*, *come*, *carry*, *focus*, and *hold* have not been semantically bleached compared to *be*. In other words, *be* is the most semantically bleached verb (i. e. unmarked) and does not impede the meaning of *on against*. Consequently, *on against* construes with *be*.

3.5 Formation and development of *be on against*

This section discusses the process of how *be on against* became established and came to have its own meaning.

First, 'concept categorisation' turns *be on against* into a PU. According to Yagi (1999: 105ff.), concept categorisation is a process in which various concepts expressed in various syntactic units are interpreted as one syntactic unit. The sentences in (12) are an example of such usage quoted from Yagi (ibid.: 105).

(12) a. John is far from the destination.
b. John is far from being honest.
c. John is far from honest. (Yagi 1999: 105)

Far from in (12a) refers to physical long distance, while *far from* in (12b) and (12c) metaphorically refers to John's being almost the opposite of honest. The process of deriving this use of *far from* from its basic use in the physical sense is as follows: The relevant part of (12a) is syntactically analysed as [far [from the destination]], in which *far* is an adverb followed by the prepositional phrase *from the destination*. It is later reanalysed as [[far from] the destination], in which *far from*, as a lexical unit, holds an idiomatic status expressing the meaning of 'a physically long way away from'. Once this idiomatic status is given to *far from*, the next step to the semantic expansion of *far from* from its typical sense to its metaphorical sense is easily taken, as can be seen in (12b) and (12c) (Inoue 2007: 125f.). In addition to the example of *far from*, phrases such as 'cultural historian' and 'change of address card' are due to the working of concept categorisation.

Let us return to the discussion of how *be on against* is formed. The process of deriving the demonstrated use of *be on against* from its basic use in its original sense is as follows: As is commonly assumed, *be on against* syntactically consists of [be] [on] [against], as in [The war] [is] [on] [against] [poverty], which means that the war to eliminate poverty continues. However, [be] [on] [against] becomes [be] [on against] by its repeatedly use and finally turns into [be on against]. [Be on against] is reanalysed as a lexical unit by its repeatedly use, coming to hold the idiomatic meaning 'compete'. For example, [the war is][on][against poverty] becomes [the war [is] [on against] poverty], and then finally become [the was [is on against] poverty] which means that the war continues as a fight against poverty. By moderate advances, any word implying 'compete' can come to appear as the subject of a sentence, so *be on against* develops the meaning that one programme or person competes with another that is counterprogrammed. Thanks to the working of concept categorisation, the lexical items *be*, *on*, and *against* are combined in the lexical unit [be on against]. Then, *be on against* is attached to a prepositional function and comes to behave like a PU consisting of '*be* + CP'.

This section also confirms that *on against* functions as a CP based on Quirk et al.'s (1985) criteria. As (3) shows, Quirk et al. (1985) set nine criteria for a unit consisting of a preposition + noun + preposition to be judged as a CP. However, in the case of *on against*, the nine criteria do not apply because the phrase does not have the structure preposition + noun + preposition. However, when we apply Quirk et al.'s (1985: 669) opinion – "In the strict definition, a complex preposition is a sequence that is indivisible both in terms of syntax and in terms of meaning" – to the case of *on against*, we see that *on against* in (5) to (11) cannot be replaced by any words or divided into segments syntactically or semantically. Also, *on against* is accompanied with a noun (phrase), which means that *on against* works as a preposition. Consequently, *on against* is formed by the working of concept categorisation and is established as a CP. This phenomenon is an interesting trend in contemporary English.

The development of such PUs, which consist of *be* and a CP, has been recently observed. Many phrases (*be in at*, *be in for*, *be in on*, *be in with*, *be on about*, *be on at*, and *be on for*) found in dictionaries are listed in the Introduction, but some of these (*be in at*, *be on about*, *be on at*, and *be on for*) are not included in the *OED*^{2} (*Oxford English Dictionary*, 2nd edition). Furthermore, other phrases (*be in of*, *be in to*, *be on against*, *in and of*) are contained neither in the *OED*^{2} nor in other dictionaries. *OED*^{2} describes only three phrases as follows: '*Be in for* A' is defined as getting involved with A, '*be in on* A' as participating in or having knowledge of A, and '*be in with* A' as agreeing or keeping in touch with A. The definitions for these three phrases as given in the *OED*^{2} show no semantic differences with definitions in other dictionaries.

4 Concluding remarks

This study introduced the new phraseological unit *be on against*, which functions as a PU, and explained the process by which it came to acquire its new meaning. The phenomenon of combining two prepositions into a single CP, which is in the complement position of *be*, is one of the interesting trends of contemporary English.

Acknowledgements

This research is supported by both Grant-in-Aid for Challenging Exploratory Research (24652097) (KAKENHI) and Grant-in-Aid for Special Research of National Defense Academy of Japan. I appreciate the Japan Society for the Promotion of Science, Ministry of Education, Culture, Sports, Science and Technology, and the Ministry of Defense.

References

Corpora

BNC: British National Corpus

WB: WordBanks*Online*

LKL: Larry King Live Corpus

COCA: The Corpus of Contemporary American English

Dictionaries

Longman Dictionary of Contemporary English, 2008. 5th edition. London: Longman.

Longman English-Japanese Dictionary, 2007. Tokyo: Kirihara Shoten.

*MED*2, 2007: *Macmillan English Dictionary*. 2nd edition. Oxford: Macmillan Education.

*OALD*8, 2010: *Oxford Advanced Learner's Dictionary*. 8th edition. Oxford: Oxford University Press.

Oxford English Dictionary, 2009. 2nd edition. CD-ROM version 4.0. Oxford: Oxford University Press.

Youth Progressive English-Japanese Dictionary, 2006. Tokyo: Syougakukan.

Bibliography

INOUE, Ai, 2007: *Present-day Spoken English: A Phraseological Approach*. Tokyo: Kaitakusha.

QUIRK Randolph et al, 1985: *A Comprehensive Grammar of the English Language*. London: Longman.

YAGI, Katsumasa, 1999: *Eigo no Bunpo to Goho – Imikara no Approach* (*A Semantic Descriptive Approach to Modern English*). Tokyo: Kenkyusha Syuppan.

Paarformeln und Binomiale im Slowenischen: Ein korpusbasierter Ansatz

EMMERICH KELIH (Wien)

Abstract

Der vorliegende Beitrag behandelt reversible Paarformeln im Slowenischen. Nach einer einleitenden Diskussion über die Gründe für die Reihenfolge der einzelnen Komponenten einer Paarformel wird näher auf die synchrone Verwendungshäufigkeit von reversiblen Paarformeln eingegangen. Das Ausmaß der Variabilität in der Reihenfolge wird in einem synchronen Referenzkorpus (*FidaPlus*) des Slowenischen untersucht. Es stellt sich heraus, dass bei reversiblen Paarformeln zwischen drei Gruppen, nämlich nicht belegbaren bzw. sehr selten vorkommenden Paarformeln und einer Gruppe mit einer relativ hohen Verwendungshäufigkeit unterschieden werden sollte. Allerdings ist die Untersuchung als exemplarisch-symptomatisch und nicht als systematisch-analytisch zu verstehen.

1 EINLEITUNG

Paar- und Zwillingsformeln erwecken immer wieder das Interesse von Vertretern aus unterschiedlichen Wissenschaftsbereichen (allgemeine Linguistik, Phraseologie im breitesten Sinne, Sprachlehrforschung, Sprachdidaktik, usw.); damit einher geht eine unterschiedliche Fokussierung und Perspektivierung auf ausgewählte Fragestellungen (phonetisch-phonologische, morphosyntaktische Struktur, lexikalische, semantische Struktur, sprach- und kulturspezifische Ausprägungen usw.). Aus linguistischer Sicht bietet es sich an, zwischen Binomialen (weitere Bezeichnungen sind binomiale Wortverbindungen, Freezes, koordinierte Nominalphrasen, cojoined expressions, phrasal conjuncts, Wortpaare, koordinative Binomiale, Binomialbildungen) und Paar- und Zwillingsformeln zu unterscheiden.

Lässt man die Feinheiten einiger konzeptueller Unterschiede zwischen diesen Bezeichnungen beiseite, so versteht man unter Binomialen (binomials) nach der Definition von Malkiel (1959: 113) eine Sequenz von zwei Wörtern, die der gleichen Formanten-Klasse angehören, auf der gleichen syntaktischen Ebene zu positionieren sind und in der Regel durch einen bestimmten lexikalischen Konnektor (z. B. Konjunktionen) verbunden sind. Ein wichtiges Merkmal von Binomialen ist die Reihenfolge der Komponenten – der jeweiligen Wortformen, die der gleichen Formantenklasse angehören. Hierbei wird zwischen irreversiblen Binomialen, die keine Freiheit und Varianz in der Abfolge der einzelnen Komponenten zulassen und reversiblen Binomialen unterschieden, die hinsichtlich der Abfolge frei und variabel sind. Als Beispiel für Binomiale im Deutschen seien *Freiheit und Gleichheit*, *Freund und Helfer* und *mit Kind und Kegel* genannt, die im Deutschen als irreversible Binomiale gelten.[1]

Die Definition von Malkiel (1959) basiert ausschließlich auf formalen Kriterien; dennoch lässt sich die konzeptuelle Nähe von Binomialen zu idiomatischen Wendungen bzw. phraseologischen Formeln im Allgemeinen nicht leugnen. In der Phraseologie wird in diesem Kontext von sogenannten Paarformeln bzw. Zwillingsformeln gesprochen. In der Regel versteht man darunter unveränderliche, durch Konjunktion oder Präposition verknüpfte Wortpaare, die häufig durch Alliteration, Assonanz, Stabreim u. ä. verbunden sind (Bußmann 1990: 984, Burger/Buhofer/Sialm 1982: 37). Es können auch noch weitere Subformen von Zwillingsformen, wie z. B. solche ohne entsprechende Konjunktion (z. B. *wennschon dennschon*), oder Zwillingsformeln mit Wiederholung der Komponenten (*Schulter an Schulter*, *Zug um Zug* usw.) unterschieden werden. Burger/Buhofer/Sialm (1982: 37) verweisen darauf, dass Zwillingsformeln insgesamt einen unterschiedlichen Grad an Idiomatizität zeigen oder nur zum Teil idiomatisch motiviert sein können.

Zusammengefasst gesagt sind Binomiale bzw. Paarformeln ein breites und in den letzten Jahren intensiv diskutiertes Forschungsfeld, insbesondere innerhalb der allgemeinen Sprachwissenschaft (Landsberg 1995, Müller 2009, Southern 2000 u. v. m.). Im vorliegenden Beitrag wird die Festigkeit der Abfolge der Komponenten von Paarformeln anhand von ausgewählten slowenischen Binomialen bzw. Zwillingsformeln im Detail analysiert. Hierbei wird die Reversibilität bzw. Irreversibilität deren einzelner Komponenten untersucht. Das Material für die Analyse stammt aus Keber (2011), dem derzeit aktuellsten slowenischen phraseologischen Wörterbuch. Die Reversibilität innerhalb von Paarformeln wird durch entsprechende Belege aus dem slowenischen Referenzkorpus *FidaPlus* belegt; besonderes Augenmerk wird auf deren Vorkommenshäufigkeit gerichtet.

2 Beschränkungen der Positionierung von Komponenten

Hinsichtlich der Abfolge der einzelnen Komponenten innerhalb von Paarformeln gibt es zwei Möglichkeiten: (1) die Reihenfolge ist fix und nicht

[1] Die Beispiele sind Lambrecht (1984) entnommen.

veränderbar, und (2) die Reihenfolge obliegt einer gewissen Variabilität, d. h. die Komponenten können ausgetauscht werden, ohne dass dies mit im weiteren Verlauf noch zu diskutierenden Konsequenzen verbunden wäre. In der Forschung zu Paarformeln ist in der Vergangenheit allerdings weniger die Aufmerksamkeit auf die Reihenfolge der Komponenten gerichtet worden, als vielmehr auf die Frage nach Faktoren, die auf die Positionierung von Komponenten einen Einfluss haben können.

Die Bandbreite möglicher Einflussfaktoren kann an dieser Stelle nicht im Detail diskutiert werden, sondern wird nur in den wichtigsten Eckpunkten zusammengefasst. Eine wichtige Rolle wird pragmatischen bzw. semantischen Faktoren zugesprochen, die Cooper/Ross (1975) plakativ zu einem „Me-First-Principle" zusammengefasst haben. Postuliert wird, dass in erster Linie bestimmte semantische Eigenschaften den Ausschlag geben, warum eine der Komponenten eher in der ersten als in der zweiten Position zu finden ist. So wird u. a. davon ausgegangen, dass z. B. Belebtes vor Unbelebtes, Männliches vor Weibliches, Erwachsenes vor Kindliches usw. gestellt wird. In dieser Hinsicht wird auch versucht, allgemeine Prinzipien zu finden, die die Positionierung der Komponenten in Paarformeln determinieren. Als ein mögliches Prinzip wird u. a. die Markiertheit der einzelnen Glieder ins Spiel gebracht (Cooper/Ross 1975: 66–67). Demnach werden jeweils diejenigen Komponenten vorangestellt, die eine breitere, allgemeinere Bedeutung haben und weniger Distributionsbeschränkungen unterworfen sind (Benor/Levy 2006: 237).

Ein weiterer wichtiger Einflussfaktor ist die phonetische bzw. phonologische Struktur der Komponenten. Unter anderem können die Vokalqualität, die Länge des Vokalnukleus, die Anzahl von initialen Konsonanten usw. eine Rolle spielen. Diskutiert werden auch suprasegmentale Eigenschaften, die eine entscheidende Bedeutung für die Positionierung haben können. So werden u. a. die prosodische Struktur, die Betonungsposition bzw. die Silbenstruktur (offene Silben im Anlaut) der betreffenden Wortformen im Allgemeinen ins Spiel gebracht (Bolinger 1962, Cooper/Ross 1975: 71).

Neben semantisch-pragmatischen, morphologischen und phonologischen Faktoren spielt auch die Länge der Komponenten eine Rolle. Es wird davon ausgegangen, dass ein organisierendes Prinzip von Zwillingsformeln darin zu sehen ist, dass die erste Komponente kürzer als die zweite ist. Gemeint ist damit die Länge von Wortformen (ohne Berücksichtigung der Konjunktionen), die in der Anzahl von Phonemen, Silben, Morphen usw. gemessen werden kann. Kurz zusammengefasst: Es wird von einem „Kurz-vor-Lang-Prinzip" ausgegangen, welches eben besagt, dass die kürzere Komponente der längeren Komponente tendenziell vorangestellt wird (vgl. u. a. Malkiel 1959: 136, Benor/Levy 2006: 242).

Eine neue Perspektive hinsichtlich der Motivierung der Positionierung der einzelnen Komponenten in Binomialen bzw. Paarformeln eröffnet Fenk-Oczlon (1989) mit ihrem Beitrag *Word frequency and word order in freezes*. In diesem Zusammenhang wird die Vorkommenshäufigkeit der einzelnen Wortformen als determinierender Faktor ins Spiel gebracht: Häufiges wird vor weniger Häufiges gestellt. Ohne Zweifel ist die Häufigkeit eine wichtige linguistische Eigenschaft (vgl. u. a. Bybee/Hopper 2001), deren Bedeutung sich auch vor allem daran zeigt, dass diese eine integrale Eigenschaft systemischer Wechselbeziehungen ist. Hinzuweisen ist auf die in der Linguistik gut bekannte Wechselbeziehung zwischen der Vorkommenshäufigkeit und der Länge von Wortformen im Sinne des Zipf'schen Gesetzes – ein Aspekt, der hier nicht im Detail besprochen werden kann. Insgesamt bleibt festzuhalten, dass sowohl die Vorkommenshäufigkeit als auch die Länge der Komponenten als Einflussfaktoren betrachtet werden können. Darüber hinaus ergeben sich in diesem Punkt interessante Querverbindungen zu Fragen der sprachlichen Ökonomie und der kognitiven Sprachverarbeitung.

Die bislang genannten Faktoren geben insgesamt Grund zu der Vermutung, dass nicht ein einziger Faktor für die Positionierung der Komponenten verantwortlich ist, sondern dass vielmehr ein ganzes Set von miteinander verwobenen Faktoren und Eigenschaften ins Spiel kommt. Darüber hinaus ist aber vor allem der Variabilität in der Abfolge der einzelnen Komponenten mehr Aufmerksamkeit zu schenken, denn – wie bereits einleitend erwähnt – ist nicht in allen Paarformeln von einer unveränderlichen Festigkeit der Reihenfolge auszugehen, sondern die Vertauschbarkeit der Komponenten ist im Gegenteil mitunter ein integrales Kennzeichen von vereinzelten Paarformeln. Dieses Phänomen soll nunmehr exemplarisch anhand einiger Beispiele aus dem Slowenischen untersucht werden.

2.1 Paarformeln im Slowenischen

Als eine der wenigen Untersuchungen, die explizit dem Thema slowenische Paarformeln gewidmet ist,[2] lässt sich Toporišič (1996, 1998) nennen. Terminologisch werden diese als *dvojčiči* bzw. *dvojčični obrazci* bezeichnet, deren Definition sich mit der in der Einleitung erwähnten (Burger/Buhofer/Sialm 1982, Bußmann 1990) weitgehend deckt. Toporišič (1996) unterscheidet mehrere

[2] Insgesamt scheint das Thema Paar- und Zwillingsformeln im Slowenischen bislang wenig untersucht worden zu sein. Diesen Schluss lässt die umfangreiche, kommentierte Bibliographie zur slowenischen Phraseologie und zur lexikographischen Verarbeitung von phraseologischen Wendungen in Keber (2000) zu.

Gruppen von Paar- und Zwillingsformeln, deren zwei wichtigste Gruppen die folgenden sind:

(1) Paarformeln mit Wiederholung der Wortform (= semantische Übereinstimmung) und ohne eine Konjunktion als Bindeglied, wie z. B. *ti ti*, *prima prima*, *tiho tiho*, *čiča čiča*, *ne ne* usw., die im Grunde genommen als Wortwiederholung (Geminatio) in Erscheinung treten. Auffällig ist bei dieser Art von Gruppierung, dass auch die Reduplikation von Interjektionen (*ha ha*, *ne ne*) als Paarformeln aufgefasst werden. Als eine weitere Sonderform wird die Aneinanderreihung von Grundform und Steigerungsform eines Adjektivs angesehen (*dober predober*, *temen pretemen*). Auch Wiederholungen der gleichen Wortform, allerdings mit einem lexikalischen Konjunktor, wie z. B. *prošnje in prošnje*, *tisoči in tisoči*, *garaj in garaj* usw. werden als Paarformeln aufgefasst. Es wird in diesem Zusammenhang auch darauf hingewiesen, dass nicht nur Konjunktionen als Bindeglied auftreten können, sondern auch Präpositionen, wie aus den angeführten Beispielen ersichtlich ist: *človek proti človeku*, *dan na dan* und *mesto ob mestu* usw. Bei dieser Art von Paarformeln stellt sich – außer bei den Adjektiven – aufgrund der Wiederholung der gleichen Wortformen die Frage einer Varianz in der Reihenfolge der Komponenten nicht.

(2) Die zweite Gruppe sind Paarformeln, in denen zwei unterschiedliche, in der Regel antonymische Wortformen auftreten (Toporišič 1996: 271), die in der Regel durch eine Konjunktion (*in*, *pa*, *ter*, *ali*) verbunden werden, wie *noč in dan*, *denar ali življenje*, *hoče in noče*, *staro in mlado*. Zu dieser Gruppe werden auch *Paarformeln* wie *mož – žena*, *oče – mati* gezählt, die an anderer Stelle als Dvadva-Komposita aufgefasst werden (Wälchli 2009). Hingewiesen wird auch auf die Möglichkeit der Verneinung der zweiten Komponente (*poklican in nepoklican*, *zgode in nezgode*, *Slovenci in Neslovenci*), welches als ein beliebtes morphologisches Verfahren in Paarformeln angeführt wird.

Des Weiteren verweist Toporišič (1996: 273) auf rhetorische Figuren, die in Paarformeln gehäuft auftreten. Insbesondere werden Reimstrukturen (*vik pa krik*, *ne bev ne mev*), Alliterationen (*bodi Peter bodi Pavel*, *papir pa pero*) und Assonanzen (*ne tič ne miš*, *dolg in širok*, *berač in kralj* usw.) hervorgehoben. In diesem Sinne liefert Toporišič (1996) in erster Linie eine theoretische Auseinandersetzung mit Paarformeln, die anhand von Beispielen aus dem Slowenischen illustriert werden.

Eine erste systematische Sammlung von slowenischen Paarformeln findet sich in dem erst kürzlich erschienen slowenischen phraseologischen Wörterbuch (*Slovar slovenskih frazemov*) von Keber (2011). In diesem Werk, dem ersten umfangreichen Wörterbuch dieser Art,[3] sind Paar- und Zwillingsformeln mit einem eigenen Qualifikator *dvoj.* (*dvojčični frazem* – zwillingshaftes Phrasem) ausgezeichnet, womit zumindest ein Kernbestand an slowenischen Paarformeln eruiert werden kann. Eine Durchsicht des Wörterbuchs ergibt in Summe etwas mehr als 100 derartiger Paarformeln. Eine Auswahl davon wird nunmehr als Ausgangspunkt für eine Untersuchung der Variabilität innerhalb der Reihenfolge der Komponenten herangezogen.

2.2 Irreversibilität in Paarformeln

Eines der konstitutiven Merkmale einer Paarformel ist die sogenannte Irreversibilität der einzelnen Komponenten; in letzter Instanz geht es um die Festigkeit der Reihenfolge von Komponenten in phraseologischen Wendungen. Müller (1997: 7) spricht in diesem Zusammenhang von einer starken Tendenz zur Irreversibilität von Paarformeln. Bei einer Änderung der Abfolge der Komponenten kann dies die Konsequenz haben, dass die entsprechend veränderte Paarformel (1) als ungrammatisch empfunden wird und (2) damit ein Verlust der Formelhaftigkeit einhergeht und so (3) die semantische Transparenz verloren gehen kann. Allerdings ist festzuhalten, dass die Festigkeit der Reihenfolge von Komponenten tatsächlich als eine Tendenz aufzufassen ist, da immer wieder eine Änderung der Reihenfolge zu beobachten ist. In anderen Worten: Offenbar erweisen sich einige Paarformeln als tolerant gegenüber einer Abänderung der Reihenfolge, ohne dass dabei die Idiomatizität oder die grammatikalische Wohlgeformtheit verletzt werden.

Um diesen Sachverhalt anhand einer in Keber (2011) verzeichneten Paarformel zu demonstrieren: Unter der Paarformel *dan in noč* (*Tag und Nacht*) (S. 152) findet sich der Querverweis auf *noč in dan* (S. 585) mit dem entsprechendem Vermerk, dass es sich hierbei um eine neutrale stilistische Variante der erstgenannten Form handelt. In beiden Fällen kann von einer Bedeutung im Sinne von ‘ununterbrochen, ohne Pause’ ausgegangen werden. Im vorliegenden Fall sind somit – zumindest suggeriert dies die lexikographische Praxis – beide Varianten mehr oder weniger als gleichberechtigt anzusehen. In jedem Fall ist man hierbei mit einem klaren Fall der sowohl vom Sprachsystem als auch vom

[3] Keber (2011) geht von einem breiten Phraseologiebegriff aus und erfasst in seinem Wörterbuch insgesamt circa 10.000 Phraseme und phraseologische Wendungen. Vgl. dazu auch Rojs (2012) mit einer Rezension zu Keber (2011) bzw. Kržišnik (2004), die eine detaillierte Rezension zu einem früheren Probeheft dieses Wörterbuchs verfasst hat.

Rezipienten zugelassenen Variabilität in der Reihenfolge der Komponenten konfrontiert.

Der Fall von *dan in noč* und *noč in dan* zeigt aber auch ein grundsätzliches Problem auf: Sofern Variabilität der Reihenfolgen zugelassen ist, kann nicht ohne weiteres von einer (phonetischen, morphologischen, semantischen) Motivierung der Reihenfolge gesprochen werden, zumindest nicht von einer einfaktoriellen oder unidimensional motivierten. So sind im obigen Fall beide Komponenten sowohl hinsichtlich der Anzahl von Silben und Lauten gleich lang; aus semantischer Sicht ist sowohl ein Prinzip „hell vor dunkel" als auch „dunkel vor hell" möglich. Dennoch muss hier nicht unbedingt von gleichberechtigten (stilistischen) Varianten gesprochen werden. Es bietet sich an, einen präferierten Typus anhand der Vorkommenshäufigkeit der einzelnen Varianten in einem sprachlichen Korpus heranzuziehen. Für das Slowenische eignet sich in dieser Hinsicht z. B. das *FidaPlus-Korpus*,[4] welches zum gegenwärtigen Zeitpunkt als wichtigstes slowenisches Referenzkorpus anzusehen ist. So haben die angeführten Paarformeln (*dan in noč* vs. *noč in dan*) im Korpus folgende Vorkommenshäufigkeit: *dan in noč* kommt 1496-mal vor, während die umgekehrte Form *noč in dan* insgesamt 2447-mal belegt ist. In diesem Fall ist – solange keine weiteren systematischen Vergleiche[5] vorliegen – davon auszugehen, dass diese beiden Formen aus der Sicht der Vorkommenshäufigkeit nicht als gleichverteilt anzusehen sind, sondern tendenziell eine Präferenz für die Form *noč in dan* festzustellen ist.

Dieses einleitende Beispiel ist Ausgangspunkt für eine systematische Analyse der Variabilität der einzelnen Komponenten in Paarformeln. Um eine entsprechende Abgrenzung des Datenbestandes vornehmen zu können, werden in die-

[4] Die Verwendung des FidaPlus-Korpus steht – nach einer entsprechenden Anmeldung – für wissenschaftliche Zwecke zur Verfügung. Keber (2011: 12–13) verweist in diesem Zusammenhang darauf, dass bei der Erstellung des Wörterbuches – welches auch kommerziellen Zwecken dient – auf dieses aktuell größte und umfangreichste Korpus des Slowenischen bei der Erstellung nicht zurückgegriffen werden konnte und auch keine Belege aus diesem Korpus angeführt werden konnten. Aus diesem Grund hat Keber (2011) neben allgemeinen Suchmaschinen vor allem auf das Korpus *Nova Beseda* (http://bos.zrc-sazu.si/s_beseda.html) zurückgegriffen, welches aber viel – und vor allem doch eher ältere –slowenische Prosa beinhaltet. Das Referenzkorpus *FidaPlus* beinhaltet hauptsächlich Texte aus den Jahren 1990 bis 2006. Auch wenn unterschiedliche Textsorten und Funktionalstile berücksichtigt sind, ist ein Überhang an journalistischen Texten (über 80 % des Materials aus 615 Millionen Wortformen) nicht zu übersehen. Details zum Projekt Fida bzw. FidaPlus vgl. u. a. Arhar/Gorjanc (2007).

[5] Dieses Ergebnis ist tentativ und als ein sehr grober Gradmesser anzusehen, da eine nichtidiomatische Verwendung nicht ausgeschlossen ist. Bei weiterführenden Analysen müsste also jeder einzelne Beleg manuell überprüft werden.

sem Beitrag ausschließlich Paarformeln[6] aus Keber (2011) untersucht, die dort explizit als erlaubte und mögliche reversible Paarformeln ausgewiesen sind.

Betrachtet man diese Gruppe von reversiblen Paarformeln und versucht zu erkennen, ob es bestimmte Merkmale gibt, die diese Gruppe auszeichnen, so ergibt sich – dies als Vorausgriff auf die entsprechenden Ergebnisse – die Möglichkeit einer dichotomen Aufteilung: (1) Einerseits handelt es sich hierbei um offenbar wenig frequente Paarformeln, die sich darüber hinaus z. T. durch eine veraltete und archaische Lexik auszeichnen. (2) Andererseits ergibt sich eine Gruppe, die durch eine relativ hohe Verwendungshäufigkeit ausgezeichnet ist, in der aber dennoch eine bestimmte Tendenz zur Festigung bzw. Stabilisierung der Reihenfolge der Komponenten festzustellen ist. Dies wird nun im Folgenden anhand einiger ausgewählter Beispiele demonstriert.

2.2.1 Reversible Paarformeln mit geringer Vorkommenshäufigkeit

Die im vorangehenden Kapitel vorgeschlagene Charakterisierung von reversiblen Paarformeln wird im Folgenden anhand von ausgewählten Fallbeispielen näher diskutiert. Begonnen wird mit der Paarformel *ne bati se ne biriča ne hudiča*, die nach Keber (2011: 271) auch in der Form *ne bati se ne hudiča ne biriča* angeführt wird [weder den Schergen noch den Teufel fürchten] bzw. im umgekehrter Reihenfolge. Somit wird von Keber (2011) eine prinzipielle Reversibilität der Komponenten vorgeschlagen – ein Vorschlag, der im wichtigsten einsprachigen slowenischen allgemeinen Wörterbuch (*Slovar slovenskega knjižnega jezika* 1970–) in dieser Form jedoch nicht zu finden ist. In erster Linie fällt auf, dass der in der Paarformel genannte *birič* [Scherge, Büttel] einer veralteten lexikalischen Schicht zuzuordnen ist, sich aber eben offenbar in dieser Paarformel in phraseologischer Bedeutung im Sinne von 'Furchtlosigkeit' erhalten hat.

Untersucht man die aktuelle Vorkommenshäufigkeit dieser Wendung im Korpus, so ist das Ergebnis einigermaßen ernüchternd: Keine der beiden vorgeschlagenen Formen ist im Korpus nachzuweisen, sodass mit aller Vorsicht eine insgesamt geringe Verbreitung im gegenwärtigen Standardslowenisch zu konstatieren wäre. Hervorgehoben werden muss aber, dass die Komponente *hudič* [Teufel] gleichzeitig produktiv bei der Bildung von lexikalischen Varianten – so die Belege im genannten Korpus – dieser Paarformeln ist (*ne smrti ne hudiča*, *ne boga ne hodiča*, *ne hudiča ne boga*, *ne hudiča niti zlega duha*).

[6] Ein Desiderat der slowenischen Phraseologieforschung ist sicherlich die systematische lexikographische Erfassung von Paarformeln. Die bei Keber (2011) erfassten Paarformeln – es sind dies ca. 100 Einträge (s. o.) – sind vermutlich nur ein Bruchteil des gesamten Bestandes.

Allerdings ist die Vorkommenshäufigkeit dieser Paarformeln sehr gering bzw. sind sie de facto nur einmal belegt; außer in der Kombination von *ne boga ne hudiča*, welches zumindest zehnmal im Korpus nachzuweisen ist.

Ein ähnlicher Fall, d. h. geringe Vorkommenshäufigkeit, liegt bei der von Keber (2011: 409) angeführten Paarformel *kri in mleko* [Blut und Milch] bzw. *mleko in kri* vor. Diese Paarformel, die hauptsächlich in Vergleichen verwendet wird (als Ausdruck für 'in einer guten gesundheitlichen Verfassung sein'; jmd. ist wie *kri in mleko*), zeigt die Korpusanalyse, dass *kri in mleko* zwar in dieser Form belegt ist (Vorkommenshäufigkeit: 16), sich hierbei aber keine idiomatische Bedeutung erkennen lässt. Es handelt sich z. B. um den Titel eines Buches bzw. Films, um den direkten Bezug auf die Aufnahme der beiden Flüssigkeiten; ähnliches lässt sich für die potenziell reversierte Form, nämlich *mleko in kri*, die insgesamt nur fünfmal belegt ist, feststellen.

Mit anderen Worten: Eine von der Lexikographie suggerierte Reversibilität in Paarformeln erweist sich bei einer näheren Analyse im Grunde genommen als empirisch nicht belegbares Phänomen. Die angeführten Paarformeln haben insgesamt eine sehr niedrige Verwendungshäufigkeit, und dies unabhängig von der Tatsache, ob sie reversibel sind oder nicht. A priori wäre anzunehmen, dass eine zugelassene Reversibilität der Komponenten mit einer höheren Verwendungshäufigkeit einhergeht, da die Komponenten in einer nicht fix vorgegebenen Reihenfolge reproduziert werden müssen, so dass die Wahrscheinlichkeit ihres Auftretens zunehmen sollte. Im gegebenen Fall – die Analyse stützt sich einstweilen allerdings nur auf das Material aus Keber (2011) – zeigt sich aber, dass man es mit Sprachmaterial zu tun hat, welches offenbar den synchronen Bestand des Slowenischen nicht (mehr) wiedergibt.

2.2.2 Reversible Paarformeln mit hoher Vorkommenshäufigkeit

Wenn im gegeben Zusammenhang von einer hohen Vorkommenshäufigkeit gesprochen wird, so hat man hier nicht die absolute Häufigkeit einer Paarformel in einem Korpus vor Augen; vielmehr ist dies in Relation zu der im vorangehenden Kapitel diskutierten niedrigen Verwendungshäufigkeit zu sehen, die darin besteht, dass keine bzw. nur vereinzelte Belege ausgemacht werden können.

Es ist auch nicht das Ziehen einer absoluten Grenze notwendig; statt dessen ergibt sich die „relative" Bedeutung der Häufigkeit in Relation zu der Verwendungshäufigkeit anderer Paarformeln. Um eine Vorstellung über die Vorkommenshäufigkeit von Paarformeln im FidaPlus-Korpus zu bekommen, sei z. B. auf die Paarformel (*pobrati*) *šila in kopita* [mit Ahle und Hufen] verwiesen, welche sinngemäß als *sich mit all seinen Sachen auf die Socken machen* übersetzt werden kann. In jedem Fall ist diese Paarformel im Korpus mit einer relativ hohen Verwendungshäufigkeit (337) nachzuweisen, und es ist auch kein einziger Beleg mit einer geänderten Reihenfolge (d. h. also *kopita in šila*) zu beobachten.

Damit kann wiederum zu der eingangs gestellten Frage nach der Reversibilität von Komponenten innerhalb von Paarformeln und deren Vorkommenshäufigkeit zurückgekehrt werden: Zu beginnen ist mit der Paarformel *jok in smeh* [Weinen und Gelächter] bzw. in umgekehrter Reihenfolge *smeh in jok*, die offenbar in beiden Varianten im Gebrauch zu sein scheint. Zumindest deuten die entsprechenden Befunde aus dem Korpus darauf hin, wonach – im Grunde genommen – keine der beiden Varianten als präferiert bezeichnet werden kann, denn *jok in smeh* lässt sich 22-mal belegen, während *smeh in jok* 38-mal vorkommt. Somit liefert die Verwendungshäufigkeit in diesem Fall keine eindeutige Entscheidungshilfe, um eine bestimmte synchrone Tendenz hinsichtlich der präferierten Verwendung der einen oder anderen Variante ablesen zu können.

In einer Vielzahl von Paarformeln, für die auf paradigmatischer Ebene (d. h. dem Lexikon) eine Reversibilität zugelassen ist, lässt sich aber festhalten, dass eine Korpus-Analyse zumindest das Erkennen bestimmter Tendenzen zulässt. Es lässt sich eine eindeutige Präferenz für die eine oder andere Form erkennen, wie aus dem folgenden Belegmaterial abzulesen ist. Die Paarformel *podnevi in ponoči* [tags und nachts] (Keber 2011: 716) – im Sinne von ‘andauernd, fortlaufend’ hat eine relativ hohe Verwendungshäufigkeit (613), während die umgekehrte Form – die von Keber (2011) übrigens nicht angeführt wird – eine weitaus geringere Verwendungshäufigkeit (130) hat. Eine ähnliche Tendenz lässt sich für *med in mleko* [Honig und Milch] bzw. *mleko in med* feststellen. Hier kann eine eindeutige Präferenz für die Variante *med in mleko* festgestellt werden (Vorkommenshäufigkeit 503), während die umgekehrte Form mit 125 Belegen weitaus seltener vorkommt.

Ein weiteres, abschließendes Beispiel für eine relativ eindeutige Verteilung einer bestimmten Reihenfolge der Komponenten ist die Paarformel *meso in kri* [Fleisch und Blut], welche nach Keber (2011: 518) auch in der Reihenfolge *kri in meso* verwendet werden kann. Allerdings ist hierbei zu beachten, dass die Verwendung dieser Komponenten je nach Kontext und insbesondere in Abhängigkeit von dem mit der Paarformel einhergehenden Verb unterschiedliche Bedeutungen[7] aufweisen kann:

[7] Nicht berücksichtigt wird dabei die Form *biti iz mesa in krvi*, da in diesem Fall eine Komponente ausschließlich im Genitiv (*mesa*) verwendet wird und somit von morphologischer Varianz auszugehen ist.

(a) *postati meso in kri* – d. h., etwas wird zu Fleisch und Blut, im Sinne von 'sich verwirklichen',

(b) *preiti komu v meso in kri* – in Fleisch und Blut übergehen, d. h. 'etwas wird jemandem zur Gewohnheit',

(c) und *meso in kri* – in der Regel verbunden mit dem Verbum *biti* [sein] und in der Bedeutung 'der Mensch mit all seinen Schwächen'.

In diesem Fall zeigt die Verwendungshäufigkeit der Paarformeln insgesamt eine eindeutige Präferenz für die Variante *meso in kri* (Vorkommenshäufigkeit 109), während *kri in meso* nicht mehr als 19-mal vorkommt. Zu beachten ist aber, dass auch bei dieser Paarformel in einigen Fällen eine eindeutige idiomatische Bedeutung überhaupt nicht gegeben ist, wenn es sich z. B. um den Titel eines Filmes, eines Buches oder ähnliches handelt. Insgesamt gilt aber, dass die Paarformel *meso in kri* als die stabile Form der Reihenfolge anzusehen ist.

Ein ähnlicher Befund gilt auch für eine Reihe weiterer Paarformeln, in denen eine jeweils bestimmte Reihenfolge der Komponenten eindeutig präferiert wird. Eine Auswahl derartiger Paarformeln findet sich zusammengefasst in der Tabelle 1, in der die jeweilige Vorkommenshäufigkeit im FidaPlus-Korpus (August 2012) verzeichnet ist. In einigen Fällen ist eine reversible Form überhaupt nicht belegt, so dass in diesen Fällen tatsächlich von irreversiblen Paarformeln gesprochen werden kann. Die jeweils präferierte Form ist mit Stern markiert.

Paarformel	Häufigkeit (abs.)	Paarformel	Häufigkeit (abs.)
brez repa in glave	139*	*prah in pepel*	127*
brez glave in repa	9	*pepel in prah*	11
ne repa ne glave	62*	*pes in mačka*	175*
ne glave ne repa	7	*pes in maček*	11
živ in zdrav	142*	*ne duha ne sluha*	824*
zdrav in živ	8	*ne sluha ne duha*	21

Tabelle 1: Häufigkeit von reversiblen Paarformeln im Fida-Plus-Korpus.

Abschließend ist noch auf einen Fall einzugehen, der insbesondere aus Sicht der morphologischen Diversifikation von Interesse ist. Es handelt sich hierbei um die insgesamt hochfrequente Paarformel *od glave do pet* [von Kopf bis Ferse], die in erster Linie hinsichtlich der Komponente *peta* [Ferse] eine auffallende morphologische Variabilität[8] aufweist, wie der Tabelle 2 entnommen werden kann. Es kann im Falle von *peta* sowohl der Akk. Pl. (*pete*) als auch der Gen. Sg. (*pete*) bzw. Gen. Pl (*pet*) verwendet werden. Vgl. dazu Tabelle 2 mit den entsprechenden Angaben zur Vorkommenshäufigkeit der entsprechenden Varianten. Trotz der hohen Variabilität innerhalb einer bestimmten Reihenfolge der Komponenten kristallisiert sich heraus, dass mit der möglichen Umkehrung der Komponenten gleichzeitig eine hohe morphologische Variabilität einhergeht.

Paarformel	Häufigkeit (abs.)	Paarformel	Häufigkeit (abs.)
od glave do pete	335*	*od pet do glave*	37
od glave do peta	322*	*od pete do glave*	17
od glave do pet	261*	*od peta do glave*	28
Gesamt	918		82

Tabelle 2: Vorkommenshäufigkeit von *glava-peta*-Paarformeln.

Allerdings ist festzuhalten, dass sich auch hier eine eindeutige Präferenz für eine Erststellung von *glava* zu beobachten ist. Mit anderen Worten: Bei zukünftigen Untersuchungen wäre zu prüfen, ob eventuell mit der hohen Verwendungshäufigkeit einer bestimmten Reihenfolge gleichzeitig auch eine hohe Varianz auf anderen sprachlichen Ebenen (morphologische, lexikalische Varianten usw.) einhergeht.

3 Zusammenfassung

Im vorliegenden Beitrag wurde die Frage der Reversibilität von Komponenten in Paarformeln exemplarisch anhand von einigen Beispielen aus dem phraseologischen Wörterbuch von Keber (2011) diskutiert. Als wichtiger Befund lässt sich festhalten, dass in vielen Fällen ein Vertauschen der Komponenten nicht unbedingt – wie zuweilen postuliert – zu einem Verlust der Formelhaftigkeit

[8] Ähnliches gilt für die Paarformeln *ločiti pleve od zrnja* bzw. *zrnje od pleva* '*die Spreu vom Weizen trennen*', wörtlich [die Spelze vom Korn] – in diesem Fall können beide Komponenten in jeweils unterschiedlichen morphologischen Formen verwendet werden (*zrnje*, *zrno*, *zrna* bzw. *plev*, *pleva*). Eine weitere Art der Varianz bezieht sich auf die Ersetzung der Komponenten durch *plevel* [Unkraut] bzw. *žito* [Getreide].

bzw. zur Ungrammatikalität führt. Offenbar kann die Reversibilität von Komponenten sogar als ein integrales Kennzeichen von Paarformeln gelten, wie anhand einer Vielzahl von Beispielen belegt werden konnte.

Der zweite Befund ist, dass die Analyse der Vorkommenshäufigkeit von Paarformeln eine produktive Möglichkeit ist, eine jeweils präferierte Form der Reihenfolge von Komponenten zu extrahieren. In diesem Sinne kann die Vorkommenshäufigkeit – unter Berücksichtigung aller Faktoren, die in Zusammenhang mit den jeweils verwendeten Korpora stehen – als ein grober Maßstab für den Grad an Variabilität herangezogen werden. In Bezug auf die aus dem Wörterbuch von Keber (2011) extrahierten Paarformeln, für die eine Umkehrung von Komponenten angeführt ist, lässt sich abschließend festhalten: Aufgrund der Vorkommenshäufigkeit lässt sich eine Gruppe von Paarformeln extrahieren, die sich durch eine mehr oder weniger archaische Lexik auszeichnet und die insgesamt eine relativ geringe (z. T. gegen Null tendierende) synchrone Verwendungshäufigkeit aufweist. In diesem Sinne handelt es sich dabei um wertvolles lexikographisches Material, welches aber für eine weitere Erforschung z. B. der Variabilität von Komponenten nur bedingt verwendbar wäre.

Eine zweite Gruppe, die sich extrahieren lässt, beinhaltet jene Paarformeln, in denen sich eine in der lexikographischen Praxis angeführte Variabilität insgesamt auch empirisch in synchronen Korpora des Slowenischen bestätigen lässt. Darüber hinaus lässt sich aufgrund der Vorkommenshäufigkeit in vielen Fällen eine relativ eindeutige Präferenz für die eine oder andere Form der Reihenfolge ableiten. In diesem Sinne ist dieser Befund nicht als eigentliches Resultat zu bezeichnen, sondern als Ausgangspunkt für weiterführende Untersuchungen zu den Gründen und Motiven für eine bestimmte Art der Reihung.

Literatur

ARHAR, Špela / GORJANC, Vojko, 2007: Korpus FidaPlus: Nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2, 95–110.

BENOR, Sarah Bunin / LEVY, Roger, 2006: The Chicken or the Egg? A Probabilistic Analysis of English Binomials. *Language* 82/2, 233–278.

BOLINGER, Dwight, 1962: Binomials and pitch accent. *Lingua* 11, 34–44.

BURGER, Harald / BUHOFER, Annelies / SIALM, Ambros 1982: *Handbuch der Phraseologie.* Berlin/New York: de Gruyter.

BUSSMANN, Hadumod (Hg.), 1990: *Lexikon der Sprachwissenschaft.* Zweite, völlig neue bearbeitete Auflage. Stuttgart: Kröner.

BYBEE, Joan / HOPPER, Paul, 2001: Introduction to frequency and the emergence of linguistic structure. Bybee, Joan / Hopper, Paul (Hg.): *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: Benjamins. 1–24.

COOPER, William E. / ROSS, John R., 1975: World order. Grossman, Robin / San, James / Vance, Timothy (Hg.): *Papers from the parasession on Functionalism*. Chicago: Chicago Linguistic Society. 63–111.

FENK-OCZLON, Gertraud, 1989: Word frequency and word order in freezes. *Linguistics* 27, 517–556.

KEBER, Janez, 2000: Raziskovanje slovenske frazeologije – sedanje stanje in zasnova frazeološkega slovarja. *Jezikovni zapiski* 6, 81–116.

KEBER, Janez, 2011: *Slovar slovenskih frazemov*. Ljubljana: ZRC SAZU.

Korpus slovenskega jezika FidaPLUS: <http://www.fidaplus.net.>. Zugriff 28. 11. 2012.

KRŽIŠNIK, Erika, 2004: Poskusni zvezek slovenskega frazeološkega slovarja. *Slavistična revija* 52/2, 199–208.

LAMBRECHT, Knud, 1984: Formulaicity, Frame Semantics, and Pragmatics in German Binomial Expressions. *Language* 60/4, 753–796.

LANDSBERG, Marge E., 1995: Semantic constraints on phonologically independent freezes. Landsberg, Marge E. (Hg.): *Syntactic Iconicity and Linguistic Freezes: The Human Dimension*. Berlin/New York: de Gruyter, 65–78.

MALKIEL, Yakov, 1959: Studies in irreversible binomials. *Lingua* 8, 112–160.

MÜLLER, Gereon, 1997: Beschränkungen für Binomialverbindungen Deutschen. *Zeitschrift für Sprachwissenschaft* 16, 1/2, 5–51.

ROJS, Jurij, 2012: Janez Keber, Slovar slovenskih frazemov, Ljubljana: Založba ZRC, ZRC SAZU, 2011. *Jezikoslovni zapiski* 18/1, 209–212.

SOUTHERN, Mark R. V., 2000: Formulaic Binomials, Morphosymbolism, and Behaghel's Law: The Grammatical Status of Expressive Iconicity. *American Journal of Germanic Linguistics and Literatures* 12/2, 251–279.

TOPORIŠIČ, Jože, 1996: Dvojčiči in podobne frazeološke zgradbe v slovenščini. *Slavistična revija* 44/3, 269–278.

TOPORIŠIČ, Jože, 1998: Zwillings- und ähnliche Strukturen im Slowenischen. Eismann, Wolfgang (Hg.): *Europhras 95: Europäische Phraseologie im Vergleich: Gemeinsames Erbe und kulturelle Vielfalt*. Bochum: Brockmeyer. 795–806.

WÄLCHLI, Bernhard, 2009: *Co-compounds and natural coordination*. 1. Aufl. Oxford: Oxford Univ. Press.

From Dictionary to Corpus[1]

Marie Kopřivová, Milena Hnátková (Prague)

Abstract

In our paper, we would like to introduce “the path of phrasemes” from the dictionary into the Czech corpus using *Slovník české frazeologie a idiomatiky* (Dictionary of Czech Phraseology and Idiomatics, hereinafter SCFI) and the large synchronous written corpus SYN2010 as an example. It is a way of converting phraseme entries that appear in a paper dictionary in their basic or most common form to such a form that it enables to identify marking the corresponding collocations in corpus texts. The software for automatically discovering collocations in corpus texts based on the SCFI works on disambiguated morphologically tagged texts. The idioms are listed in tables which are then used by the software. A linguistic analysis determining specific conditions must be entered for individual idioms. The automatically identified collocations are marked and can then be searched for with the help of the corpus concordancer Bonito.

1 Introduction

Marking phrasemes and collocations in corpora of the Czech National Corpus is done using a program for the automated discovery of collocations, which was developed and is constantly being improved by Hnátková (2011: 171). This program was created on the basis of the SCFI. The fourth volume of this phraseological dictionary contains sentential expressions expanding the series started by the previous three volumes, which focused on non-verbal expressions, similes and verbal expressions. In this paper, we focus on identifying propositional phrasemes, therefore we will first briefly introduce the fourth volume of that dictionary. Secondly we will present the program itself and the process of identifying phrasemes in the corpus. Using specific examples, we will demonstrate the variability of propositional phrasemes. Phrasemes were found and identified in the SYN2010 corpus containing one hundred million entries, which include fiction, journalism and professional texts. Labeled phrasemes can be searched for, just like anything else, using the corpus manager *Bonito* (Rychlý 2007), which is used by the Czech National Corpus.

[1] The study was conducted within the framework of the Charles University Programme of Development of Scientific Disciplines (PRVOUK).

2 Dictionary of Czech Phraseology and Idiomatics

The Dictionary of Czech Phraseology and Idiomatics consists now of four books: Similes (1983), Non-verbal Expressions (1988), Verbal Expressions (1994) and Sentential Expressions (2009). The whole dictionary has a unitary concept and was created under the leadership of F. Čermák. In Čermák's conception of phraseology and idiomatics the emphasis is placed on the anomalous nature of formal or semantic components and consistency of phrasemes (Čermák 2009: 9).

Phraseological data from various dictionaries are gradually incorporated into the program FRANTA (*Phraseology Annotation and Text analysis*), to be used also later on for similar tasks. In this paper, we focus on propositional phrasemes, which are in the 4th volume of the SCFI. It contains a total of 10,056 entries, of which 2 769 are interpretative, 6 188 reduced and 1 066 linked. The dictionary is supplemented by a semantic index.

The selection of phrasemes for the SCFI was limited to contemporary language, excerption of the textual resources thus focused on the 2nd half of the 20th century, taking into account frequency and colloquialisms. The basis for this was the authors' phraseological file, which drew on a variety of sources such as song lyrics, poems, popular literary works, spoken language, plays, movies and existing written sources such as collections of proverbs. The database was closed in 2004 and was then verified against the corpus of Czech existing at the time, which, however, included only written language, while many propositional phrasemes are mainly used in spoken language. The recorded propositional phrasemes are of different kinds, including quotations, proverbs, weather lores, but also various other types that do not have fixed names, such as consistently and frequently used expressions of surprise, disapproval, etc.

In all full entries of the dictionary, information about grammatical use is provided (e. g. in which persons, tenses and moods the verb, which is part of the phraseme, is not used; in the case of proverbs, the lemma has a fixed structure), information on intonation; description of the prototypical situation in which the phraseme is used (who says what to whom, what do they do to each other, when, where, how and why it is used); explanation of the meaning, example of use, possible origin of the phraseme (e. g. a quote, where it comes from); link to semantic register; some of the entries contain also type labels, such as proverbs. At the end of the entry equivalent phrasemes in English, German, French and Russian are usually given. Reduced entries contain only a description of the communication situation, an explanation of the meaning, the origin of the phraseme and a link to the semantic register.

Full entry example:

Dobrá rada nad zlato. (literally: Good advice is beyond gold)

(kol – neutr) **0** neměnné ~ 1c kles. ozn. (od 3. slova)

(*Čl. uznale vůči druhému a jeho nápadu pro řešení dané situace, problému ap. n. bezradně toužící po podobném nápadu v těžké situaci ap.*:) *konstruktivní, použitelný návrh, nápad řešení problému jsou potřebné a velmi vítané.*

A: Radši si kupte prací prášek do zásoby, četla jsem, že se to má zdražovat. – B: Díky. D. r. nad z.

S rada užitečná je cenná A teď, babo raď!

Přísloví.

A (Good advice is beyond all price.) Good counsel has no price.

N Guter Rat ist Goldes wert.

F (Bon conseil vaut mieux que de l'or.), (Bon conseil est plus précieux que l'or.)

R Совет дороже золота.

3 Variations of phrasemes

When searching in the corpus, one of the problems consists in the possible variations of phrasemes (Nováková 1998), some to be found directly in the dictionary, others encountered in a particular text, for example in journalism, where a specific new version may be found. Identifying and labeling phrasemes in corpora also helps us to see how each version of the phraseme is used, we can determine which one is the most common and in which types of texts the individual variations occur. Marking phrasemes also helps to improve disambiguation in the context of the morphological analysis of corpora.

Some variations are presented in the following examples.

3.1 Grammatical variants: morphological variants, word order

Abys to ***nezakřik****. – Abys to* ***nezakřiknul****.* (Don't speak too soon; literally: Not to shout it down)

být na hlavu ***padlý*** *– být* ***padlej*** *na hlavu* (What/who does he take me for?; literally: to be fallen on one's own head)

Mráz kopřivu *nespálí. –* ***Kopřivu mráz*** *nespálí.* (He's got some nerve; literally: The frost does not burn the nettle)

3.2 Lexical replacement, insertion, deletion

> ***Nic** si na něm **nevemeš**. – **Co** si na něm **vemeš**?* (You can't take anything from him; What can you take from him?)
>
> *Nic mu **do toho** není – Nic mu **po tom** není*. (It's none of his business; literally: Nothing into/after it is for him)
>
> *Pro **korunu/krejcar/groš/halíř/haléř** by si **dal/nechal** koleno vrtat.* (He's a skinflint; literally: He would have his knee drilled for a crown/dime/penny)

3.3 Updating

> *Není na světě člověk ten, **aby se zalíbil lidem všem**.* (There's no pleasing some folk; literally: There is no such person in the world whom everyone would like)
>
> *Není na světě člověk ten, **aby ukradl všechno všem**.* (literally: There is no such person in the world who would steal everything from everyone)

3.4 Transformation

Some entries seem to repeat in different parts of the dictionary. These are transformations (Čermák 2009: 20), whose meaning is similar but not identical, i. e. it is functionally different. This feature makes it more complicated to determine the exact type in a specific text; therefore the basic version is used in search.

> *špička ledovce – **To je** špička ledovce.* (Tip of the iceberg)
>
> *hlava na hlavě – **Je tam** hlava na hlavě.* (You can't move for people; literally: There is a head on a head)
>
> *padlej na hlavu – **bejt** padlej na hlavu/na hlavu padlej – **Nejsem** na hlavu padlej. – **To bych byl/musel** bejt na hlavu padlej (kdybych/abych…)* (What/who does he take me for?; literally: fallen on one's own head (to be, I am not, I would have to be)).

4 Automatic discovery and annotation of phrases and special fixed key phrases based on the dictionary

We shall now briefly introduce the program FRANTA designed to automatically search for and annotate phrasemes in morphologically annotated corpora. We will describe how the idioms which are listed in the phraseological dictionary are incorporated into the software and we will introduce the results of the automatic search for and identification of idiomatic expressions and word collocations in the SYN2010 synchronous corpus.

4.1 The software for automatic annotation for collocations in corpus texts

The FRANTA software for automatic discovery and annotation of collocations in corpus texts (based on the Dictionary of Czech Phraseology and Idioms; Hnátková 2011: 171) is part of a complex procedure of automatic morphological disambiguation and works on morphologically disambiguated texts. The idioms are listed in tables which the software then accesses. A linguistic analysis determining specific conditions must be entered for the individual idioms. A program for searching for discontinuous phrases allows you to specify morphological information for each word of the combinations studied, it allows specifying a variable as a lexical unit and it allows determining a change in the word order. It is determined whether one deals with a continuous or discontinuous combination of words, i. e. the program allows to mark positions in a sentence where any words may be located that are not part of idioms, and, conversely, the software allows to specify that in a given position a certain word or part of speech cannot occur.

The automatically identified collocations are marked and then they can be searched for with the help of the corpus concordancer Bonito using the special *kolok* (collocation) attribute and the collocation type as specified by the 16th and 17th position of the morphological tag: the letter K indicates components of nonverbal collocations, the letter V indicates word of verbal collocations, the letter P indicates word of similes, the letter M indicates word of proverbs, the letter S indicates word of sentence expressions.

Table 1 shows examples of phrases labelled in the corpus:

16th position	**17th position**	**meaning**	**example**
K	Z/H*	word/main word* of **nonverbal collocation**	Dopadení je jen otázkou-KZ času-KH. Capture is only a matter of time.
V	Z/H*	word/main word* of **verbal collocation**	Nenecháme-VZ lidi na-VZ holičkách-VH. We won’t leave people in the lurch.
P	Z/H*	word/main word* of **simile**	Byl-PZ tvrdý-PZ jako-PZ kámen-PH. He was hard as a rock.
M	Z/H*	word/main word* of **proverb**	Bez-MZ práce-MZ nejsou-MZ koláče-MH. No pain no gain.
S	Z/H*	word/main word* of **sentential expression**	Tudy-SZ cesta-SZ nevede-SH. This way is not the way.

Table 1: Marking phrases in the corpus (the * marking the main word of a collocation is there only for technical reasons: to evaluate the frequency of the combination of words in Bonito).

The FRANTA software is a program for the automated discovery and annotation of phrases and special fixed key-phrases in digitized written texts, based on the first three volumes of the Dictionary of Czech Phraseology and Idioms. Based on the first volume of the dictionary, *Similes* (SCFI1), the program indicates similes in the text, covering a total of 4 842 basic variants of similes. The second volume *Non-verbal Expressions* (SCFI2) is used by the program for identifying nonverbal idioms and collocations, i. e. nominal phrases, adverbs, particles, composite conjunctions and compound prepositional phrases; in sum, the program includes a total of 6 744 basic variants of nonverbal collocations. Verbal idioms are automatically annotated on the basis of the third volume *Verbal Expressions* (SCFI3); the program includes a total of 15 769 basic variants of verbal collocations.

Originally, occurrences of sentence idioms in the corpora of the *Czech National Corpus* (a family of large electronic corpora of mainly written Czech) were used as the source of sentential expressions, non-systematically supplemented by some proverbs and sentential idioms.

However, using the fourth volume of the Dictionary of Czech Phraseology and Idioms – *Sentential Expressions* (SCFI4) made it possible to expand existing tables for the discovery and annotation mainly of idioms including proverbs, and, to a lesser extent, of sentential similes and of other verbal idioms.

After incorporation of the data from the fourth volume of the phraseological dictionary, the program now contains a total of 978 basic variants of proverbs and 3 692 basic variants of sentential annotation.

4.2 Processing dictionary entries of the *Dictionary of Czech Phraseology and Idioms* (SCFI4)

We shall now describe the procedure of processing the data in the dictionary (SCFI4) and the procedure for expanding the FRANTA program.

4.2.1 Automatically generated list of variants of sentential idioms

A list of all variants of sentential idioms (expressions) in the dictionary (called SENTEX, as in SENTential EXpressions) was automatically created from the list of dictionary entries. The dictionary contains a total of 10 056 entries. After the variants had been generated, a total of 17 456 variants of sentential phrases was available.

The list of idioms (SENTEX) was automatically morphologically tagged and enabled identification of all occurrences of the verbal and nonverbal idioms, similes and proverbs that have so far been automatically identified by the program for automatic discovery and annotation of collocations. Using frequency in the corpus texts, all unmarked sentential idioms from the SENTEX list were then added to automatic annotation program. After this addition, the program could serve for the annotation of specific entries in the dictionary, but also their variations, for example variants in negation, person, gender, number, tense and word order.

The SCFI4 dictionary contains similes, proverbs, quotes, etc. On the other hand, a large part of these idioms seem to be just particular uses of general verbal collocations that are already included in the automatic phrase discovery program (they were mentioned in the preceding volumes of the dictionary). Therefore, the parts of sentential idioms in SENTEX were automatically identified as general nonverbal or verbal idioms, and, accordingly, they were also annotated as similes and proverbs.

After incorporation of the sentential expressions from the SCFI4 dictionary into the FRANTA software, a total of 9 209 occurrences of collocations were automatically annotated in the SENTEX list, consisting of 1 396 different verbal idioms, 685 different nonverbal idioms, 374 different variants of similes, 680 different proverbs and 2 946 sentential idioms.

4.2.2 The specific entries in the phraseological dictionaries

We will now give an impression of the specific entries in the phraseological dictionary of sentential expressions:

(1) Occurrence of the pronoun *ten* – it

These are mainly sentential expressions specific to spoken language. As it is typical for spoken language, the substitute pronoun *ten* – it is used in the statements, especially its ambiguous forms: *to*, *tím*, *toho*, *tomu*, which are morphologically homonymous (the ambiguity mainly concerns case and gender):

to	it neuter sing. nominative, neuter sing. accusative
tím	it neuter sing. instr., masculine anim. sing. instr., masculine inanim. sing. instr.
toho	it masculine anim. sing. accusative, masculine anim. sing. genitive, masculine inanim. sing. genitive, neuter sing. genitive
tomu	it masculine inanim. sing. dative, masculine animate sing. dative, neuter sing. dative

Of all the variants in the automatically generated SENTEX list that contain some of these forms of the pronoun *ten*, about 5716 entries, i.e. one third of all entries, have been recorded to belong here. This is very difficult for any automatic morphological disambiguation system. Furthermore, it is difficult to automatically determine in which cases a general (substitute, determinative) pronoun *to* – it simply replacing object, property, or event may be found – see example (1), and in which cases it has a specific phraseological function – see example (2).

Example (1)

In the SCFI4 dictionary the following item is given:

To vyjde nastejno [It comes to the same thing, It makes no odds].

In the corpus one will find specific variations of this idiom:

Obojí vyjde nastejno [Both things come out the same];

Oba kandidáti vyjdou nastejno [Both candidates come out the same].

Example (2)

There is also an alternative way of expressing the same thing found in the sentential expression:

Co to povídám! [What am I talking about!];

Kde jsme to přestali? [Where were we?].

In this case, however, you cannot replace the pronoun *to* with another word.

Especially in written texts, the given thing is made more specific; in dialogues, the context helps to make it clear what *to* refers *to*. It is interesting to see that the pronoun *to* – *it* is not frequently used in proverbs, in quotations or in snippets of songs.

(2) Occurrence of sentential idioms such as <sentential expressions> <comma> that / but / so / if / when …

This is actually the valence of the main verb in the form of the subject phrase, where the collocation requires a specific form of continuation and textual complementation. About 600 of these variants are registered in the SENTEX list, variants including the conjunction *že* – that add up to about three hundred. A total of 19675 occurrences of these sentential expressions with the conjunction *že* were annotated in the SYN2010 corpus, with 138 different variants.

The most frequent collocations of this type in the SYN2010 corpus are with the verb *říci* – to say:

Example (3)

> ***Řekl bych, že*** … [I would say that …] (2 027 occurrences in the SYN2010 corpus)
>
> ***Nedá se říct, že*** … [It cannot be said that …] (1 706 occurrences in the SYN2010 corpus)

In the case of sentential connecting expressions ***řekl bych, že …*** [I would say that …] there is only a typical collocation found without any metaphor.

Conditional forms are actually the most common type of phrase, such as *I would do that ...*, followed by the expression ***přísahal bych, že …*** [I could swear, that…]; only 68 occurrences in the SYN2010 corpus.

(3) Occurrence of sentential idioms such as *It is …*

Some parts of sentential idioms (especially nominal phrases) in the SENTEX list are automatically annotated as nonverbal idioms or similes, in variants of idioms such as:

> *Je/Bylo to* (nonverbal idiom) [Is/Was it …]
>
> *To je/bylo* (verbal idiom) [It is/was …]
>
> *Je to jako* (noun or nominal phrase) [It is like …]

The parts are annotated as nonverbal phrasemes and similes, or as a type of verbal phrasemes such as: to be <noun/nominal phrase>. So this case represents a particular use of the verbal idioms or the similes (there are about 450 variants of this type in the SENTEX list).

Examples (4)

> ***Byla to láska na první pohled*** (nonverbal idiom: love at first sight) [It was love at first sight]
>
> ***Je to živá kronika*** (nonverbal idiom: a walking memory-bank) [He's a walking memory-bank]
>
> ***Je to k pláči*** (verbal idiom: to be fit to make one weep) [(The sight of) it would make you weep]
>
> ***To je jako zlý sen*** (similes: to be like a nightmare) [It's a nightmare!]

(4) Latin or English fixed terms, proverbs and quotes

Latin or English fixed terms, proverbs and quotes, included in the dictionary, are easily automatically identified in the text.

Examples (5)

never more, memento mori, Veni, vidi, vici, In vino veritas.

(5) One-word sentence idioms

One-word sentential phrasemes listed in the dictionary (specific to spoken language) are not identified by the program for automatic discovery and annotation of idioms, because the program is designed to search for and identify multiword collocations only; in order to identify the meanings of individual words as defined in the dictionary, a semantic analysis of the whole utterance would be necessary. However, the question whether these are phrasemes is not relevant, as these are only variants of larger combinations.

Examples (6)

Dost! [Enough!]
Jasan! [Sure! / Of course!]
Jasně. [Go ahead! / Sure! / But of course!]
Nashle! [See you (then)]
Padla! [It's knocking-off time]
Platí! [You're on! / Agreed!]
Čest! [Hello / Good morning]

4.2.3 Results of the annotation of the SYN2010 corpus

Only the stable combinations that occur in SYN2010 and whose automatic identification is possible were incorporated into the FRANTA program for the discovery and annotation of phrasemes. Thus the occurrence of the string of words not in the text in a non-idiomatic meaning is excluded. A total of 1 218 585 occurrences of collocations were automatically identified by the automatic discovery and annotation of collocations during the automatic morphological disambiguation of the SYN2010 corpus of one hundred and twenty million words (121 666 531 positions in total). A total of 3 104 250 words were identified as part of automatically labeled collocations (cca 2.5 %).

Table 2 shows frequent idioms found automatically, the number of occurrences of different variants and the most common example.

SYN2010	number of different variants	the most common example
nonverbal phrases	5 572	ve skutečnosti in reality
verbal phrases	11 139	mít možnost have the opportunity
proverbs	1 900	Účel světí prostředky. The end justifies the means.
similes	620	jako dříve as previously
sentential idioms	3 021	To není možné. It is not possible.

Table 2: Frequent idioms found automatically.

4.2.4 Conclusions based on the results of automatic annotation of idioms and collocations

Based on the evaluation of the results of automatic annotation of stable collocations in the SYN2010 corpus, we can study the most frequent variants of sentential expressions in the corpus data (while the corpus consists of written texts, the dictionary was based mainly on the spoken language). This has been made possible by comparing the results of the corpus annotation with the entries in the SCFI4 dictionary. Here are some examples:

A) Examples indicating whether the variant of a sentential or verbal phrase included in the dictionary is or is not the most frequently identified variant of the idiom in a corpus. The colloquial variant of the sentence expression *Jak se to veme* [It all depends] occurs only once in the corpus; the more formal variant *Jak se to vezme* is much more frequent.

Examples (7)

Jak se to veme / vezme. [It all depends.]

Jak se to veme. (*colloquial form*), 1 occurrence in corpus

Jak se to vezme. 85 occurrences in corpus

In the dictionary, only the colloquial forms of the questions *Vo co de?* [What is it?] or *Vo co kráčí?* [(All right,) what's going on then?] are presented, while in the corpus, the standard versions *O co jde?* and *O co kráčí?* are more frequent.

Examples (8)

Vo co de? (*colloquial form*) [What is it?]

Vo co de?! 16 occurrences in corpus

O co jde? 131 occurrences in corpus

Vo co kráčí? (*colloquial form*) [(All right,) what's going on then?]

Vo co kráčí? 1 occurrence in corpus

O co kráčí? 5 occurrences in corpus

B) Examples confirming that the variant listed in the dictionary is not the most frequent variant in the corpus. In some cases, a particular dictionary variant does occur in the corpus, but there are other variants with the same meaning. Thus, some variants listed in the dictionary may not reflect the true frequency in written texts. Examples show that the word order variant of the idiom *Fakt je, že* ... [The fact is, that …] listed in the dictionary is not the most frequent variant in the corpus; in fact, the variant with the opposite word order occurs significantly more often in the corpus.

Examples (9)

Fakt je, že … [The fact (of the matter) is, (that) …]

Fakt je, že … 163 occurrences in corpus

Je fakt, že … (*change in word order*) 237 occurrences in corpus

C) Examples of additional sentential phrases which do not occur in the dictionary:

The dictionary gives the sentential idiom: *Hladina se uklidnila* [The water surface calmed down], but the frequent statement *Situace se uklidnila* [The situation calmed down] is not listed in the dictionary as a separate item. The sentential idiom *Uhni mi z vlny* [Get out of my wave] does not appear in the corpus, but *Uhni mi z cesty!* [Out of my way!], which is frequent, is not in the dictionary.

Examples (10)

SCFI4: ***Hladina se uklidnila.*** [The water surface calmed down]

SYN2010: ***Situace se uklidnila.*** [The situation calmed down]
SCFI4: ***Uhni mi z vlny.*** [Get out of my wave]
SYN2010: ***Uhni mi z cesty.*** [*Out of my way*]

5 Conclusion

There will be more work done on adding idioms into the search program and annotating them in the corporas, eventually including corpora of spoken language. Based on the texts of the corpora, the most frequently used options as well as other less frequent ones can be detected, for inclusion in an eventual web phraseme dictionary. We hope that the designation of idioms in the corpora will help phraseologists in their research.

Bibliography

ČERMÁK, František et al., 1983: *Slovník české frazeologie a idiomatiky 1. Přirovnání*. Praha: Academia.

ČERMÁK, František et al., 1988: *Slovník české frazeologie a idiomatiky 2. Výrazy neslovesné*. Praha: Academia.

ČERMÁK, František et al., 1994: *Slovník české frazeologie a idiomatiky 3. Výrazy slovesné*. Praha: Academia.

ČERMÁK, František et al., 2009: *Slovník české frazeologie a idiomatiky 4. Výrazy větné*. Praha: Leda.

SYN2010, 2010*: Český národní korpus*. Praha: Ústav Českého národního korpusu FF UK. <http://www.korpus.cz>.

HAJIČ, Jan, 2004: *Disambiguation of Rich Inflection (Computational Morphology of Czech). Vol. 1.* Praha: Karolinum Charles University Press.

HNÁTKOVÁ, Milena, 2011: Výsledky automatického vyhledávání frazému v autorských korpusech. Petkevič, V. / Rosen, A. (eds.): *Korpusová lingvistika Praha 2011. Gramatika a značkování korpusů.* Praha: NLN, 171–185.

JELÍNEK, Tomáš, 2008: Nové značkování v Českém národním korpusu. *Naše řeč* 91, 13–20.

NOVÁKOVÁ, Marie, 1998: All That Glitters in the Newspapers Is Not Proverb. Ďurčo, P. (ed.): *Phraseology and Paremiology.* Bratislava: Akadémia PZ. 250–256.

KOPŘIVOVÁ, Marie, 2008: Frazeologie v mluvených korpusech na základě PMK, Kopřivová, M. / Waclawičová, M. (eds): *Čeština v mluveném korpusu* Praha: NLN a ÚČNK. 149–160.

PETKEVIČ, Vladimír, 2006: Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. Šimková, M. (ed): *Insight into the Slovak and Czech Corpus Linguistics*. Bratislava: Veda, 26–44.

RYCHLÝ, Pavel, 2007: Manatee/Bonito – A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masarykova univerzita, 65–70.

Digitalisierung der Phraseologie und der Benutzer-Aspekt

Nataša Kralj (Maribor)

Abstract

Despite phraseology being an established linguistic discipline and phrasemes, which as expressions with figurative meaning significantly mark everyday communication, phraseological phenomena are still rather neglected in language teaching. This is partly due to the lack of specialized phraseological learning materials, which would encourage and enable self-study of phrasemes as well as learning them in language teaching classes. The information technology era provides new possibilities in the domain of phraseology digitalization; practice shows that to varying degrees, the emerging phraseological e-materials exploit the options provided by modern media. This article presents the result of usability aspect research – potential users and their needs, methods of using phraseological e-materials, their polyfunctionality, their applicability and topicality – by an example of phraseological learning materials for German as foreign language (emphasis on EPHRAS learning materials).

1 Einführung

Die Etablierung der Phraseologie als linguistische Disziplin in den letzten Jahrzehnten (u. a. Palm 1997, Burger 2003, Hessky 2007 und Kühn 2007) scheint noch immer zu wenig Einfluss auf die Praxis im Fremdsprachenunterricht zu haben. Die phraseologischen Phänomene werden in der Fremdsprachendidaktik weiterhin mit der Begründung vernachlässigt, dass man auch ohne sie in der Alltagskommunikation gut auskommen kann. Jedoch hängt der Status der Phraseologie in fremdsprachlichen Lernprozessen nicht nur von der Einstellung der Lehrkräfte ab, sondern auch davon, ob und wie die Phraseologie in den Lehr- und Lernmaterialien zum Gegenstand des Spracherwerbs gemacht wird. Dazu ist es zu evidentieren, dass sie z. B. in den DaF-Lehrwerken noch immer unsystematisch thematisiert werden (Anisimova 2012: 1, Jesenšek 2006a: 123).

Hierbei stellt sich die Frage, ob die modernen Medien der Phraseodidaktik und besonders der lernorientierten Phraseographie mit der Ausarbeitung computergestützter phraseologischer Lehr- und Lernmaterialien einen neuen Schwung geben könnten. Um die Frage zu beantworten, wird zunächst ein kurzer Überblick über die Möglichkeiten der Digitalisierung im Bereich der Phraseologie gegeben. Im Vordergrund stehen Benutzer-Aspekte, die anschließend am Beispiel des elektronischen phraseologischen Lehr- und Lernmaterials EPHRAS überprüft und erläutert werden.[1]

2 Digitalisierung der Phraseologie

Die Ära der IKT ermöglicht neue Wege bei der Konzipierung, Entwicklung und Erarbeitung elektronischer phraseologischer Materialien. So entstehen entsprechende Datenbanken und Lernmaterialien, die man autonom und nur computergestützt anwenden kann.[2] Obwohl in der Zwischenzeit schon mehrere Anwendungsprinzipien für die elektronischen phraseologischen Nachschlagewerke geliefert worden sind (Dobrovol'skij 1989, Bergenholtz 2006, Hallsteinsdóttir 2004a und 2004b, Ďurčo 2007, Jesenšek 2009), ist eine Diskrepanz zwischen den theoretischen Erkenntnissen der Phraseologieforschung und der praktischen lexikographischen Arbeit festzustellen. Almind u. a. (2006: 163) sehen den Hauptgrund für diese Divergenz vor allem darin, dass die Wörterbuchfunktionen und die eigentlichen Benutzerbedürfnisse nicht genug Beachtung finden, vgl.:

> The main reason is first of all the lack of incorporating and not taking into account the functions of dictionaries, including users prerequisites, i. e. linguistic and cultural previous knowledge and actual needs.

Aspekte, die man im Rahmen der computerorientierten Beschreibung der Phraseme in Betracht ziehen soll, sind mehrere. Einer davon betrifft die wörterbuchgemäße Metasprache, die den vorausgesetzten Benutzergruppen angepasst werden soll. Die Analyse der Log-Dateien des dänischen elektronischen phraseologischen Wörterbuches (*The Danish Idiom Dictionary*) hat z. B. gezeigt, dass sogar 25 % aller Phrasemsuchen mit Wortverbindungen verbunden sind, die keine Phraseme darstellen, sondern zu Kollokationen oder sogar zu Sprichwörtern gehören (Almind u. a. 2006: 167). Diese Feststellung

[1] Der Beitrag bezieht sich auf Erkenntnisse der empirischen Untersuchung im Rahmen der Dissertation mit dem Titel *Phraseologie und Phrasegraphie in der Zeit der IKT-Ära* (Universität Maribor, Philosophische Fakultät, Mentorin Prof. Dr. Vida Jesenšek).

[2] Vgl. Kralj (2012: 5): „Die elektronische Datenverarbeitung, die Entstehung der korpusbasierten Datenbanken, Online- und Offlineumgebungen, die Entwicklung der hypertextuellen Struktur, Log-Dateien ermöglichen /…/ die Konzipierung neuer, aktueller, zeitgemäßer Phraseologie-Materialien, die an die Bedürfnisse unterschiedlicher Zielgruppen zugeschnitten werden können."

ist ein Indiz dafür, dass für viele Benutzer die Terminologie im erwähnten Nachlagewerk nicht angemessen bzw. nicht auf ihr Benutzerprofil zugeschnitten ist, vgl.:

> In principle, it is a good (and user-friendly) device to avoid using too much lexicographic and linguistic terminology in a dictionary, at least if it targets lay persons, not linguist. (Heid 2011: 298).

Bei der Ausarbeitung elektronischer phraseologischer Wörterbücher für Lernzwecke setzt man sich hauptsächlich mit zwei Problembereichen auseinander: mit der Phrasemauswahl und mit den Anordnungsprinzipien sowie mit der Phrasembeschreibung, vgl. Heid (2011: 289):

> Obviously, the lexicographer has a double task to accomplish, in order to provide adequate lexicographic data: he/she has to select carefully the data that best fit the user's needs, and he/she has to present such data in a convivial, pedagogical and easy-to-use way.

Da der computergestützten Phraseologie und damit auch den elektronischen phraseologischen Wörterbüchern platzmäßig kaum Grenzen gesetzt sind, ist eine ausführliche Phrasembeschreibung mit umfassenden Angaben durchaus realisierbar, mehr noch, denn

> darüber hinaus können phraseologische Datenbanken auch unterschiedliche phraseologische Wörterbuchtypen (z. B. produktive und rezeptive) integrieren und mit den Möglichkeiten der Hypertexttechnologien die Bedürfnisse verschiedener Adressatengruppen optimal befriedigen (Kralj 2012: 23).

Bei elektronischen Nachschlagewerken wird eine optimale Gestaltung der Interaktion zwischen Mensch und Computer angestrebt, die an der sog. Benutzerschnittstelle realisiert wird (Abel 2003: 184). Dafür müssen allerdings verschiedene Benutzer-Aspekte in Betracht gezogen werden.

2.1 Die Benutzerebene in der E-Umgebung

Bei der Konzipierung eines phraseologischen Nachschlagewerkes gilt es als Erstes, potenzielle Benutzerprofile vorzusehen. Die Qualität eines elektronischen Nachschlagewerkes und die entsprechende Relation zu den jeweiligen Benutzern können mit den sog. Usability-Tests und Log-Dateien (engl. *log files*) unter die Lupe genommen werden. Mit der Überprüfung der Effektivität und Effizienz eines Produktes wird die Gebrauchstauglichkeit einer Software oder Hardware seitens potenzieller Benutzer getestet. Bei der Effektivität einer Software geht es um die Frage, ob die richtigen Daten in der richtigen Menge bereitgestellt werden. Die Effizienz misst man mit der verbrauchten Zeit zur Ausführung einer Aufgabe, die sich als typisch für die ausgewählte Applikation herausstellt. Dabei geht es also um das Testen der Zugriffsmodi: je schneller die gesuchten Daten gefunden werden können, desto effizienter ist die Applikation.

Ein Beispiel eines durchgeführten Usability-Tests stellt die Fallstudie von C. Bank vor (Heid 2011: 290), in dem drei Onlinewörterbücher (OWB) (ELDIT[3], BLF[4] und OWID[5]) auf ihre Gebrauchstauglichkeit in einem Usability-Labor überprüft worden sind. Die Ergebnisse des Tests haben gezeigt, dass manche Studenten folgende Eigenschaften der OWB als vorteilhaft angegeben haben: schneller und einfacher Zugriff auf die lexikographischen Daten, einfache Benutzung, aktueller Inhalt und reiches Angebot an lexikographischen Daten. Aus den empirischen Ergebnissen konnte man jedoch eine niedrigere Effizienz bei den komplexen Nachschlagehandlungen, wie z. B. beim Nachschlagen von *c'est une question de vie ou de mort*, feststellen. Die meisten Testpersonen versuchten nämlich, die vollständigen Mehrwortverbindungen direkt in das Suchfeld einzugeben, anstatt den phraseologischen Kern bzw. die Kollokationsbasis zu erkennen. Wenn jedoch das elektronische Werk die Eingaben von Mehrwortverbindungen und die tolerante Suche (mit typografischen Fehlern) zuließ, waren bis zu 80 % der Testpersonen bei der gestellten Aufgabe erfolgreich. Die Integration einer fehlertoleranten Suche nach dem Vorbild von Suchmaschinen verhindert nämlich, dass etliche Lexeme bei der Sucheingabe wegen der orthographischen Fehler nicht gefunden wären (Bergenholtz u. a. 2005: 131).

Ein nützliches lexikographisches Instrument stellen auch die Log-Dateien vor, die uns ermöglichen, die Benutzungsdaten zu sammeln und auszuwerten; es sind z. B. Daten zur Anzahl der Zugriffe auf das jeweilige System, zu den Sucheingaben, zu den eingeschlagenen Lesewegen, zu den abgebrochenen Benutzungssituationen, zu den nicht gefundenen Lemmata, vgl.:

> Undoubtedly, the most obvious way that log files can be used to improve internet dictionaries is a tool to discover lemma lacuna (Bergenholtz u. a. 2005: 136–137).

Ein weiterer wichtiger Aspekt ist die Möglichkeit, die Softwareumgebung zu beeinflussen. Dabei unterscheidet Kemmer (2010: 25) zwei Möglichkeiten der Einwirkung: 1. Die Software-Anpassung an individuelle Benutzerbedürfnisse (Kommunikation zwischen Computer und Benutzer; beliebige Datenauswahl

[3] ELDIT, Elektronisches Lernerwörterbuch Deutsch-Italienisch.
[4] BLF, Base Lexikale du Français.
[5] OWID, Online-Wortschatz-Informationssystem Deutsch.

und Lesewege) und 2. User-Mitarbeit bzw. das Benutzer-Feedback (Kommunikation zwischen Erstellern und Benutzern mittels E-Mail-Kontakt, Foren, Chatrooms oder im Rahmen einer kollaborativen Lexikographie), wodurch die ständige Möglichkeit zur Überarbeitung von Inhalten ermöglicht ist. Ein solches Bespiel stellt das dänische phraseologische Wörterbuch *Ordbogen over faste vendinger* dar (Almind u. a. 2006: 176). Im elektronischen phraseologischen Wörterbuch mit mehr als 8000 Phrasemen steht dem jeweiligen Benutzer im Hinblick auf seine Bedürfnisse und die ausgewählte Suchoption folgendes Informationsangebot zur Verfügung:

- Textrezeption: das Lemma, die Bedeutung;
- Textproduktion: das Lemma, Angaben zum Phrasemtyp und Stil, zur Bedeutung und Grammatik, zu Kollokationen, Beispielen und Synonymen;
- mehr über Phraseme: alle in Verbindung zum ausgewählten Lemma stehenden Informationen (einschließlich Kommentare und Assoziationswörter).

Derartige Anpassungsmöglichkeiten führen zu einer benutzerfreundlichen Entwicklung der elektronischen Materialien und ebenso zur Steigerung der Interaktivität.

2.2 Polyfunktionalität

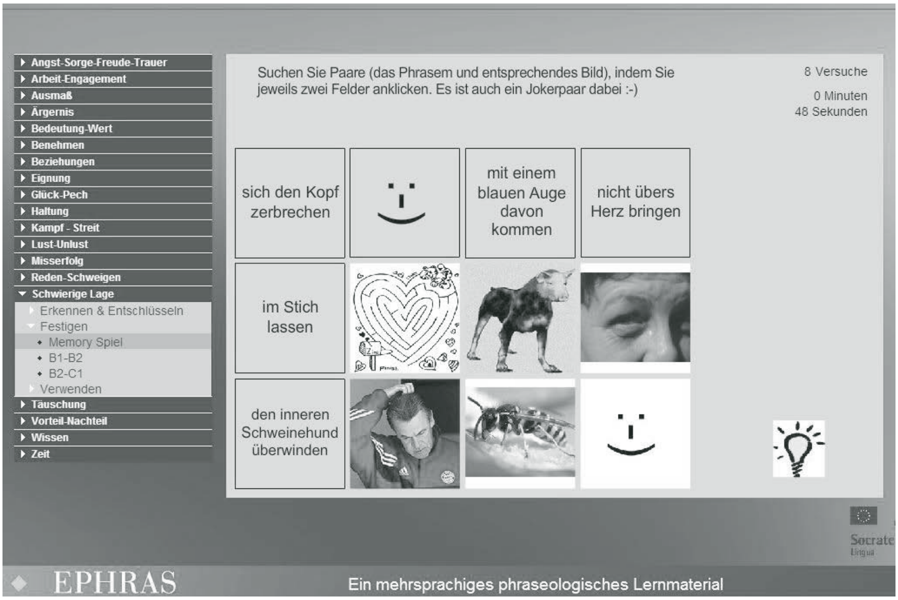
Die nächste wichtige Eigenschaft der elektronischen phraseologischen Materialien ist ihre Polyfunktionalität. Polyfunktionelle bzw. multifunktionelle Nachschlagewerke stellen solche Werke dar, die mehrere Funktionen parallel erfüllen können. Diese Eigenschaft kommt bei elektronischen Materialien umso mehr zum Vorschein, weil die neuen Technologien mit umfangreichen Retrievalmöglichkeiten die Aufbereitung lexikographischer Daten für verschiedene Benutzer und Benutzungssituationen möglich machen (Hallsteinsdóttir 2009: 215; Kemmer 2010: 24). Beim Einloggen wird demzufolge ein benutzerspezifisches Profil kreiert und daraufhin eine entsprechend zugeschnittene Benutzeroberfläche erstellt, damit die lexikographischen Daten nach eventuellen Selektionskriterien schnell zugänglich bzw. einfach abrufbar sind.

Auch im phraseologischen Bereich ist die Tendenz die gleiche. Nach Jesenšek (2006a: 64) erfüllen polyfunktionelle phraseologische Lexika und Lernmaterialien:

> (1) wissensbezogene Funktionen zu Zwecken einer Erweiterung und/oder Weiterentwicklung der fremdsprachlichen Kompetenz /…/ (2) textbezogene Funktionen in rezeptiven und produktiven Benutzungssituationen /…/ einschließlich Situationen der problemorientierten Konsultation bei der Übersetzung.

3 Zum Benutzer-Aspekt am Beispiel von EPHRAS

EPHRAS (2006) ist ein mehrsprachiges elektronisches phraseologisches Lernmaterial. Es besteht aus einer viersprachigen Phrasem-Datenbank (Deutsch als Ausgangssprache, Slowenisch, Slowakisch und Ungarisch) mit 4000 ausführlich linguistisch beschriebenen Phrasemen und einem niveaudifferenzierten und onomasiologisch aufgegliederten Übungsteil (A1–C2) zu 95 ausgewählten Phrasemen (Abbildung 1).



Suchen Sie Paare (das Phrasem und entsprechendes Bild), indem Sie jeweils zwei Felder anklicken. Es ist auch ein Jokerpaar dabei :-)

8 Versuche

0 Minuten
48 Sekunden

sich den Kopf zerbrechen

mit einem blauen Auge davon kommen

nicht übers Herz bringen

im Stich lassen

den inneren Schweinehund überwinden

Abbildung 1: Onomasiologische Anordnung und Struktur der Übungen in der EPHRAS-Datenbank.

Die interaktiven Übungen sind nach dem didaktischen Dreischritt von Peter Kühn (1992: 178) konzipiert und durch die zusätzliche Phase des Festigens erweitert: entdecken – entschlüsseln – festigen – verwenden. Eine ausführliche Beschreibung des EPHRAS-Konzeptes liefert Jesenšek (2008: 111–143). Die zuerst auf einer CD-ROM erschienene Applikation ist in der Zwischenzeit für die potenziellen Benutzer im Internet frei zugänglich.[6]

[6] Erreichbar unter: http://projects.ff.uni-mb.si/frazeologija/static/ephras/.

An der empirischen Untersuchung zu den Benutzer-Aspekten, die im Rahmen meines Promotionsssstudiums zwischen 2008 und 2010 stattfand, nahmen vier Zielgruppen teil: slowenische Schüler, Studenten (mit unterschiedlichen Deutschkenntnissen und Bedürfnissen), Deutschlehrer und Hochschullehrer, die Deutsch als Fremdsprache unterrichten. Insgesamt haben 450 Testpersonen beim Bewerten der Phrasem-Datenbank und des Übungsteils von EPHRAS mitgemacht. Eine ausführliche Darstellung der Untersuchung mit empirischen Ergebnissen liefert Kralj (2012: 75–134); an dieser Stelle wird nur ein kleiner Teil der Resultate vorgestellt.

Das Vorurteil, dass Phraseme nur für fortgeschrittene Lerner reserviert seien, wurde in der empirischen Untersuchung von EPHRAS teilweise widerlegt. Aus den Ergebnissen geht hervor (Abbildung 2), dass auch Anfänger (Lerner auf Niveaustufen A1–A2) durchaus im Stande sind, Phraseme zumindest rezeptiv wahrzunehmen, zumal der modulare Aufbau und die hypertextuelle Struktur von EPHRAS manche Hilfeleistungen anbieten (Kralj 2012: 98).

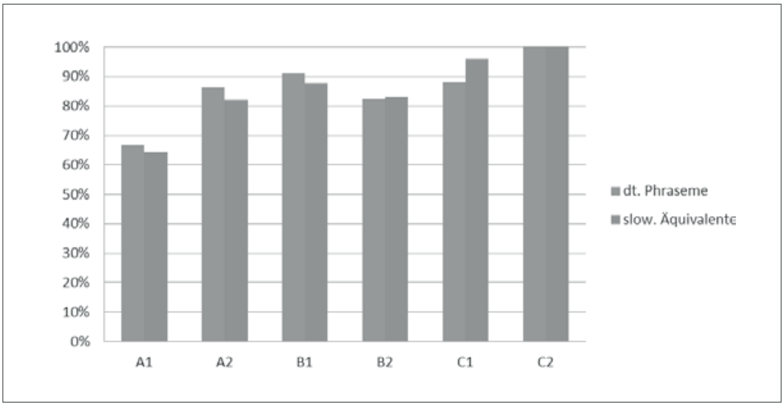


Abbildung 2: Die richtige Angabe der geübten deutschen Phraseme und der slowenischen Äquivalente in Korrelation mit dem Sprachniveau (A1 bis C2).

Alle Phraseme der EPHRAS-Datenbank werden nach einem einheitlichen Beschreibungsmodell ausführlich linguistisch beschrieben. Die Angaben sind aus der Abbildung 3 ersichtlich. Als besonders wichtige Informationsträger (sowohl für Lehrer als auch für Lerner) gelten demnach Angaben zur Bedeutung, Textbeispiele und Äquivalenzangaben. Die Voraussetzung, dass im Rahmen einer präzisen systematischen Beschreibung der Phraseme nicht alle darin gespeicherten Informationen für einen Benutzer von gleicher Bedeutung sind (Dobrovol'skij 1989, 530), hat sich an dieser Stelle zwar bestätigt, jedoch überrascht eine starke Korrelation zwischen den Variablen *Auswahl der Angaben* und *Adressatengruppen* (Lehrer vs. Lerner), weil man von unterschiedlichen Benutzergruppen auch unterschiedliche Ziele und Bedürfnisse erwarten würde.

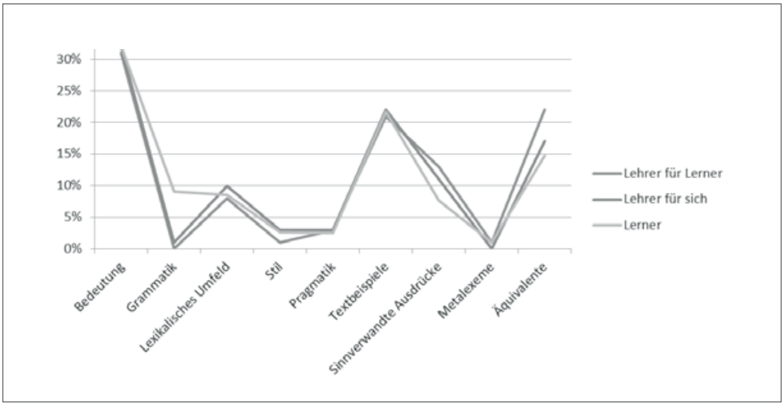


Abbildung 3: Die Auswahl der Angaben zu Phrasemen: Lehrer vs. Lerner.

Bei der empirischen Untersuchung des Übungsteils hatten vor allem Probanden mit geringeren Deutschkenntnissen in der Phase *Erkennen & Entschlüsseln* große Dekodierungsprobleme gehabt. In diesem Lernschritt kommen nämlich nur authentische Phrasem-Kontexte vor, die für viele DaF-Lerner eine harte Nuss vorstellen. Deshalb sollten Lehrer und Lerner bei drei angegebenen Phrasemen (*den inneren Schweinehund überwinden – P1*, *blau machen – P2* und *Bock haben – P3*), die jeweils in vier verschiedenen Kontexten vorkommen, diejenige Textpassage auswählen, die sich als besonders angebracht für die Erschließung des jeweiligen Phrasems anbietet. Der Vergleich der Resultate zeigt, dass wiederum eine starke Korrelation zwischen den Variablen *Benutzergruppe* und *Kontextauswahl* für die drei Phraseme zu sehen ist (Abbildung 4). Das Ergebnis lässt die Schlussfolgerung zu, „dass die Auswahl der authentischen Textbeispiele, in denen Phraseme vorkommen, überlegt erfolgen muss: die Kontexte sollen (sowohl lexikalisch als auch syntaktisch) an die potenziellen Benutzergruppen angepasst werden. Das ist für die Beispiele, die zur Einübung der Phraseologie ausgewählt werden, umso wichtiger, denn eine missglückte Auswahl kann in der Anfangsphase des Lernens mit nicht angemessenen bzw. zu anspruchsvollen Texten bei Lernern eine Abneigung gegenüber der fremdsprachigen Phraseologie verursachen, was zur Folge haben kann, dass der Lernprozess abgebrochen wird“ (Kralj 2012: 144).

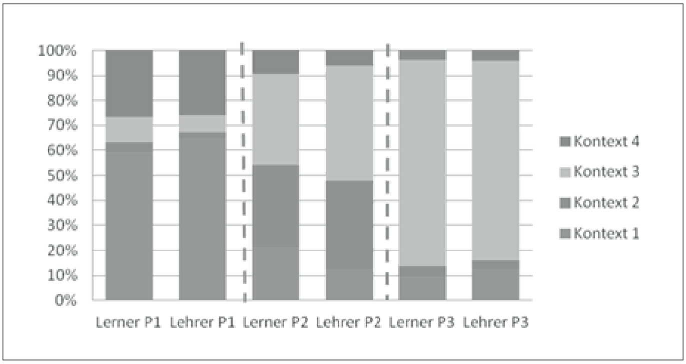


Abbildung 4: Die Auswahl der Phrasem-Kontexte: Lehrer vs. Lerner.

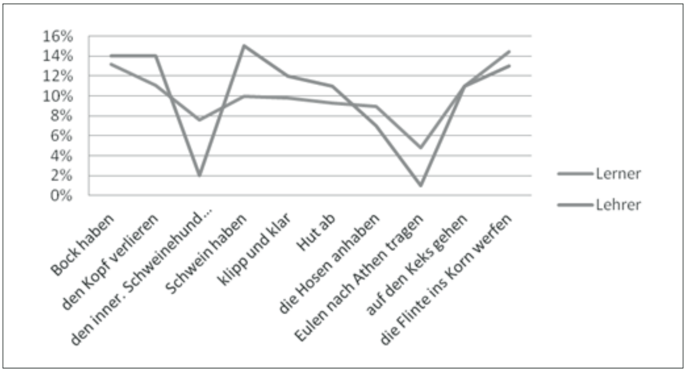


Abbildung 5: Die Auswahl der fünf Phraseme für den Deutsch-Unterricht: Lerner vs. Lehrer.

Da ein phraseologisches Optimum bzw. „ein repräsentativer Ausschnitt aus der Phraseologie einer Sprache für fremdsprachendidaktische Zwecke“ (Jesenšek 2006a: 61) kein endgültiges Phrasem-Verzeichnis darstellen kann, sollten die Testpersonen von zehn angegebenen deutschen Phrasemen fünf Phraseme wählen, die sie lernen bzw. im Unterricht vermitteln würden. Dabei soll erwähnt werden, dass die ersten fünf Phraseme der EPHRAS-Datenbank entstammen, im Gegensatz zu den zweiten fünf Phrasemen, die in der erwähnten Datenbank nicht zu finden waren. Obwohl man an dieser Stelle von keiner generellen Korrelation ausgehen kann, ist aus der Abbildung 5 evident, dass die Phraseme *den inneren Schweinehund überwinden* und *Eulen nach Athen tragen* bei beiden Adressatengruppen nachgestellt oder zumindest als weniger bekannt empfunden werden, was bereits die Untersuchungen von Hallsteinsdóttir et. al. (2006) bestätigt haben. Auf der anderen Seite gibt es aber Phraseme wie z. B. *jmdm. auf den Keks gehen*, *die Flinte ins Korn werfen*, die durchaus als bekannt und frequent bewertet werden. Somit wird anhand eines praktischen Beispiels der Bedarf nach ständiger Aktualisierung einer elektronischen Phrasem-Datenbank evident.

4 Schlussfolgerungen

Die Ergebnisse der Forschungsarbeit zum Thema Digitalisierung der Phraseologie, die sich auf Deutsch als Fremdsprache beschränkt hat, zeigen, dass es an empirischen Studien mit Betonung auf Benutzungsforschung in der elektronischen Umgebung mangelt. Die Benutzungsforschung und die daraus resultierenden Erkenntnisse sollen jedoch eine Grundlage für die Weiterentwicklung der elektronischen phraseologischen Nachschlagewerke darstellen. Sie tragen entscheidend zur Weiterentwicklung der elektronischen phraseologischen Materialien bei. Daher plädiere ich für eine intensivere Benutzungsforschung, die samt der Ausnutzung der IKT wichtige Kenntnisse über die sog. E-Benutzer liefern könnte.

Literatur

ALMIND, Richard et al., 2006: Theoretical and Computational Solutions for Phraseological Lexicography. *Linguistik online* 27/2, 159–181.

ANISIMOVA, Elena V., 2002: *Phraseologismen im Unterricht Deutsch als Fremdsprache.* <http://www.daad.ru/wort/wort2002/Anisimova.Druck.pdf>. Zugriff 25. 11. 2012.

BERGENHOLTZ, Henning / JOHNSEN, Mia, 2005: Log Files as a Toll for Improving Internet Dictionaries. *Hermes, Journal of Linguistics* 34, 117–141.

BERGENHOLTZ, Henning, 2006: Idiomwörterbücher und ihre Benutzer. Breuer, Ulrich et. al. (Hrsg.): *Wörter-Verbindungen: Festschrift für Jarmo Korhonen.* Frankfurt am Main: Peter Lang. 19–30.

BURGER, Harald, 2003: *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Berlin: Erich Schmidt Verlag.

DOBROVOL'SKIJ, Dmitrij, 1989: Linguistische Grundlagen für die computergestützte Phraseographie. *Zeitschrift für Germanistische Linguistik* 10, 528–536.

ĎURČO, Peter, 2007: Phraseologie und allgemeines Konzept für eine komplex strukturierte Datenbank. Jesenšek, Vida / Fabčič, Melanija (Hrsg.): *Phraseologie kontrastiv und didaktisch. Neue Ansätze in der Fremdsprachenvermittlung.* Maribor: Slavistično društvo Maribor. 169–181.

EPHRAS, 2006: < http://projects.ff.uni-mb.si/frazeologija/static/ephras/index.html>. Zugriff 6. 12. 2012.

HALLSTEINSDÓTTIR, Erla, 2004a: Konzeption und Erstellung eines computergestützten Phraseologiewörterbuchs Isländisch–Deutsch. Häcki Buhofer, Annelies et. al. (Hrsg.): *Phraseology in Motion I. Methoden und Kritik*. Baltmannsweiler: Schneider Verlag Hohengehren. 101–112.

HALLSTEINSDÓTTIR, Erla, 2004b: A bilingual electronic dictionary of idioms. Gottlieb, Henrik et. al. *Dictionary Visions, Research and Practice*. Copenhagen: Benjamins. 97–106.

HALLSTEINSDÓTTIR, Erla et. al., 2006: Phraseologisches Optimum für Deutsch als Fremdsprache. Ein Vorschlag auf der Basis von Frequenz- und Geläufigkeitsuntersuchungen. *Linguistik online* 27/2, 117–136.

HALLSTEINSDÓTTIR, Erla (2009): Zweisprachige Lernenrphraseographie aus funktionaler Sicht. Mellado Blanco, Carmen (Hrsg.): *Theorie und Praxis der idiomatischen Wörterbücher*. Tübingen: Niemeyer. 209–231.

HEID, Ulrich (2011): Electronic dictionaries as tools: towards an assessment of usability. Fuertes-Olivera, Pedro A. et. al. (Hrsg): *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum. 287–304.

HESSKY, Regina, 2007: Perspektivwechsel in der Arbeit mit Phraseologie im DaF-Unterricht. Jesenšek, Vida / Fabčič Melanija (Hrsg.): *Phraseologie kontrastiv und didaktisch. Neue Ansätze in der Fremdsprachenvermittlung*. Maribor: Slavistično društvo Maribor. 9–17.

JESENŠEK, Vida, 2006a: Phraseologie in der Fremdsprache Deutsch. Krumm, Hans-Jürgen et al. (Hrsg.): *Schwerpunkt: Innovationen–neue Wege im Deutschunterricht*. Innsbruck, Wien, Bozen: Studien Verlag. 117–129.

JESENŠEK, Vida, 2008: *Begegnungen zwischen Sprachen und Kulturen. Beiträge zur Phraseologie*. Bielsko-Biala: Akademia Techniczno-Humanistyczna.

JESENŠEK, Vida, 2009: Phraseologische Wörterbücher auf dem Weg zu Phraseologiedatenbanken. Mellado Blanco, Carmen (Hrsg.): *Theorie und Praxis der idiomatischen Wörterbücher*. Tübingen: Niemeyer. 65–83.

KEMMER, Katharina, 2010: *Onlinewörterbücher in der Wörterbuchkritik. Ein Evaluationssraster mit 39 Beurteilungskriterien*. <http://pub.ids-mannheim.de/laufend/opal/pdf/opal2010-2.pdf>. Zugriff: 5. 12. 2012.

KRALJ, Nataša, 2012: *Phraseologie und Phraseographie in der Zeit der IKT-ÄRA: dargestellt am Beispiel eines elektronischen phraseologischen Materials*. Dissertation. Maribor: Universität Maribor.

KÜHN, Peter, 2007: Phraseologie des Deutschen. Zur Forschungsgeschichte. Burger, Harald et al. (Hrsg.): *Phraseologie / Phraseology. 2. Halbband*. Berlin/New York: de Gruyter. 619–643.

PALM, Christine, 1997: *Phraseologie. Eine Einführung*. Tübingen: Günter Narr.

Prosodic Aspects of Proverb Change in English: Panini's Principle

CLAUDIA LÜCKERT (Münster)

Abstract

In earlier publications it has been noted that there are semantic and phonological constraints which determine the structure of set phrases. While the focus in past research has been on binomials, this pilot study shall shed some light on phonological constraints, Panini's Principle in particular, with regard to English proverbs with a parallel structure. It is argued in favour of the notion that the influence of Panini's Principle seems quite pervasive in such proverbs, but that further phonological principles may be at work at the same time. Proverbs with a parallel structure, for instance, appear to favour the "syllabic symmetry principle" in the juxtaposed lexical kernels. The corpus-based case study of *Here today, gone tomorrow* centres on synchronic variation, whereas the case study of *Many a little makes a mickle* draws on diachronic data and traces inventorisation processes. The data will be discussed in light of the concepts of iconicity and linguistic economy.

1 INTRODUCTION

It has been observed that set phrases tend to show specific patterns of semantic and phonological sequences (Cooper/Ross 1975, Fenk-Oczlon 1989, Heath 2004, Benor/Levy 2006). Cooper and Ross, however, make clear that there are exceptions, and that particular principles may override others (1975: 77). The role of one of these structuring principles – namely Panini's Law, or more correctly Panini's Principle – has been explored repeatedly over the years in a number of studies.

Most publications so far, however, have discussed this phenomenon in view of binomials (Cooper/Ross 1975, Fenk-Oczlon 1989, Benor/Levy 2006) or coordinated constructions (Heath 2004, Schlüter 2005) in general. Proverbs have at most been mentioned in passing in connection with Panini's Principle (for example Norrick 2007: 382).

This contribution intends to shed some light on the role of prosodic aspects in varying and changing proverb structure. As proverbs come in many different forms it stands to reason to choose a particular sub-set that is rather homogeneous in terms of lexico-grammatical structure. The focus in the present study, therefore, is on proverbs with a clearly parallel structure generated by a syntactic "proverb pattern".

The topic at hand will be approached in form of two case studies which exemplarily demonstrate the workings of prosodic factors that may be observed in varying proverb structure. In a follow-up study I plan to substantiate my views on the role of phonological structuring principles in English proverbs with a corpus which offers more data.

This study has been inspired by observations in diachronic data of English proverbs (Aurich 2012) and will therefore include a diachronic case study as a case in point. In general, studying the workings of prosodic constraints may benefit greatly from access to larger corpora. Historical corpora of English, however, are still very limited in size. Therefore I thought it necessary to carry out a case study which draws on a wealth of data testifying to synchronic variation. For this purpose I decided to work with COCA (Davies 2008) and the World Wide Web as these are modern corpora which make such a wealth of data available in spite of some limitations – see Colson (2007: 1072) on the limitations of the World Wide Web as a corpus.

2 Proverb Structure and Panini's Principle

It has been widely noted that proverbs are characterised by a stability of their form (for example Norrick 2007: 381) which makes these set phrases recognisable in actual use. At the same time it has been highlighted that this stability is only relative. Proverbs may after all circulate in variant forms. This paper addresses the question of which factors may be held responsible for linguistic choices in such variant forms and focuses on phonological constraints.

A working assumption adopted here is that if there are variants of a given proverb they often differ in frequency – some may be more commonly used than others. It is further assumed that if there are indeed differences in frequency across such proverb variants these may not be entirely random. Studying the particular interrelatedness of a variety of factors coalescing in a given proverb and its variants may enable us to identify an underlying motivation.

In order to collect relevant data, it may seem reasonable to document and analyse a given proverb and its variant forms. Differences in frequency attested in corpora may then point to preferred structures (that is the more frequent variants) and dispreferred structures (that is the less frequent variants). There are – it goes without saying – drawbacks to retrieving data on the token frequency of proverb variants from corpora, as these may give a (slightly) wrong impression of how frequently (and for that matter, in which contexts) the proverb variants are really used by a speech community.[1] Nonetheless, large-scale corpora may point to tendencies at least.

When it comes to the structure of proverbs, we may identify a number of characteristics that are frequently used. Mieder, for example, mentions alliteration, rhyme, parallelism, and ellipsis (2007: 396). Such patterned structures make possible that we may more easily memorise and recognise such an item (Norrick 2007: 382). In the following, the focus shall be on parallelism.

Parallelism in proverbs may be the result of two different strategies. For one thing, a syntactic "proverb pattern" may impose parallelism structurally. This may be observed in proverbs such as *Better late than never* which draws on the well-known pattern 'better *x* than *y*'. For another, proverbs may show parallelism that is generated by the lexical meaning of the juxtaposed elements. This "patternless" construction may be identified in the historical English proverb *As fast as one goeth another cometh in ure* ('As fast as one goes, another one comes running in') from Heywood's sixteenth-century proverb collection (No. 110). Although we find a "proverb pattern" ('as *x*, *y*') the meaning is governed by the antonymic relation which holds between *to go* vs. *to come*.

Proverbs with a parallel structure based on a "proverb pattern" are not as restricted in the possible variations of the form when compared to proverbs which generate parallelism with the help of synonyms or antonyms. This is because of the slots provided by "proverb patterns" which may be filled quite freely – and which may allow the creation of proverb variants that show different choices in these slots. In contrast, drawing on "patternless" constructions that rely on antonyms, for instance, will not allow variations as easily. In the two case studies below, I will therefore look into proverbs which rely on "proverb patterns".

The various factors that may determine the structure of a set phrase in general include semantic constraints (Cooper/Ross 1975, Fenk-Oczlon 1989, Heath 2004, Benor/Levy 2006), frequency constraints (Fenk-Oczlon 1989, Benor/Levy 2006) and phonological constraints (Cooper/Ross 1975, Fenk-Oczlon

[1] I am much indebted to Peter Grzybek for his comments on this paper and my workshop contribution at the EUROPHRAS conference in Maribor in August 2012. He pointed out, that we need to bear in mind in this respect, for instance, the use of a proverb variant as the title of an album or a movie – this may be behind a great number of tokens if using the World Wide Web or actual language corpora. Quoting the title of a movie may not count as a "natural" use of the proverb in the strictest sense, but would surely contribute to the item's familiarity, and possibly use.

1989, Benor/Levy 2006). Research has shown that these constraints are not all on a par but that some factors may take precedence over others:

> The main trend we found in our data was the prominence of semantic over metrical, and metrical over frequency constraints. (Benor/Levy 2006: 271)

As these observations are entirely based on binomials and coordinated constructions, the ranking may look slightly different in other types of set phrases. What is more, it may be worthwhile to explore particular factors within these constraints in light of a possible weighing.

In this respect, we may ask whether all phonological constraints are equally influential in diverse contexts. In the case studies below, the focus will be on the role of Panini's Principle, that is the tendency that words with fewer syllables come before words with more syllables – or on a more general level, that short elements are placed before long ones.[2] What is more, the Principle of Rhythmic Alternation – that is the tendency that stress clash is avoided (Schlüter 2005: 124–129) – shall be dealt with. Moreover, proverbs with a parallel structure generated by a syntactic "proverb pattern", seem to prefer a prosodic pattern that draws on symmetry. The juxtaposed lexical kernels in the slots of the "proverb pattern" often sport a similar, or even identical, number of syllables. This may be the result of a "syllabic symmetry principle" at work in this kind of context.[3]

3 Case Studies

3.1 *Here today, gone tomorrow*

In this section, a familiar English proverb, namely *Here today, gone tomorrow*, shall be analysed with a special focus on variant forms and differences in frequency. The variant forms shall be considered against the background of prosodic aspects – in particular Panini's Principle, the Principle of Rhythmic Alternation, and the "syllabic symmetry principle".[4] I am well aware that a

[2] Fenk-Oczlon suggests that this goes hand in hand with the tendency that high frequency structures precede those with a low frequency (1989: 522).

[3] Elsewhere I have discussed this prosodic principle as "syllabic harmony principle" (Aurich 2012). However, "syllabic symmetry principle" may be a more fitting label.

[4] Again, I would like to thank Peter Grzybek for his comments. With regard to the structural variants of this proverb, he made me aware of the forms with *and* (e. g. *Here today and gone tomorrow*) and pointed out that the *and* further supports the symmetrical construction.

variety of further factors may leave their mark on proverb structure. This is why the analysis of the proverb at hand is sketchy. Most importantly, in this case study the significance of the contextual use of the individual variants attested in a variety of sources will be neglected, although different contexts may require modifications in a proverb.[5]

So as not to base the discussion of proverb variants and their frequency on intuition only, two corpora – COCA (Davies 2008) and the World Wide Web – were chosen to extract relevant data. With some 450,000,000 words COCA is, at present, the largest freely accessible online corpus of English. Despite this huge size, the variants of *Here today, gone tomorrow* are, however, at best sparsely documented in this corpus (cf. table 2). Therefore, it appeared necessary to collect data elsewhere. According to Colson, the advantages of studying variant forms of set phrases on the Web lie in the possibility to "analyze a profusion of variants and modifications of set phrases" (2007: 1075). And indeed, the Web offered data on the structural variants of the proverb in question.

The limitations of the Web as a corpus have been discussed elsewhere: the contents of the Web have not been controlled by a linguist, the total amount of words in the corpus is not known – search engines "actually make a copy of only a part of the Web", and the naturalness of internet language is highly doubtful (Colson 2007: 1072). The Google advanced search engine takes a "snapshot", as it were, of only a part of the Web. If the same search is carried out anew, another part of the Web may be copied and files with hits may have been uploaded or deleted in the meantime. For these reasons searches on the Web cannot be considered an exact corpus-linguistic tool as results cannot be reproduced. Nonetheless, queries carried out on the Web may still give hints at tendencies.

The variants of *Here today, gone tomorrow* considered in the present study are listed with their token frequencies in both COCA and the World Wide Web in table 2. Before we turn to a discussion of which variants occur more frequently or less frequently, respectively, and which underlying principles may be held responsible for frequency differences in these variants, let us look at the structure of the canonical proverb variant itself.

The construction of *Here today, gone tomorrow* is characterised by a number of structuring principles. Among these we may identify semantic constraints, namely both the spatial sequencing in *here* vs. *gone* and temporal sequencing

[5] I am grateful to Natalia Filatkina for her comments on this paper. She stressed the relevance of taking contextual aspects into account.

in *today* vs. *tomorrow* (Fenk-Oczlon 1989: 532ff.). Norrick (2007: 382) mentions this proverb as an example of Panini's Principle, though he does not go into any details where exactly in the proverb this principle holds. Below, it shall be demonstrated that there are indeed two layers which show Panini's Principle. Furthermore, it may stand to reason to assume that related "proverb patterns", say 'today *x*, tomorrow *y*' (*Today here/x, tomorrow the world*) or similar structures in frequent set phrases may play a role.

The most frequent variant in the COCA turns out to be the canonical form in (1a) *Here today, gone tomorrow.* The distribution of Panini's Principle and the Principle of Rhythmic Alternation in this variant may be charted as follows:

	Here today, gone tomorrow
Stress (Schlüter 2005)	x x x x x x x x x x x x x
Layers of Panini's Principle	short-long short-long short long

Table 1: Distribution of stress and Panini's Principle in *Here today, gone tomorrow.*

In this proverb, Panini's Principle seems quite pervasive inasmuch as the monosyllabic elements come first in both parts (*here*, *gone*) and the first part (*here today*) is shorter than the second part (*gone tomorrow*). This holds true in the other variants included in Table 2 which show a quite clear picture. Based on the Web data, we may see that variants (a) and (b) – which adhere to the order of elements described in Table 1 – are favoured over variants (c) and (d) by far.[6]

The stress pattern for this proverb is actually variable to a certain extent but usually takes the form given in Table 1. What may be observed in this proverb is that stress clash is clearly avoided: cf. the low number of tokens of a variant such as *Today here, gone tomorrow* (not included in Table 2) where heavily stressed *here* is immediately followed by *gone* which is stressed just as much (on the Web, Google advanced search on 26/11/2012 yielded 605 tokens in hits for the language 'English' and region 'United States').

In the data from the World Wide Web, variant (2b) *Here today and there tomorrow* with a token frequency of 2,700,000 stands out as the most frequent variant by far (even though it does not occur in COCA). Example 2 (b) may be

[6] Cf. frequency differences in 1 (a), 1 (b), 2 (a), 2 (b), 3 (a), 3 (b) in contrast to 1 (c), 1 (d) etc.

favoured over all other variants considered here because of Panini's Principle in the proverb pattern, the word length of *there* (which is shorter than the disyllabic *elsewhere* in variants 3 (a) etc.) and word frequency of *there* (COCA word rank of *there* is 116 in contrast to *elsewhere* which is 2,299), the "syllabic symmetry principle" observed in the juxtaposed kernels *here* and *there* which is further supported by *and*, as well as because of the Principle of Rhythmic Alternation (Schlüter 2005) which is again associated with the coordinator *and* in this construction.

But maybe as importantly, variant 2 (b) may be preferred because of the highly frequent binomial *here and there*.[7] This binomial may strengthen the proverb which is quite similar in form and meaning. Moon (2008) has discussed in how far different types of set phrases may influence each other. 'Collocational reinforcement' may strengthen similar structures, whereas "collocational interference", that is a type of clash, may weaken a given set phrase (Moon 2008). With regard to variant 2 (b) the meaning of both binomial and proverb is similar enough so as not to be an obstacle.

Proverb Variants of ***Here today, gone tomorrow***		**COCA**	**World Wide Web**
1	(a) Here today, gone tomorrow	(a) 21	(a) 646,000
	(b) Here today and gone tomorrow	(b) 17	(b) 205,000
	(c) Today here, tomorrow gone	(c) ---	(c) 15,200
	(d) Today here and tomorrow gone	(d) ---	(d) 5
2	(a) Here today, there tomorrow	(a) ---	(a) 1,490,000
	(b) Here today and there tomorrow	(b) ---	(b) 2,700,000
	(c) Today here, tomorrow there	(c) 1	(c) 209,000
	(d) Today here and tomorrow there	(d) ---	(d) 12,900
3	(a) Here today, elsewhere tomorrow	(a) ---	(a) 6,680
	(b) Here today and elsewhere tomorrow	(b) ---	(b) 30,000
	(c) Today here, tomorrow elsewhere	(c) ---	(c) 98
	(d) Today here and tomorrow elsewhere	(d) ---	(d) 1

Table 2: *Here today, gone tomorrow*: Token frequency of variants in COCA and World Wide Web.[8]

[7] A search on the World Wide Web on 22/11/2012 rendered a token frequency of 62,900,000 for *here and there* if restricted to 'English' and region 'United States'. Data retrieved from COCA (complete corpus searched on 22/11/2012) offers 3,269 tokens.

[8] The search for the variants was carried out in all of COCA that is the whole 450 million-word corpus (Davies 2008) on 21/11/2012. The data from the World Wide Web was retrieved on 21/11/2012 with the help of the Google advanced search engine. The queries for the variants were restricted to hits for the language 'English' and region 'United States'.

Variant 1 (b) is almost as well suited with regard to word length and word frequency of *gone* (COCA word rank of *to go* is 35), rhythmic alternation, increased symmetry associated with *and*, but it lacks an equivalent, frequent fixed expression – a binomial, say, **here and gone* (the string *here and gone* is found much less frequently than the binomial *here and there* on the World Wide Web) – which may support the proverb.[9]

3.2 *Many a little makes a mickle*

In section 3.1 we could observe the interplay of a number of prosodic structuring principles, individual word frequencies of lexis which may vary synchronically across different variants of the same proverb, word length, and the workings of "collocational reinforcement" (Moon 2008). In the present section, however, the focus will be on diachronic variation and inventorisation processes. The case study of *Many a little makes a mickle* shall exemplarily demonstrate in how far prosodic structuring principles may be involved in bringing about change in a given proverb.

Proverb Variants of *Many a little makes a mickle*	
1250	/…/ of ***lutel muchel*** waxeth. ('Of little much grows')
1390	/…/ manye ***smale*** maken a ***greet***. ('Many small (things) make a great (one)')
1545	Many a ***little*** maketh a ***great***. ('Many little (things) make a great (one)')
1614	Many a ***little*** makes a ***mickle***.
1822	Many a ***little*** makes a ***mickle***.
1905	/…/ many a ***pickle*** maks a ***muckle***.
1979	Many a ***pickle*** (or ***little***) makes a ***mickle***.

Table 3: Variants across time of *Many a little makes a mickle*.[10]

If we trace the development of the proverb at hand, a number of interesting aspects may be identified. For one thing, the proverb which is still in use today has a quite long tradition which reaches back to at least the mid-thir-

[9] The query for *here and gone* on 22/11/2012 on the Web yielded a token frequency of 220,000 if limited to 'English' and region 'United States'. COCA (complete corpus searched on 22/11/2012) rendered 16 tokens.

[10] For information on the sources see Aurich (2012: 202–204).

teenth century. For another, variant forms attested in historical sources show that elements have been replaced which still were motivated. This is why substitutions of the kind observed when comparing, say, the variant dating from 1545 (which has *little* and *great*) to that one from 1614 (which has *little* and *mickle*), cannot be accounted for by the necessity of "up-dating" archaic structures. The reason – or, for that matter, the reasons – for this modification must lie elsewhere.

In the variant preserved in a document from about 1250 Panini's Principle shows in the ordering of a short element (the first part *of lutel* with three syllables) followed by a slightly longer element (the second part *muchel waxeth* with four syllables). The two juxtaposed kernels *lutel* and *muchel* are placed next to each other. The two items, which are characterised by assonance and an identity in syllable number, are placed in a temporal sequence (cf. semantic constraint). Their proximity does, in fact, not go against the Principle of Rhythmic Alternation inasmuch as *lutel* is stressed on the first syllable only – *muchel*, then, actually follows an unstressed syllable.

In more recent sources, the proverb, however, takes a different form. The syntactic proverb pattern 'many *x* make *y*' seems to have been used ever since the late-fourteenth century. The two kernels in the slots of the pattern have been repeatedly exchanged for other ones. These substitutions may look quite haphazard at first glance. Taking prosodic aspects into account, however, may explain these decisions.

Again, Panini's Principle may have been one of the reasons why the variant of 1390 has been modified in later tradition: The first part *manye smale* with three syllables is indeed shorter than the second part *maken a greet* with four syllables; the sequencing within each part, however, violates Panini's Principle (e. g. *manye* with two syllables vs. *smale* with one syllable). What may be especially noteworthy here is the replacement of *great* by *mickle* which is first attested in the variant from 1614. That ever since *mickle* – a Northern English dialect word – has been favoured in this proverb over *great* may be attributed to the "syllabic symmetry principle".

This preference may still be observed in modern data: *Many a little makes a mickle* had a token frequency of 90,000 on the World Wide Web on 27/11/2012 if limited to 'English' and the region 'United States' – the variant *Many a little makes a great* on the other hand turned up just six times when running the same search. This may be taken as a hint that the individual word frequency (*great* after all is highly frequent) may be of a lesser impact when compared to the "syllabic symmetry principle".

4 Conclusion: Iconicity and Linguistic Economy

This study was meant to shed some light on the role of Panini's Principle in varying and changing proverb structure with the help of two little case studies. At the same time it was my aim to highlight the importance of paying attention to further prosodic principles such as the Principle of Rhythmic Alternation and the "syllabic symmetry principle".

There are, unmistakably, drawbacks to the approach adopted in this paper: two sketchy case studies which neglect a variety of factors do not allow for general conclusions. They may at best generate an interest into a more detailed consideration of the issues at hand. What is surely needed is large-scale corpus data on both structural variants of proverbs and their frequencies. In a more detailed study, a large number of proverbs with a parallel structure (and possibly with a number of structural variants) would have to be subjected to the analysis used in section 3.1.

While this may work quite nicely for modern English proverbs on the basis of COCA and the World Wide Web, historical English proverbs are hardly ever attested in language corpora. Moreover, an analogous diachronic study carried out in a way equivalent to the analysis in section 3.1 would require data on the frequency of proverb variants, the frequency of individual words (especially of the lexical kernels), the frequency of similar set phrases ("collocational reinforcement" and "collocational interference", see Moon 2008), as well as data on likely lexical alternatives for the lexical kernels (is there any choice at all in opting for a given structure?). This approach may, it goes without saying, already fail in the first step – in identifying a string of words in an historical document as a proverb.

This pilot study cannot furnish evidence on whether particular prosodic structuring principles are all on a par in determining the structure of parallel proverbs or whether some prosodic principles may override others in this context. It is my assumption, on the one hand, that the lexical kernels in proverbs with a parallel structure imposed by a syntactic "proverb pattern" favour the "syllabic symmetry principle". The overall structure of such a proverb, on the other hand, seems to prefer Panini's Principle for the sequencing of elements (say, in a two-part parallel construction a shorter first part would be placed before a longer second part).

A few observations on what may underlie these tendencies are in order. As pointed out in section 2 above, the patterned structure of proverbs makes it easier to memorise and recognise such set phrases. With regard to patterns, it may be claimed that in particular set phrases, such as binomials, repetition is avoided. This may be explained on the basis of perception phenomena – experimental data has shown that "subjects are prone to fail to perceive repeated morphemes, words or semantic concepts" (Schlüter 2005: 265).

In proverbs, however, "repeated" – that is identical or at least similar – structures tend to be not in that much proximity: compare the binomial ***here** and* ***there*** to the proverb ***Here** today and* ***there*** *tomorrow.* Therefore, we may hold that "repeating" a given structure in a proverb does not inhibit perception – it may rather contribute to a stronger joining of the lexical kernels. We may even speculate whether his strengthened bond of the juxtaposed lexical kernels in proverbs would be required even more strongly the further apart these kernels are in a given proverb.

The "repetition" of structures in proverbs with a parallel construction may, for example, take the form of juxtaposed lexical kernels that are characterised by an identity in syllable number, rhyme, assonance, or, say, alliteration. Generally, it may involve less effort to find lexis with an identity in syllable number only than to find rhyming words etc. The principle of language economy may, indeed, best explain why proverb structure seems to rely heavily on the "syllabic symmetry principle" in this context. It has been observed that iconic coding is easier for speakers and hearers (Fenk-Oczlon 1989: 535). In the case of parallel proverbs we may describe the choice of lexical kernels which agree in syllable number as iconic coding and formulate an iconic principle "increase of formal identity = increase of conceptual closeness" (cf. the notion of "conceptual closeness" in Haiman 1983: 783, Haspelmath 2008: 63).

Bibliography

AURICH, Claudia, 2012: *Proverb Structure in the History of English: Stability and Change. A Corpus-Based Study.* Baltmannsweiler: Schneider Verlag Hohengehren.

BENOR, Sarah / LEVY, Roger, 2006: The Chicken or the Egg? A Probabilistic Analysis of English Binomials. *Language* 82/2, 233–278.

COLSON, Jean Pierre, 2007: The World Wide Web as a Corpus for Set Phrases. Burger, H. et al. (eds.): *Phraseologie. Phraseology. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research.* 2. Halbband. Volume 2. Berlin/New York: de Gruyter. 1071–1077.

COOPER, William E./ROSS, John R., 1975: World Order. Grossman, Robin et al. (eds.): *Papers from the Parasession on Functionalism*. Chicago: Chicago Linguistic Society. 63–111.

DAVIES, Mark, 2008–: *The Corpus of Contemporary American English: 450 million words, 1990-present* COCA: <http://corpus.byu.edu/coca/>.

FENK-OCZLON, Gertraud, 1989: Word Frequency and Word Order in Freezes. *Linguistics* 27, 517–556.

HAIMAN, John, 1983: Iconic and Economic Motivation. *Language* 59/4, 781–819.

HASPELMATH, Martin, 2008: Reply to Haiman and Croft. *Cognitive Linguistics* 19/1, 59–66.

HEATH, Jeffrey, 2004: Coordination. An Adaptationist View. Haspelmath, M. (ed.): *Coordinating Constructions*. Amsterdam, Philadelphia: Benjamins. 67–88.

MIEDER, Wolfgang, 2007: Proverbs as Cultural Units or Items of Folklore. Burger, H. et al. (eds.): *Phraseologie. Phraseology. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research.* 1. Halbband. Volume 1. Berlin/New York: de Gruyter. 394–414.

MOON, Rosamund, 2008: Conventionalized As-Similes in English. A Problem Case. *International Journal of Corpus Linguistics* 13, 3–37.

NORRICK, Neal R., 2007: Proverbs as Set Phrases. Burger, H. et al. (eds.): *Phraseologie. Phraseology. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research.* 1. Halbband. Volume 1. Berlin/New York: de Gruyter. 381–393.

SCHLÜTER, Julia, 2005: *Rhythmic Grammar. The Influence of Rhythm on Grammatical Variation and Change in English*. Berlin/New York: de Gruyter.

Acerca de la (in)traducibilidad de las unidades fraseológicas en la interpretación de conferencias

JASMINA MARKIČ (Ljubljana)

Abstract

The article deals with phraseological units and conference interpreting. The differences between translation and interpreting are essentially associated with the cognitive stress interpreters face under pressure of time. In their work interpreters must bridge the linguistic, cultural and conceptual gaps separating the participants in a meeting. Whether working in consecutive or simultaneous, the interpreters have first to listen to the speaker, understand and analyse the speech and re-express it in a different language. The difficulties arise in both phases: understanding and re-expressing. The interpreters cope with these difficulties using different tactics and strategies. Phraseological units appear in different types of speeches and represent one of the linguisitc and cultural barriers to communication. Based on the Corpas Pastor's definition and classification the article analyses some examples of colocations, locutions, paremias and formulas which appear in Spanish speeches and are interpreted into Slovene.

1 INTRODUCCIÓN

El intérprete de conferencias se encuentra regularmente con aspectos culturológicos de los significados de palabras y textos. Las unidades fraseológicas de la lengua de origen pueden ser un hueso duro de roer o, como se dice en esloveno, pueden *ser una nuez dura* (*so trd oreh*) cuando hay que trasladarlas a la lengua meta. El presente trabajo trata de las estrategias que usan los intérpretes para trasladar adecuadamente el significado de las unidades fraseológicas que encuentran en los discursos que interpretan, es decir, intenta vincular dos disciplinas: la fraseología y la interpretación de conferencias.

2 DIFERENCIAS Y SEMEJANZAS ENTRE LA INTERPRETACIÓN Y LA TRADUCCIÓN

Ambos, tanto el intérprete como el traductor, tienen la misma tarea: la de transmitir un mensaje de la lengua de origen a la lengua meta. Es obvio que los intérpretes y los traductores trabajan en condiciones y niveles diferentes y que existen diferentes modalidades de traducción e interpretación. Sin embargo, su actividad básica es la misma:

> Beneath the diversity of performance levels and conditions, interpretation and translation can be defined as performing essentially the same function, namely reexpressing in one language what has been expressed in another (Gile 1995b: 2).

Pero también está claro que existen diferencias, tanto en el marco temporal y espacial como en las técnicas de trabajo. La más evidente es que la interpretación es oral, por lo tanto, el nivel fonético, tanto en la percepción como en la producción, desempeña un papel importantísmo. El lenguaje hablado tiene una serie de elementos (Torres Díaz 1996: 15) de los que carece el lenguaje escrito: prosódicos (entonación, ritmo, acentuación, pausas) y paralingüísticos e indéxicos (características de la voz, lenguaje no verbal, actitud del hablante). Es menos denso que el escrito, con más elementos deícticos y de relación, con términos de alta frecuencia, tendencias a la redundancia y menos complejidad gramatical. En una situación ideal interpretar significa transmitir el mensaje oral de los oradores, no obstante, en muchos casos, el intérprete se enfrenta a textos escritos leídos a gran velocidad, lo que quiere decir que el intérprete transforma el lenguaje escrito en lenguaje oral. El intérprete hace frente a un estrés cognitivo debido a la presión de tiempo. Durante la fase de la comprensión auditiva y análisis el intérprete escucha activamente el mensaje una sola vez (en la modalidad consecutiva escucha el mensaje entero, mientras que en la modalidad simultánea no – en este caso le falta además el contexto entero). El texto (mensaje) es dinámico y el intérprete le sigue activamente. En la modalidad consecutiva las fases de comprensión (escucha activa y análisis) y de transmisión se siguen sucesivamente, mientras que en la simultánea estas dos fases se traslapan: el intérprete transmite el mensaje anteriormente escuchado, al mismo tiempo que escucha un mensaje nuevo. El traductor no conoce este tipo de estrés, el texto sobre el que trabaja es estático, puede volver a leerlo y corregirlo (*ibídem* 18). El traductor tiene tiempo para investigar y preprarar la traducción, buscar la terminología, redactar y corrregir lo redactado, mientras que el intérprete debe hacer todas las preparaciones antes (la fase de documentación y preparación activa es anterior), ya que durante la interpretación no puede hacerlo y debe confiar en su memoria a corto y largo plazo, en sus conocimientos, sus capacidades, estrategias y técnicas. Es muy importane que el intérprete esté al tanto de todo lo que ocurre en su entorno y en el mundo y que enriquezca su base de conocimiento. El intérprete simultáneo puede documentarse durante la interpretación mientras trabajan sus compañeros de cabina, analizando los documentos, escuchando a los oradores y tomando notas de la terminología, consultando los diccionarios y glosarios *on line*, lo que no es posible en el caso de la interpretación consecutiva.

En la segunda fase de la interpretación la transmisión del mensaje se puede realizar una sola vez, mientras que la traducción se puede rehacer y corregir. El intérprete, a diferencia del traductor, trabaja *in situ* (también puede trabajar a distancia en teleconferencias, videoconferencias…) y recibe pronto el *feedback*, el traductor no. La comunicación es inmediata puesto que existe una interacción entre el orador, el intérprete y el público. El intérprete a diferencia del traductor debe reaccionar rápidamente y tomar decisiones instantáneas.

3 Discursos y fraseología

El intérprete de conferencias actúa en un ambiente de congresos internacionales, conferencias, encuentros y trabaja con discursos (ponencias, discursos, ruedas de prensa, declaraciones, debates, mesas redondas…). El discurso oral es, en la mayoría de los casos, un discurso escrito leído o presentado por el orador. Los textos orales espontáneos ocurren sólo en los debates, preguntas y respuestas. Los discursos con los que trabaja el intérprete son de diferente índole: entre los más frecuentes figuran los discursos en los que el orador presenta los pros y los contras (o bien navega entre unos y otros, o bien presenta primero los pros y luego los contras), los discursos lógicos en los que el orador presenta y argumenta lógicamente un punto de vista, los discursos de tipo narrativo o cronológico, los discursos descriptivos, los discursos polémicos con una enérgica defensa de una opinión y el rechazo de otra, los discursos retóricos en los que la forma prevalece sobre el contenido (discursos de inauguración y clausura, de bienvenida y despedida, brindis…).[1] A primera vista se podría pensar que la fraseología y la interpretación de discursos no tienen puntos comunes y que las unidades fraseológicas no tienen cabida en este tipo de textos. Sin embargo, los discursos políticos, científicos, económicos están llenos de unidades fraseológicas: a los oradores les gusta introducir en sus discursos técnicos locuciones, refranes, citas de obras literarias, fórmulas rutinarias. El intérprete debe estar preparado para este tipo de "sorpresas" realizando con antelación sus "glosarios" particulares de las unidades fraseológicas más usadas en sus lenguas de trabajo, sobre todo de las locuciones y fórmulas rutinarias que suelen aparecer más a menudo en este tipo de trabajo: terminología de conferencias, felicitaciones, condolencias, expresiones de solidaridad en caso de acontecimientos extraordinarios (catástrofes naturales, accidentes) etc.

[1] Para más detalles véase Jones (1998: 16–24).

La lengua es una herramienta para el intérprete y debe dominar perfectamente tanto la(s) lengua(s) activas(s) como la(s) pasiva(s). Para poder interpretar es necesario conocer a fondo la cultura[2] de los países donde se hablan las lenguas que interpreta y de su propio país y su lengua materna. Para Newmark (1992: 133) "la cultura es el modo de vida propio de una comunidad que utiliza una lengua particular como medio de expresión y las manifestaciones que ese modo de vida implica". El mencionado autor distingue el lenguaje cultural[3] del lenguaje universal y personal. Sin este conocimiento del *background* cultural la interpretación no se puede llevar a cabo de una manera satisfactoria. En el caso del español, por ejemplo, el intérprete debe conocer el contexto cultural de veinte países.

4 Las unidades fraseológicas (UFS) y la interpretación de conferencias

En la fraseología española existen varias definiciones de lo que son las unidades fraseológicas y cómo se denominan. Por una parte Corpas Pastor defiende una definición amplia de las unidades fraseológicas, objeto de estudio de la Fraseología: "son unidades léxicas formadas por más de dos palabras gráficas en su límite inferior, cuyo límite superior se sitúa en el nivel de la oración compuesta" (1996: 20) que se caracterizan por su alta frecuencia de uso, su institucionalización, su estabilidad y su idiomaticidad. La autora las divide en dos grupos (las UFS que no forman actos de habla ni enunciados y las que los forman) y tres esferas: dos esferas en el primer grupo (colocaciones y locuciones) y una en el segundo (enunciados fraseológicos) (ibíd.: 51–52). Mario García Page (2008: 7–14), por su parte, defiende la concepción estricta de la Fraseología y afirma que las unidades fraseológicas abarcan únicamente las locuciones.

Para este intento de análisis contrastivo español-esloveno en la interpretación de unidades fraseológicas nos apoyamos en la definición amplia, puesto que para un intérprete de conferencias todas estas construcciones hechas representan dificultades en la interpretación. Se trata de 'piezas prefabricadas' que reflejan significados culturales y pragmáticos a veces difíciles de resolver,

[2] En el sentido amplio de la palabra. El DRAE define la cultura como "Conjunto de modos de vida y costumbres, conocimientos y grado de desarrollo artístico, científico, industrial, en una época, grupo social etc."

[3] El lenguaje cultural abarca, según Newmark, diferentes categorías culturales: ecología, cultura material, cultura social, organización social política y administrativa, términos históricos, términos internacionales, términos religiosos, términos artísticos, gestos y hábitos (1992: 136–145).

teniendo en cuenta las circunstancias en las que trabajan los intérpretes (sobre todo el estrés, la velocidad y las dificultades acústicas e intelectuales (Jones 1998: 72)).

4.1 Colocaciones

En cuanto a las colocaciones[4] el intérprete de conferencias debería estar al tanto de estas estructuras definidas por Corpas Pastor como

> /…/ unidades fraseológicas que, desde el punto de vista del sistema de la lengua, son sintagmas completamente libres, generados a partir de reglas, pero que, al mismo tiempo, presentan cierto grado de restricción combinatoria determinada por el uso (cierta fijación interna). /…/ Al igual que las locuciones, no constituyen enunciados ni actos de habla por sí mismas. (Corpas Pastor 1996: 53)

Las colocaciones son parte integrante de una lengua que un intérprete debería conocer en todas sus lenguas de trabajo. Sin embargo, en un contexto de interpretación aparecen errores por intereferencias lingüísticas, calcos o traducciones directas. Algunas colocaciones suelen repetirse con frecuencia en el lenguaje de conferencias y hacen parte de la terminología de conferencias, otras, aunque se empleen con frecuencia en las conferencias, pertenecen también a otras especialidades como derecho, economía, administración, informática etc., o a la lengua en general. Pocas veces las colocaciones en español y esloveno coinciden, como por ejemplo en las siguientes pertenecientes a la terminología de conferencias: aprobar el orden del día [*sprejeti dnevni red*], solicitar la opinión de [*prositi za mnenje*], procesamiento de datos [*obdelava podatkov*], sesión de apertura, de clausura [*otvoritvena*, *zaključna seja*], sesión plenaria [*plenarna seja*], a puerta cerrada [*za zaprtimi vrati*], conceder, retirar la palabra [*dati*, *odvzeti besedo*], tener la palabra [*imeti besedo*], ceder la palabra [*prepustiti besedo*], revocar una decision [*razveljaviti sklep*].

En la mayoría de los casos las colocaciones no coinciden y el intérprete debe conocerlas y automatizarlas para no perder tiempo y energía buscando expresiones adecuadas durante la interpretación o tratando de parafrasear. En muchos casos el esloveno utiliza una sola palabra para las colocaciones españolas, p. ej.: guardar silencio [*molčati*], otras cuestiones [*razno*], ejercer la presidencia, ejercer el cargo de presidente [*predsedovati*], con plenos poderes

[4] Pejović (2010: 13) opina que, "en el sentido estrecho de la palabra, la colocabilidad supone la particularidad de algunas palabras para formar combinaciones léxico-semánticas restringidas, que se caracterizan por ser estables, habitualizadas y típicas, parcialmente composicionales, fijadas en la norma".

[*polnopravno*], hacer uso de la palabra [*spregovoriti*], dar a conocer [*seznaniti*], papeleta, boletín de voto [*glasovnica*], celebrar sesión [*zasedati*], voto de confianza [*zaupnica*], moción (voto) de censura [*nezaupnica*]. Muchas colocaciones difieren en menor o mayor grado en las lenguas contrastadas, p. ej.: rescindir el contrato [*odstopiti od pogodbe*, *preklicati pogodbo*], levantar acta [*sestaviti zapisnik*], entrar en vigor [*stopiti v veljavo*], infringir el reglamento [*kršiti pravilnik*], acervo comunitario [*evropski pravni red*], llamar al orden [*pozvati*, *opozoriti na red*], presidente en ejercicio [*sedanji*, *delujoči predsednik*], presidente saliente [*odhajajoči predsednik*], relaciones públicas [*stiki z javnostmi*], miembro vitalicio, honorario [*častni*, *dosmrtni član*], con voz y voto [*s pravico do glasovanja*], interventor de cuentas [*finančni nadzornik*], periodo de sesiones [*zasedanja*], convocar una sesión [*sklicati sejo*], reanudar la sesion [*nadaljevati sejo*], levantar la sesión [*zaključiti sejo*], pronunciar un discurso [*imeti govor*], prestar ayuda [*nuditi pomoč*], ceñirse al tema [*držati se tematike*], cuestión de fondo [*vsebinsko vprašanje*], entablar el debate [*odpreti*, *sprožiti razpravo*], plantear una pregunta [*zastaviti vprašanje*], tomar nota [*vzeti na znanje*], tomar medidas [*sprejeti sklepe*], someter a votación [*dati na glasovanje*], dar su opinión [*izraziti mnenje*], desatar una polémica [*sprožiti polemiko*], desempeñar una función [*opravljati funkcijo*], desempeñar un papel, rol [*igrati vlogo*] y un largo etcétera (véase Markič/Ljeskovac 2011).

4.2 Locuciones

Corpas Pastor (1996: 88) define las locuciones como "unidades fraseológicas del sistema de la lengua con los siguientes rasgos distintivos: fijación interna, unidad de significado y fijación externa pasemática. Estas unidades no constituyen enunciados completos y, generalmente, funcionan como elementos oracionales". Aparecen muy a menudo en todo tipo de discursos inclusive en los más especializados y representan diferentes grados de dificultad para el intérprete. Algunas locuciones españolas tienen equivalentes o semi-equivalentes en esloveno, como por ejemplo: apretar el cinturón [*zategniti pas*], ensuciarse las manos [*umazati si roke*], tener a (la) mano [*imeti pri roki kaj*], estar al alcance de la mano [*biti na dosegu roke*], conocer de primera mano [*kaj izvedeti iz prve roke*], a hierro y fuego [*z ognjem in mečem*], estar con el alma en vilo [*biti na trnih*], ponerse manos a la obra [*vzeti v roke kaj*]. Las locuciones como moverse a gatas [*plaziti se*], mover cielo y tierra [*storiti vse*], darle vueltas al asunto [*premlevati zadevo*], dar gato por liebre [*prevarati*, *oslepariti*], llevarse el gato al agua [*zmagati*, *uspeti*] etc. no se traducen con locuciones en esloveno (en estos ejemplos la traducción al esloveno es con verbos). Muchas locuciones no tienen equivalentes en la otra lengua y no se pueden traducir literalmente sino con paráfrasis: mover cielo y tierra [*storiti vse*], cuatro gatos [*zelo malo ljudi*], tener malas pulgas [*vzkipeti*, *razdražljiv biti*, *ne razumeti nobene šale*], haber gato encerrado [*nekaj smrdeti*, *biti nekaj prikrito*].

Algunas, como por ejemplo coger, tomar el toro por los cuernos (las astas), de origen incierto, se han difundido por todo el mundo, aparecen con gran frecuencia y se traducen literalmente en todas las lenguas (en esloveno: *zgrabiti bika za roge*). Se podría hablar de una cierta globalización de algunas locuciones que se van repitiendo en las conferencias y congresos, en la prensa, la radio y la televisión, la red. Una locución, antes no usada en esloveno, que últimamente aparece con frecuencia en la prensa eslovena y en los discursos (probablemente se trata de un calco de otra lengua), es *z rokami*, *s prsti v marmeladi*[5] con el significado de *in fraganti* y corresponde a la locución española con las manos en la masa.

Toda una serie de locuciones son comunes a diferentes lenguas y en ese caso el intérprete debe conocer las que se usan con más frecuencia. Estas locuciones tienen origen común en la tradición greco-latina (el talón de Aquiles [*Ahilova peta*], la manzana de la discordia [*jabolko spora*], el sufrimiento de Tántalo [*Tantalove muke*], dormirse ne los laureles [*(za)spati na lovorikah*]) o bíblicas y religiosas (llorar como una Magdalena [*močno jokati*], llevar la cruz [*nositi križ*]). En ocasiones pueden presentar ligeras variaciones.

Frente a estas locuciones universales existen también las locuciones culturales que se basan en las realidades peculiares de una cultura como es el caso de los toros en español. El mundo de los toros y corridas ha dado origen a numerosas locuciones que se usan en todos los niveles de la lengua. Para un intérprete del español al esloveno las locuciones provenientes del mundo de los toros son un obstáculo 'peligroso', puesto que la tauromaquia es algo ajeno a la cultura eslovena. Locuciones parecidas o equivalentes no existen en esloveno, por lo tanto, la única salida del apuro es conocerlas para poder parafrasearlas. Algunas de estas locuciones presentes a menudo en los discursos deben parafrasearse en esloveno, p. ej.: andar de capa caída [*biti potrt*, *imeti velike težave*], dejar a alguien en las astas del toro [*nekoga postaviti v nevaren položaj*], mirar los toros desde la barrera [*opazovati z varne razdalje*, *ne se izpostaviti*], hacer la verónica [*umakniti se*].[6]

[5] *Zaloten s prsti v marmeladi* (Nedeljski 22/3/2011); *Zaloten s prsti v marmeladi* (Dnevnik 28/1/2011); *Predsednik SD: Janković vzel zakon v svoje roke, Janša z roko v orožarski marmeladi* (Planet Siol.Net 28/9/2012).

[6] Esta sentencia, en cambio, pone en su sitio la verdad y la equidad, puesto que el Parque Nacional Galápagos, demandante y beneficiario de la indemnización, apuntó hacia otro lado, siguió el juicio contra Petrocomercial, que no tenía pito que tocar en este asunto, pero que, al parecer, se creyó que era la institución adecuada, rica y manirrota, para

4.3 Enunciados fraseológicos

Los enunciados fraseológicos[7] se dividen, según Corpas Pastor en paremias y fórmulas rutinarias (1996: 132) y constituyen enunciados completos. Los oradores se sirven con frecuencia de refranes, citas literarias, juegos de palabras, alusiones a hechos históricos. El intérprete, durante el proceso de interpretación, no tiene tiempo de buscar los equivalentes en su propia lengua y en este caso también depende de su conocimiento general, su saber enciclopédico, las preparaciones previas y sus propios “glosarios fraseológicos” bilingües y/o multilingües.

Como ocurre con las locuciones, ciertos refranes tienen su equivalente en esloveno (p. ej.: a caballo regalado no le mires los dientes [*podarjenemu konju se ne gleda na zobe*], cuando el gato no está, los ratones bailan [*ko mačke ni doma, miši plešejo*]), otros lo tienen parcialmente o no lo tienen, p. ej.: el que a hierro mata a hierro muere / con la misma vara que mides serás medido [*kdor drugemu jamo koplje, bo sam vanjo padel*], cada oveja con su pareja [*enako se z enakim druži*], la cabra siempre tira al monte [*vedno greš, kamor te nosi srce*].

Las citas de obras literarias o alusiones a ellas son muy frecuentes, por lo tanto el conocimiento de la cultura de las lenguas de trabajo es imprescindible para poder hacer una buena interpretación. “Europa es un proyecto que se inscribe a largo plazo y que hace su camino al andar” – este breve fragmento de un discurso de Josep Borell en el Parlamento Europeo hace alusión al conocido poema de Antonio Machado *Caminante*, que no es posible interpretar al esloveno con todas las implicaciones culturales sin concerlo; y aún así, sin una breve explicación adicional (que no se puede hacer siempre en la interpretación simultánea por falta de tiempo, pero sí se podría hacer en la consecutiva) no se captaría el significado de lo dicho. Cuando se trata de obras literarias de importancia universal (Shakespeare y Cervantes se citan a menudo), la transmisión del mensaje es más fácil. En el caso siguiente el esloveno conoce la misma expresión *mlini na veter* que hace alusión a Don Quijote:

> Soy consciente de que voy a enfrentarme a molinos de viento, pero si no lo hago no podría mirarme al espejo. (El Norte de Castilla, 21/03/2001 CREA consulta el 29/8/2012).

pagar los daños producidos por otros, que tenían un muy buen paraguas de protección: la compañía naviera dueña del buque Jéssica, que hizo una verónica a sus responsabilidades con el visto bueno de altísimas esferas; las autoridades de la Marina Mercante encargadas de controlar el transporte de combustible a Galápagos /…/ (CREA *Expreso de Guayaquil*, 04/10/2002: Barajando los días)

[7] La línea divisoria entre las unidades fraseólogicas y las locuciones puede es borrosa.

Las fórmulas rutinarias, a diferencia de las paremias, carecen de autonomía textual (Corpas Pastor 1996: 171). Aparecen en situaciones comunicativas precisas y para poder transferirlas a otra lengua hay que conocer los marcos socio-culturales. En los discursos, sobre todo en los retóricos, suelen aparecer con mucha frecuencia. Se trata de fórmulas de apertura y cierre del discurso, de fórmulas de agradecimiento, de pésame, felicitaciones, brindis.[8] Son piezas hechas, expresiones prefabricadas y convencionales. Para una buena interpretación de dichas fórmulas es imprescindible conocerlas en todas las lenguas de trabajo para poder transmitirlas adecuadamente.

5 Conclusión: ¿qué puede hacer el intérprete frente a las unidades fraseológicas?

Frente a las dificultades que aparecen en la interpretación simultánea y consecutiva (la velocidad con la que muchos oradores presentan sus discursos, la situación de estrés, la mala pronunciación de muchos oradores, la complejidad del texto, el texto leído, preparado o no de antemano) un intérprete busca salidas y tácticas de interpretación. A pesar de las estrategias de preparación aparecen problemas en la situación de interpretación a causa, como lo afirma Gile (1995b: 191), de las limitaciones y errores en las capacidades de procesamiento y las lagunas en la base de conocimientos del intérprete. Muchos de estos problemas aparecen con regularidad también en el caso de intérpretes con larga experiencia profesional. La dificultades surgen tanto en la fase de comprensión como en la de producción. Los intérpretes deben ser conscientes de ellos y adoptar tácticas para salir de las dificultades que se les van presentando a lo largo de su carrera profesional. Algunas de las tácticas en la interpretación simultánea en la fase de comprensión son la táctica de retraso, cuando se espera un segundo o dos para comprender mejor u obtener más contexto antes de interpretar (pero el riesgo es perder el fragmento siguiente) y/o la ayuda del colega en la cabina (sólo es posible si se trata de un término,

[8] Brindis de Su Majestad el Rey Juan Carlos en la cena de gala ofrecida por Presidente de la República de Eslovenia el 4 de julio de 2002: /…/ Deseo agradecerle, Señor Presidente, las amables palabras que nos ha dirigido, inspiradas por los sentimientos de simpatía y sincera amistad que unen a nuestros dos países, y de los que hemos tenido cumplida muestra en el recibimiento que nos ha dispensado el pueblo de Eslovenia en estos días. /…/ Haciéndome eco de esta amistad, permítame que levante mi copa y brinde por su ventura personal y la de su esposa, por las relaciones entre nuestros dos países y por nuestro futuro común como europeos. Muchas gracias. <http://www.casareal.es/GL/actividades/Paginas/actividades_discursos_detalle.aspx?data=5016> (consultado en agosto 2012).

número, locución). Las tácticas en la producción son varias: apuntar términos difíciles, nombres y números, segmentar el texto, resumir y parafrasear, a veces omitir información. En el caso de las unidades fraseológicas el intérprete puede reemplazar la unidad fraseológica de la lengua de origen con una equivalente en la lengua meta; si no hay equivalentes o no los recuerda procede a la praráfrasis. En el caso extremo la omite. Lo más importante es prepararse bien antes y tener a mano todo tipo de "glosarios" fraseológicos que puedan ayudar en caso de emergencia.

Bibliografía

Casa de Su Majestad el Rey: <http://www.casareal.es>. Consultado en agosto 2012.

CORPAS PASTOR, Gloria, 1996: *Manual de fraseología española*. Madrid: Gredos.

CORPAS PASTOR, Gloria (ed.), 2000: *Estudios de fraseología, fraseografía y traducción*. Granada: Comares.

GARCÍA-PAGE SÁNCHEZ, Mario, 2008: *Introducción a la fraseología española. Estudio de las locuciones*. Barcelona: Anthropos.

GILE, Daniel, 1995a: *Regards sur la recherche en interprétation de conférence*. Lille: Presses Universitaires de Lille.

GILE, Daniel, 1995b: *Basic concepts and models for interpreter and translator training.* Amsterdam/Philadelphia: John Benjamins.

JONES, Roderick, 1998: *Conference interpreting explained*. Manchester: St. Jerome.

LUQUE DURÁN, Juan de Dios / PAMIES BERTRÁN, Antonio (ed.), 2005: *La creatividad en el lenguaje: colocaciones idiomáticas y fraseología*. Granada: Método.

LUQUE DURÁN, Juan de Dios / MANJÓN POZAS, Francisco José, 2002: Claves culturales del diseño de las lenguas: Fundamentos de tipología fraseológica. *Estudios de la lingüística del español*, vol. 16. <http://elies.rediris.es/elies16/Claves.html>. Consultado en septiembre–diciembre 2012.

MARKIČ, Jasmina, 2009: El papel de la traducción y la interpretación en el mundo pluricultural y plurilingüe actual. Miguel Aparicio, Elena de (ed.): *La pluralidad lingüística: aportaciones sociales, culturales y formativas* (Aulas de Verano, Série Humanidades). Madrid: Ministerio de Educación. 217–236.

MARKIČ, Jasmina / LJESKOVAC, Nevenka, 2011: *Konferenčna terminologija*. Ljubljana: Znanstvena založba Filozofske fakultete.

NEWMARK, Peter, 1992: *Manual de traducción*. Madrid: Cátedra.

PEJOVIĆ, Andjelka, 2010: *La colocabilidad de los verbos en español con ejemplos contrastivos en serbio*. Kragujevac: Filološko-umetnički fakultet.

REAL ACADEMIA ESPAÑOLA: *Corpus de Referencia del Español Actual* (CREA). <http://corpus.rae.es/creanet.html>. Consulta en noviembre-diciembre 2012.

REAL ACADEMIA ESPAÑOLA: *Diccionario de la lengua española*. Edición electrónica 22ª edición. <http://www.rae.es/rae.html>. Consulta en noviembre 2012.

SECO, Manuel (dir.), 2004: *Diccionario fraseológico documentado del español actual. Locuciones y modismos españoles*. Madrid: Aguilar.

Slovar slovenskega knjižnega jezika, 1970–1991. Ljubljana: SAZU, DZS.

TORRES DÍAZ, María Gracia, 1996: *Manual de interpretación consecutiva y simultánea*. Málaga: Universidad de Málaga Granada Lingvistica. Método Ediciones.

Online Questionnaire Providing Information on most Well-known and Well-understood Proverbs in Slovene Language

Matej Meterc (Ljubljana)

Abstract

The project of an online questionnaire on familiarity of the Slovene paremiological units is represented in the article. The main goal of the questionnaire is to gain empirical data to establish a Slovene paremiological optimum in order to compare it with the Slovak paremiological optimum presented by Ďurčo. Ďurčo's approach was adapted and enriched because of the development of computer technology and internet usage. The preparation and the progress of the project are described along with its main goal. In addition, the article provides estimation what further research possibilities this empirical approach is opening. Therefore it is necessary to add the information about the structure of databases. Based on empirically gained data, these databases are created by modernized software and allow us to use different combinations of filters and other tools.

1 The aim of the article

The aim of the article is to show what the empirical approach (developed by Peter Ďurčo in the early 2000s) offers to us nowadays. Ďurčo's model of the questionnaire was used to gain valuable data on most known paremiological units in the Slovak language in the past. Now it is being used in order to establish a list of the most known Slovene proverbs for Slovene paremiological optimum. Therefore, it needs to be explained how this concept of a paremiological questionnaire is now, approximately ten years after it was established, being modernized. The modernization of the Ďurčo's model allows us to develop a very useful and elegant tool, with which we can gain a large amount of data on proverbs and sayings. It offers new possibilities when modifying some of its details and adapting it into software which corresponds with the state of the information age in 2012. In the following paragraphs, the preparation and the progress of the project will be described. It is necessary to point out what further research possibilities there are in this empirical approach and in the structure of its modernized software. The project of the online questionnaire represents the core of the demographical research in dissertation entitled, Comparison of the Paremiology in Slovene and Slovak Language on the Basis of the Paremiological Optimum.

2 The preparation of the project and the progress of the survey

2.1 Preparation of the online questionnaire and its structure

In order to prepare an experimental corpus represented in the questionnaire, critical reflection of the paremiological material was needed. While writing about qualitative and quantitative parameters, Grzybek and Chlosta (1995: 69) point out that an experimental corpus must contain all potentially known proverbs of a certain culture and it must not be too large. There are two main reasons why only two lexicographical sources were chosen as the basis of our experimental corpus instead of proverb collections. First reason is a plan to do an empirical (demographical) research similar as the Ďurčo's research. We want to have a good basis for comparative study of Slovene and Slovak paremiology. According to Ďurčo's criteria in his questionnaire, it is necessary to focus on basic lexicographical sources. In addition to main Slovak standard language dictionary (*Kratký slovník slovenského jazyka*) and a smaller phraseological dictionary (*Malý frazeologický slovník*), published by Smiešková in 1977, Ďurčo (2002: 51) also used paremiological database which includes reduced material of Zaturecký's collection, reviewed by Mlacek and Profantova in 1997. Unfortunately, there was no such critically reflected paremiographical work done in Slovene language so far. Biggest collections, published in last 50 years are more or less a result of gathering the material from older collections, created in 19th or in the beginning of 20th century. That is the objection in the case of collection *Pregovori in reki na Slovenskem* (Proverbs and sayings from Slovenia), published by Bojc in 1987. Units represented in Marija Makarovič's research from 1974 would perhaps be a slightly better choice. As Grzybek argues (2008: 25) her study is the first attempt to go in the direction of the modern empirical research of Slovene paremiology. Nevertheless, the core of her research (100 proverbs shown to the respondents) is still based on the individual intuition of the folklorist. Makarovič's material consists of additional units, which came to the mind of the respondents after they were confronted with the above mentioned 100 proverbs. After comparing the structure of the material given by individual respondents, it becomes evident that its diversity is quite restricted. We can assume that this is a result of association processes, which were influenced by the set of 100 proverbs, represented to the respondent earlier.

The paremiological units in experimental corpus therefore come from two lexicographical sources: *Slovar slovenskega knjižnega jezika* (Standard Slovene Dictionary) and *Frazeološki slovar v petih jezikih* (Phraseological dictionary in five languages). The second source, Pavlica's dictionary from 1960, is quite archaic and has therefore already been subject to criticism in the past. Nevertheless, because of the lack of other lexicographical (or phraseographical) sources, it is useful, since there are 506 paremiological units included in it. A significant number of units, that is 187, were either the same as or variants of those found with systematical research of SSKJ (599). Alltogether, the questionnaire consists of 918 units in full text presentation (FTP).

Even though the experimental corpus was made on the basis of the main lexicographical source in Slovene language (SSKJ) and one additional lexicographical source (Pavlica's dictionary), there are some paremiological units that were not included in the corpus. This is also why the extra questions at the end of the survey might help with providing a wider picture of the older units which are not found in the dictionaries, but are still well-known and used, and the new ones which appeared in the language not so long ago and are therefore also not found in the dictionaries. Therefore, this database, consisting of answers to additional questions[1] can be understood as an attempt to establish a modern collection of paremiological units in Slovene language. The association effect described in connection with Makarovič's paremiological data is lowered because of two factors – the length of the questionnaire itself and the fact that the respondent is able to add answers to the extra questions long after he marked the 918 units in the core of the questionnaire. It is also an attempt to establish a corpus of paremiological units which Slovene speakers borrow from other languages.

The questionnaire is located on the webpage <http://vprasalnik.tisina.net/>. In the case of Ďurčo's survey, respondents had to download the questionnaire database and afterwards sent it back either by email either by post on floppy disc. The form of the questionnaire itself was prepared on the basis of Ďurčo's survey (2002: 53ff. and 2004: 59ff.) but it also includes some improvements and additional options. The paremiological units are introduced to respondents separately in the form of a FTP as in Ďurčo's survey. The possible answers for each individual paremiological unit are methodologically based on Ďurčo's model: 1. I know it and I use it; 2. I know it, but I do not use it; I do not know it, but I do understand it; 4. I do not know it and I do not understand it; 5. A possibility to add a variant form that one knows. Every respondent remains

[1] These answers are also tagged according to different demographical data (age, sex, level of education and regional groups of dialects).

anonymous. During the registration respondents are asked to specify five pieces of personal data about them: year of birth, sex, level of education, as well as both the regional group of Slovene dialects in which they grew up and the regional group in which they live now. The question about seven regional groups of dialects was not further divided due to the big number (40) of Slovene dialects (Logar, 1966: 134) and the problem of self-identification due to even smaller local variants of dialects or mixed areas. Both questions about the regional group can also be answered with the option "other", which includes a possibility to write down a comment.[2] After the registration all respondents receive their unique code, which allow them to log in again, whenever they want. This is useful since there is quite a large number of proverbs (918) included in the survey. The answers are saved automatically when a respondent logs off, and when he logs in, the program automatically shows the unit, which follows next on his individual list of evaluation. It is of great importance that in the modernized software the respondent is able to change his answers in a separate window whenever he wants. He is also able to see the percentage of the answered questions.

An important improvement of this survey in comparison to Ďurčo's model is that the units are shown to the respondents by a random key. This allows us to see some partial results and tendencies even without using the filter which allows us to choose only the answers of those respondents, who have already completed the survey. It practically means that on a particular day (the situation on 28th August 2012), with 305 surveys out of the 1396 started being completed, one unit was marked approximately 500 times as a result of its random distribution. The basic statistics on the administrator's page show the number of all the respondents taking part in the survey, the number of all the finished surveys and the number of all the answered questions. We are also able to see the complete list of the respondents together with the demographical data. There are also two windows with the number of respondents taking part and the number of all the finished surveys according to days. Each paremiological unit's statistics can be seen separately on the administrator's page.

Another important improvement of this survey in comparison to Ďurčo's model are the filters. Not using the filters, we see the percentage of all the answered questions, even the answers from the respondents who have not finished the survey. Modernized software includes special filters according to the percentage of the fulfilled questionnaire and different demographical data (age, sex, level of education and regional groups of dialects). They can be combined with one another. That allows us for example to choose just the

[2] A lot of respondents from Ljubljana and Slovene speakers living abroad use this option.

answers to a certain proverb by female respondents, aged from 18 to 65, with the level of education higher than 2 and lower than 6, who first lived in the Styrian region group and are now living in Lower Carniolian region group of dialects.[3] These filters are not essential for the paremiological optimum itself (except for the first filter), but they will enable different research in the future. In order to form a paremiological optimum, we only need answers from those respondents who have finished the survey. Here is an example from the survey for the unit *Kakršna mati, takšna hči* [Like mother like daughter] before using filters:

1. I know it and I use it:	278 | 52.7 %
2. I know it, but I do not use it:	203 | 38.4 %
3. I do not know it, but I do understand it:	34 | 6.4 %
4. I do not know it and I do not understand it:	1 | 0.2 %
5. I know a variant (possibility to add a variant):	12 | 2.3 %
Sum:	528 | 100 %

Other answers (given variants) were *Like father like son* (7 entries) and *An apple doesn't fall far from the tree* (5 entries). Answering with another proverb is, of course, a result of a misinterpretation of the term *variant* by a part of the participants, but it gives us valuable information about the synonymy.

On the administrator's page all the proverbs can be seen in a table together with the percentage of each individual proverb. If we want to get a list of the most used proverbs, we arrange them according to the percentage of the first answer. The projection based on respondents' intuitive estimation will prove interesting (as an extra investigation) in comparison with the situation in language corpora (the Slovene corpus *FidaPLUS* and the *Slovak national corpus*). However, this second piece of information about frequency in corpora, will be the one needed to form the paremiological optimum. For the paremiological optimum, we only need a list of the most well-known proverbs, without regard to the usage. This list will be a sum of the percentages of the answers number 1, number 2 and number 5.

The last window on the administrator's page is a set of options of how to arrange the answers from the last, additional section of the questionnaire. This allows us to easily arrange the proverbs that were not mentioned in the survey

[3] Of course this is just an example of what the software allows us to do – we will probably not need to apply all of the filters during one research, but perhaps one, two or maybe three of them.

and other entries that were added by the respondents. Ďurčo (2002: 53) was asking respondents to write down their most favorite proverbs and sayings. The modernized version also includes a question about the least favorite ones and question about units from other languages that respondents use when they are speaking Slovene. Another question which was already present in Ďurčo's survey was to write down humorous proverb actualizations or jokes which include a paremiological unit. Along with a big number of word-plays, anecdotes etc., there is also quite a big set of anti-proverbs added by the respondents. Other comments on the questionnaire are also possible.

2.2 The progress of the survey

The questionnaire was launched on 21st April 2012. To sum up, it contains 918 units, which have to be marked in one out of five possible ways. Nevertheless, the most enthusiastic respondents completed the questionnaire in one single day or even five hours. However, most of them prefer to deal with it from time to time, even as a sort of a free time activity. The font of the webpage was intentionally made as easy and as clear as possible. We already mentioned the advantage because of the technical improvements which are now possible thanks to development of information technologies. On the other hand, new possibilities are brought by the power of social networks and therefore offer a bigger number of respondents and also their feedback after they have already fulfilled the questionnaire. Some respondents are sending lists of paremiological units even months after they finished the core of the survey.

First expectations for the number of participants per month were soon exceeded. After less than one month, the number of the participants who have completed the questionnaire reached 100 (on 6th May 2012) and after exactly one month it doubled to 200 with rapid growth. Later on, the number of the fulfilled surveys started to decrease during the summer period, but nevertheless, in the second half of August 2012 the number of fully completed surveys reached 300. This number is a good basis to freeze the data on a separate page. With some additional filters it is possible to establish the paremiological optimum of the Slovene language in September 2012.

The invitations were sent in three major phases – first to colleges, than to a quite big number of Slovene associations and clubs dealing with culture, language and pedagogical activities, student clubs and organizations of Slovenes living abroad (emigration, Diaspora, etc). The third step was sending the invitations to a long list of email addresses, which included almost all email addresses of Slovene primary and secondary schools. For this goal, online list of primary schools and secondary schools with email contacts <http://www.dijaski.net/> and Online database of Slovene associations AJPES <http://www.zdos.si/register-drustev/> were used. The peak of the first phase brought 76 registrations per day, the peak of the second phase 46 and the third one 62. The oldest respondent so far was born in 1933, the youngest in 2000. A week after the internet survey was available online a Facebook profile *Pregovori Slovenski jezik* (Proverbs Slovene language) was created. With modern social networks it is possible to popularize information about the research. The profile has around 1700 friends already and it provides a lot of feedback and new phraseological material as well as its perception in the eyes of native speakers in form of chats, statuses, comments and other interlinking.

In order to establish the list of most known proverbs, and further on, the paremiological optimum, special databases should be frozen on a separate page at a certain moment and according to the main criteria. Since it would also be interesting to investigate the difference between the data, with or without using some of the above mentioned filters (mainly the filter of 100 % individual questionnaires fulfilled), there should be (at least two) different versions of the databases saved in one particular moment. For example, when 305 respondents fulfilled the whole questionnaire, it is also interesting to see the list of the most well-known proverbs according to the answers also of those 1091 people, who never fully completed the survey (1396 × 918) and also the list of the most well-known proverbs only according to the answers of those 305, who completed the survey (305 × 918).

3 Paremiological optimum of Slovene language and further research

3.1 The online questionnaire as a part of doctoral research – the paremiological optimum

Proverb familiarity in both languages will be, later on, confronted with the data on the frequency of the units shown by the Slovene corpus *FidaPLUS* and the Slovak national corpus (*Slovenský národný Korpus*). A paremiological optimum as defined by Ďurčo (2006: 17) is a larger set of most well-known paremiological units, arranged as a correlation between the familiarity based on an empirical demographical research and their frequency based on the corpus research. By a comparison of this correlation, we will follow an important conclusion (Grzybek/Chlosta 2008: 102), that the old and often misunderstood phenomenon – the popularity of paremiological units – must be divided into two different categories – frequency and familiarity, which depend on each other in a form of a regulating circle.

Later on, investigation will allow us to compare semantic and construction differences between individual paremiological units from Slovene and Slovak optimum. An interesting point in the research will represent the cases, in which there is a referential gap between the paremiology of those two Slavic languages and the situation of an idiosyncrasy as described by Ďurčo. Slovak and Slovene equivalents will be compared from both optimums according to Ďurčo's latest typology of phraseological equivalents (Ďurčo 2011) and some diasystematic differences introduced by Ďurčo (2012: 376). The paremiological databases created as a result of demographical research and structured as described above will allow us analyze diasystematic differences between Slovene paremiology and Slovak paremiology. This will shed some light on diafrequent, diachrone and diamedial as well as partially also on dianormative and diaevaluative differences.

3.2 Possibilities for further research based on the paremiological optimum and other data, provided by the questionnaire

In the future, the results of the online questionnaire can serve as a source for different paremiological research and different phraseodidactical or phraseographical purposes. The optimum itself can later be a basis for further research and different paremiological minimums as Ďurčo (2006: 3) suggests. The idea of paremiological minimum can also be connected with the idea of comparison of minimums in different languages. Each minimum (and possibly also optimum) is merely a projection as Mokienko (2012: 83) claims. Nevertheless it is a good working field for modern paremiology. It will allow us to compare Slavic languages among each other from the point of view of genetic connection on the one hand and similar or different cultural influences on the other. A comparison with minimums from non-Slavic languages would of course also be an appealing chance if such optimums from the same methodological basis would be established.

An important task will be to show how the familiarity of paremiological units depends on demographical factors. We should probably focus mainly on the age factor, which was already proven to be very significant by Ďurčo (2002: 53f.) and Grotjahn/Grzybek (2000: 122). What is also interesting is the influence of the factor of education (ibid.). The active usage of Slovene paremiological units still have to be the subject of a research carried out from the point of view of the variant forms given by the respondents. The variant forms for each unit should be analyzed and compared with the zero variant given in the main dictionary of standard Slovene (SSKJ). This aspect can, due to the structure of the online survey database and filters, also be seen from the point of view of Slovene dialect regions or other demographical factors (age, education, sex). Another interesting question would be to compare the optimum for different region dialect groups, as well as of those respondents who are migrating between two of them.

A more detailed analysis of additional answers will also be needed. Let us just name some of them. Where do new units come from? What is the percentage of units formed from commercial, political slogans or famous film quotes etc.? What does the hierarchy of the most favourite and the least favourite proverbs show us? How does the structure of paremiological units borrowed from foreign languages look like according to the variety of criteria? Which language (auto) stereotypes are shown by the answers and added opinions? We are not only interested in the question of which additional units, added by the respondents as proverbs and sayings, are indeed phraseological units. We can turn this question around and ask ourselves which units in the language are nowadays also marked as proverbs and sayings by Slovene speakers. Do these answers show us some tendencies and to what degree do respondents respect the border between paremiological units and non-textemes?

The online questionnaire will, of course, still be open for more respondents after the making of the paremiological optimum of Slovene language (in September 2012). This means that there is a possibility to make optimum 2.0 (probably in 2014) according to Mokienko's vision of a dynamic system (a sphere of known units) instead of the minimum as a static list (Mokienko 2012: 83). Comparison of the results would help to estimate the significance of a bigger number of respondents. It will be possible to observe how this affects the differences between the list of the most well-known proverbs made on the basis of all the 100 % fulfilled questionnaires on the one hand and of the whole amount of answers (including respondents, who did not finish it) on the other. The same could be done with the lists of the paremiological units marked as most used by the respondents, from 2012 and 2014.

References

BOJC, Etbin, 1987: *Pregovori in reki na Slovenskem*. Ljubljana: Državna založba Slovenije.

ČERMÁK, František, 2003: Paremiological Minimum of Czech: The Corpus Evidence. Burger, Harald/Häcki Buhofer, Annelies/Greciano, Gertrud (eds.): *Flut von Texten – Vielvalt der Kulturen. Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie*. Baltmannsweiler: Schneider Verlag Hohengehren. 15–31.

ĎURČO, Peter, 2002: K výskumu súčasnej živej slovenskej paremiológie. Jozef Mlacek (ed.): *Studia Academica Slovaca, 31: Prednášky XXXVIII. letnej školy slovenského jazyka a kultúry*. Bratislava: Stimul. 51–60.

ĎURČO, Peter, 2004: Slovak Proverbial Minimum: The Empirical Evidence. Földes, Csaba (ed.): *Res humanae proverbiorum et sententiarum. Ad honorem Wolfgangi Mieder*. Tübingen: Narr. 59–69.

ĎURČO, Peter, 2006: Methoden der Sprichwortanalysen oder Auf dem Weg zum Sprichwörteroptimum. Häcki Buhofer, Annelies / Burger, Harald (Hrsg.): *Phraseology in Motion. Methoden und Kritik. Akten der Internationalen Tagung zur Phraseologie (Basel, 2004)*. Baltmannsweiler: Schneider Verlag Hohengehren. 3–20.

ĎURČO, Peter, 2011: Extensionale und intensionale Äquivalenz. Kübler, Natalie / Benayoun, Jean-Michel / Zouogbo, Jean Philippe (eds.): *Tous les chemins mènent à Paris Diderot. Actes du Colloque international de Parémiologie*. Baltmannsweiler: Schneider Verlag Hohengehren.

ĎURČO, Peter, 2012: Diasystematische Differenzen von Sprichwörtern aus der Sicht der kontrastiven Parömiografie. Steyer, Kathrin (ed.): *Sprichwörter multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömiologie*. Tübingen: Narr. 357–379.

ĎURČO, Peter / STEYER, Kathrin, 2012: Ein korpusbasiertes Beschreibungsmodell für die elektronische Sprichwortlexikografie. Kübler, Natalie / Benayoun, Jean-Michel / Zouogbo, Jean Philippe (eds.): *Tous les chemins mènent à Paris Diderot. Actes du Colloque international de Parémiologie*. Baltmannsweiler: Schneider Verlag Hohengehren.

GROTJAN, Rüdiger / GRZYBEK, Peter, 2000: Methodological Remarks on Statistical Analyses in Empirical Paremiology. *Proverbium* 17, 121–132.

GRZYBEK, Peter, 2008: Fundamentals of Slovenian paremiology. *Traditiones* 37/1, 23–46.

GRYZBEK, Peter / CHLOSTA, Christoph, 1995: Empirical and Folkloristic Paremiology: Two to Quarrel of to Tango? *Proverbium* 12, 67–85.

GRYZBEK, Peter / CHLOSTA, Christoph, 2008: Some Essentials on the Popularity of (American) Proverbs. McKenna, Kevin (ed.): *Festschrift on the Occasion of Wolfgang Mieder's 65th Birthday*. Burlington VT: University of Vermont. 95–110.

LOGAR, Tine, 1966: Slovenska narečja. *Jezik in slovstvo* 11/5, 134–140.

MAKAROVIČ, Marija, 1975: *Pregovori, življenjske resnice*. Ljubljana: Kmečki glas.

MLACEK, Jozef, 1983: Problémy komplexného rozboru prísloví a porekadiel. *Slovenská reč* 48, 129–140.

MLACEK, Jozef, 2001: *Tvary a tváre frazém v slovenčine*. Bratislava: Edícia Studia Academica Slovaca.

MOKIENKO, Valerij, 2012: Russisches parömiologisches Minimum: Theorie oder Praxis? Steyer, Kathrin (ed.): *Sprichwörter multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömiologie*. Tübingen: Narr. 79–99.

Paremiologická databaza: <https://data.juls.savba.sk/paremiografia>. Access 10. 8. 2012.

PAVLICA, Josip, 1960: *Frazeološki slovar v petih jezikih*. Ljubljana: Državna založba Slovenije.

PERMJAKOV, Grigorii L'vovich, 1989: On the Question of a Russian Paremiological Minimum. *Proverbium* 6, 91–102.

Večbesedni termini v turističnem terminološkem slovarju

Vesna Mikolič (Koper)

Abstract

The paper focuses on classifiying idioms (i. e. idiomatic expressions) by using selected examples which may be included in the Slovenian-English dictionary of tourism terminology. Chosen phrases were extracted from the list of frequently used multiword expressions. When arranging the entries, the distinction between a collocation and a phrase is very important, as only the latter is included as an entry (including idiomatic or non-idiomatic expressions), whereas terminological collocations are stated in the dictionary entry at one of the entry word. For example, *successful tourist season* as an example of a collocation of an adjective with the term that is a nominal phrase, in comparison with *high tourist season* as the dictionary entry. The above stated distinction was often not easily recognizable from the list of chosen phrases, therefore we had to analyse the actual use of the selected phrase in texts collected in Multilingual Corpus of Tourist Text (TURK) manually.

1 Namen prispevka

Tvorjenje strokovnega izrazja je ena temeljnih nalog razvite jezikovne skupnosti, ki svoj jezik uporablja na različnih sporazumevalnih področjih. Enako seveda velja tudi za slovenski jezik, vendar pa so opazne razlike med področji, ki hkrati z razvojem dosledno skrbijo tudi za razvoj slovenske terminologije, npr. informacijska tehnologija in mobilna telefonija (Stramljič Breznik 2003), in nekaterimi drugimi področji, ki predvsem pod vplivom globalizacijskega jezikovnega enotenja opuščajo razvijanje slovenske terminologije (Humar 2009: 42). Ker je bilo področje turizma z vidika terminologije dolgo deficitarno oziroma terminologija zaradi hitrega razvoja zelo dinamične turistične stroke ni bila izdelana, se še vedno pogosto pojavlja zmeda bodisi zaradi kopice sinonimnih izrazov bodisi zaradi različnih pomenskih razmerij med slovenskimi in prevzetimi izrazi, največkrat iz angleščine, tudi francoščine. Prav zato je bila potreba po terminološkem slovarju velika.

Seveda se je bilo kot pri vsakem terminološkem slovarju tudi pri turističnem slovarju potrebno soočiti z vprašanjem zasnove in načina oblikovanja. Logar (2009: 321) navaja tri faze v nastajanju terminološkega slovarja, in sicer:

opredelitev besedil, iz katerih bo pridobljena leksika; oblikovanje meril, na podlagi katerih bo leksika vključena v geslovnik; določitev podatkov, ki bodo vključeni v geselski članek.

Oblikovanje meril terminološkosti je še posebej pomembno pri razlikovanju večbesednih terminov od terminoloških kolokacij. Prav temu vidiku je posvečen prispevek, ki predstavlja postopek urejanja slovarja s poudarkom na urejanju stalnih besednih zvez kot potencialnih terminov za vključitev v *Turistični slovensko-angleški terminološki slovar*.[1]

2 Postopek urejanja turističnega terminološkega slovarja

Osnova za izdelavo slovarja je bil specializirani jezikovni korpus s področja turizma TURK <http://jt.upr.si/turk>, ki zajema 30.000.000 besed.[2] Iz korpusa, ki je bil oblikoslovno označen in lematiziran z orodjem TOTALE (Erjavec 2006), je bila avtomatsko izluščena lista besed (monogramov, tj. enobesednih enot) in besednih zvez (bigramov in trigramov, tj. večbesednih enot), urejena po padajoči pogostosti. V drugem koraku smo za vsak seznam pripravili nabor besed, ki naj jih program avtomatsko izloči, saj smo zanje lahko z gotovostjo predpostavljali, da niso kandidati za termine. Take besede so bile vezniki, predlogi, členki, naklonski izrazi, pomožni glagoli, zaimki in prislovi. Kljub pričakovani samostalniškosti terminov – terminološka stroka si je namreč edina, da so termini po besedni vrsti v glavnem samostalniški (Cabré in Estopà v Košir 2010: 7; Logar/Vintar 2008: 5–10; Vidovič Muha 2000: 310–328) – smo pred ročno analizo gradiva v naboru terminoloških kandidatov obdržali tudi avtomatsko izluščene glagole in pridevnike ter samostalniške besedne zveze.

[1] Ta slovenski razlagalni turistični terminološki slovar z angleškimi ustreznicami je nastal v okviru projekta *Turistični terminološki slovar* Znanstveno-raziskovalnega središča Univerze na Primorskem (financer ARRS, projektni partner STO, nosilka Vesna Mikolič, 2008–2011) in je prosto dostopen na specializiranem terminološkem portalu Termania <http://www.termania.net>. Slovar je rezultat timskega dela, pri katerem so sodelovali Vesna Mikolič, Maja Smotlak, Klara Šumenjak, Jana Volk, Mojca Kompara, Martina Rodela, Elena Šverko, Jernej Vičič.

[2] Korpus je nastal v okviru temeljnega raziskovalnega projekta *Večjezični korpus turističnih besedil – informacijski vir in analitična baza slovenske naravne in kulturne dediščine* istega izvajalca, financerja in iste nosilke v letih 2006–2008 (Mikolič idr. 2008). Gre za elektronsko zbirko turističnih besedil v slovenskem, italijanskem in angleškem jeziku, oblikovano na osnovi statistično relevantnih kriterijev in z nizom pomožnih jezikovnih orodij. Pri gradnji večjezičnega korpusa turističnih besedil je bilo nujno ločevanje med besedili glede vrste in oblike turizma (Mikolič/Beguš 2010: 233–235).

Termini so namreč lahko eno- ali večbesedni, med slednjimi prevladujejo dvo- ali tribesedne stalne besedne zveze vrstnih pridevnikov s samostalniki.

Za tako prečiščene sezname enobesednih in večbesednih enot je bilo nato potrebno oblikovati kriterije terminološkosti, pri čemer je bila pri obojih poleg pogostosti merodajna predvsem ustrezna opredelitev sporazumevalnega okolja, v katerem se lahko pojavi termin (Pearson v Logar/Vintar 2008: 9–10). Odločilen kriterij je tako bil pomen terminološkega kandidata za specifično strokovno/družbeno (pod)področje, ki se vključuje v področje turizma na eni strani in za turizem v najširšem pomenu besede na drugi strani. Pri tem smo se odločili, da bo slovar vseboval predvsem terminologijo turistične industrije ali turizma v ožjem pomenu besede, poleg tega pa tudi nekatere termine, ki izvorno prihajajo z drugih področij, npr. zgodovine, arhitekture, geografije, ekonomije, ki pa so s turizmom kot sestavljeno dejavnostjo neločljivo povezani in z vstopom na področje turizma pridobijo vsaj deloma specifičen pomen.[3] Vintar (Logar/Vintar 2008: 13) opozarja, da je terminološkost besedne zveze izrazito subjektiven pojem, odvisen od vsakokratnega uporabnika terminologije. Prav zato je bilo potrebno vse prečiščene sezname ročno pregledati. Pri tem smo izvedli dodatno korpusno poizvedbo: preverili smo pogostost besede in njeno besedilno okolje; besedna skica, ki jo nudi programska oprema SketchEngine <http://www.sketchengine.co.uk/>, nam je pokazala leksemske povezave, pomembno pa je bilo tudi posvetovanje s strokovnjaki z različnih turističnih (pod)področij.

Sledilo je oblikovanje geselskega članka, pri čemer se je bilo treba odločiti o vrstah informacij, ki bodo vključene v članek in o načinu njihovega vključevanja. Poleg standardnih slovarskih informacij, kot so npr. oznake za besedno vrsto, definicije in prevodne ustreznice v tujem jeziku (pri večjezičnih slovarjih), smo se tu odločali še za oznake, specifične za terminološki slovar (npr. kvalifikatorji za strokovno (pod)področje, terminološke kolokacije, sopomenke itd.). Izhajali smo iz dinamičnega razumevanja geselskega članka, pri katerem so nekatere kategorije obvezne, druge pa ponujene kot možnost dodatnega označevanja iztočnice, kadar je to potrebno (Mikolič/Beguš 2010: 237–238).

Z avtomatskim izpisom podatkov na vmesnik portala Termania smo dobili delno izpolnjen geselski članek, v katerem so bili izpisani naslednji podatki:

- iztočnica,
- oznaka besedne vrste (brez navajanja slovničnih posebnosti),
- kvalifikator za turistično zvrst,

[3] Prim. Hoffmannov model specifičnih in splošnih strokovnih izrazov (Košir 2010: 6–7 in 37–38).

- kvalifikator za področje družbene dejavnosti, ki se vključuje v turizem,
- terminološke kolokacije (iz korpusa TURK),
- prevodna ustreznica v angleškem jeziku (Amebisov Presis).

Za končno urejanje je bil potreben ročni pristop, kjer smo avtomatski izpis preverili in po potrebi dopolnili, v geselski članek pa vključili še naslednje elemente:

- normativno oznako (po potrebi),
- definicijo,
- dodatne kolokacije,
- sinonime (po potrebi),
- slovarske povezave (po potrebi),
- vire.

Pri večbesednih enotah je bilo ugotavljanje sistemskih lastnosti strokovnih besednih zvez (po Vidovič Muha 1988) v nasprotju s prepoznavanjem terminoloških kolokacij kot »zgolj« ponovljivih oz. ustaljenih sopojavitev dveh ali več leksikalnih enot v neposrednem skladenjskem razmerju (po Bartsch v Logar/ Vintar 2008: 12) še težavnejši postopek. Ker tudi terminološke kolokacije predstavljajo pomemben del strokovnega jezika, se v slovarskem geslu navajajo kot ilustrativno gradivo ob ustrezni terminološki iztočnici. Vendar pa je prav ločevanje terminoloških kolokacij od terminoloških stalnih besednih zvez, bodisi frazeoloških bodisi nefrazeoloških, od sestavljalcev slovarja zahtevalo posebno pozornost.

3 »Mehke« meje med terminološkimi kolokacijami ter frazeološkimi in nefrazeološkimi večbesednimi termini

Odločitev, katere večbesedne enote vključiti v slovar in katere ne, je seveda »mehka« oziroma v določeni meri subjektivna. Kot opozarjata Logar in Vintar (2008: 14), so »korpusi strokovnih besedil razrahljali več mej: mejo med terminološko in neterminološko leksiko, mejo, do katere še govorimo o večbesedni poimenovalni enoti in čez katero je že prostor kolokacij, ter mejo, ki določa, ali gre za termin področja, ki ga obravnavamo, ali ne«. Prav programska oprema SketchEngine je še posebej koristno orodje za ugotavljanje pogostosti določenih skladenjskih vzorcev in analizo ustaljenosti besednih zvez, kar lahko pomaga pri razlikovanju med kolokacijami in večbesednimi termini.

Na izbranih primerih besednih zvez naj pokažemo na zahtevnost procesa urejanja stalnih besednih zvez kot potencialnih terminov za vključitev v slovar. Vzeli bomo primer dvobesednih zvez z zelo pogostim pridevniškim določilom *turistični*. V priloženem seznamu (slika 1) so s krepko pisavo zapisani termini, ki smo jih v slovar uvrstili kot iztočnice, bodisi kot terminološko stalno besedno zvezo (obe besedi pisani s krepko pisavo), bodisi kot kolokacijsko besedno zvezo pri pridevniškem terminu *turistični* (krepko je pisan samo pridevnik), bodisi kot kolokacijsko besedno zvezo pri samostalniški iztočnici (krepko je pisan samo samostalnik).

turistična agencija ● ***turistična borza*** ● ***turistična dejavnost*** ● ***turistična destinacija*** ● ***turistična geografija*** ● ***turistična industrija*** ● ***turistična infrastruktura*** ● ***turistična kapaciteta*** ● ***turistična kmetija*** ● ***turistična konkurenca*** ● ***turistična motivacija*** ● ***turistična nadgradnja*** ● ***turistična*** *oddaja* ● ***turistična panoga*** ● ***turistična ponudba*** ● ***turistična potreba*** ● ***turistična pisarna*** ● ***turistična prireditev*** ● ***turistična privlačnost*** ● ***turistična rekreacija*** ● ***turistična sezona*** ● ***turistična soba*** ● ***turistična storitev*** ● ***turistična taksa*** ● ***turistična zmogljivost*** ● ***turistična znamenitost*** ● ***turistična zveza*** ● ***turistični agent*** ● ***turistični aranžma*** ● ***turistični cilj*** ● ***turistični delavec*** ● ***turistični izdelek*** ● ***turistični kraj*** ● *turistični* ***krožek*** ● ***turistični management*** ● ***turistični motiv*** ● ***turistični namen*** ● ***turistični obisk*** ● ***turistični paket*** ● ***turistični*** *pomen* ● ***turistični produkt*** ● ***turistični program*** ● ***turistični proizvod*** ● ***turistični ponudnik*** ● ***turistični*** *razvoj* ● ***turistični resurs*** ● ***turistični sejem*** ● ***turistični*** *sistem* ● ***turistični spominek*** ● ***turistični tehnik*** ● ***turistični tok*** ● ***turistični trg*** ● ***turistični vodič*** ● ***turistični vodnik*** ● ***turistično*** *delovanje* ● ***turistično društvo*** ● ***turistično gospodarstvo*** ● ***turistično jahanje*** ● ***turistično naselje*** ● ***turistično območje*** ● *turistično* ***podjetje*** ● ***turistično posredništvo*** ● ***turistično povpraševanje*** ● ***turistično trženje*** ● ***turistično tržišče*** ● ***turistično vodenje***

Slika 1: Primeri frazeoloških in nefrazeoloških večbesednih terminov in terminoloških kolokacij s pridevnikom *turistični.*

Pri ugotavljanju, ali gre pri besednih zvezah za kolokacijske zveze, za nefrazeološke stalne besedne zveze ali za frazeme v ožjem pomenu besede (idiome), smo izhajali iz frazeoloških kriterijev ločevanja besednih zvez. Po Kržišnikovi (1994: 92) frazeološke enote v ožjem pomenu besede opredeljujejo vse štiri osnovne lastnosti: nadbesednost, stalnost, idiomatičnost in ekspresivnost; stalne besedne zveze, ki ne izpolnjujejo vseh teh kriterijev, pa sodijo v frazeologijo v širšem pomenu besede. Predvidevali smo, da bo frazemov v tem ožjem smislu tudi med turistično terminologijo malo, saj ekspresivni in preneseni pomeni, ki so pogosto povezani z idiomatičnostjo, niso osnovna lastnost terminov, ki težijo k enoznačnosti in nevtralnosti. Na drugi strani pa smo morali kot termine upoštevati številne stalne besedne zveze, saj je prav stalnost oz.

ustaljenost značilna tudi za strokovne besedne zveze, ki se pretežno po tej lastnosti ločujejo od terminoloških kolokacij.

Če v besednih zvezah nismo prepoznali niti nadbesednosti, niti stalnosti oblike in pomena in niti idiomatičnosti, je najverjetneje šlo za terminološko kolokacijo, kot v naslednjih primerih dvobesedne zveze: *turistična oddaja*, *turistični pomen*, *turistični razvoj*, *turistični sistem*, *turistično delovanje*. Kot ugotavljata Logar in Vintar (2008: 7–8), pridevniki samostojno sicer niso termini, vendar pa vrstni pridevniki skupaj s samostalniki lahko tvorijo stalne besedne zveze. Tako smo v turističnem terminološkem slovarju tiste (izključno) vrstne pridevnike, ki na področju turizma nastopajo pogosto, in to tako v terminoloških stalnih besednih zvezah kot v kolokacijskih zvezah, uvrstili kot samostojne iztočnice. Pridevnik *turistični* je gotovo eden takih pridevnikov; primere, ko besedne zveze ne zadoščajo kriterijem terminološke stalne besedne zveze, pač pa gre za kolokacije, smo tako uvrstili med ilustrativno gradivo k terminu *turistični*, v seznamu na sliki 1 pa so v teh primerih s krepko pisavo izpisani samo pridevniki.

Drugače je pri kolokacijskih besednih zvezah, kjer se med turistične termine, ki so sicer primarni termini drugih strokovnih področij, uvrščajo lahko že sami samostalniki, tako npr. *krožek*, *podjetje*. Pridevnik *turistični* določi samo, da gre za področje turizma, sam pomen samostalniškega termina pa se ne razlikuje od pomena, ki ga ima na drugih področjih. Prav tako je tudi v turizmu pogosta tudi njihova samostojna raba. V teh primerih so iztočnice lahko samostalniki (v seznamu na sliki 1 so s krepko pisavo pisani samo samostalniki), besedna zveza s pridevnikom pa se kot terminološka kolokacija pojavi v ilustrativnem gradivu. Pearson (v Košir 2010: 8) trdi, da je termine, ne glede na področje, ki so mu prvotno pripadali, potrebno razumeti tudi kot del terminologije drugega področja, takoj ko postanejo del slovarja tega področja. Tako smo tudi mi kot terminološke kandidate upoštevali tudi nekatere termine, ki so primarni na nekem drugem področju in v okviru turizma ne spremenijo pomena, vsaj ne bistveno, pri tem je bil odločilen kriterij pogostosti rabe.

Nasprotno pa je pri nekaterih terminih drugih strok pridevnik *turistični* nepogrešljiv, saj s tem pridevniškim določilom samostalniški termini pridobijo drugačen pomen, kot so ga imeli v drugih strokah in postanejo kot stalna besedna zveza samostalnika z vrstnim pridevnikom del turistične terminologije: geografija: *turistična geografija*, ekonomija: *turistična borza*, *turistični management*, *turistični ponudnik*, *turistično gospodarstvo*, *turistično povpraševanje*, sociologija: *turistični tok*, izobraževanje: *turistični tehnik*.

Podobni so tudi vsi ostali primeri, v katerih smo prepoznali večbesedne termine, saj je šlo za stalne besedne zveze z nadbesednim in stalnim pomenom in obliko, kot npr. *turistična agencija*, *turistična ponudba*, *turistična pisarna*, pri katerih so na sliki 1 obe sestavini pisani v krepki pisavi. Pri prepoznavanju nadbesednosti je bilo odločilno mnenje strokovnjakov s področja turizma in njegovih (pod)področij.

Dokaj maloštevilni pa so frazeološki termini v ožjem pomenu besede, kjer bi poleg nadbesednosti in stalnosti lahko ugotovili tudi idiomatičnost in ekspresivnost. Tak je v zgornjem seznamu lahko *turistični paket* (iz neposrednih pomenov besede *paket* ne moremo izpeljati pomena, da gre za sestavljen proizvod, ki se na trgu ponuja po skupni ceni in je oblikovan tako, da zadovoljuje določeno potrebo in željo turistov), sicer pa k tem lahko dodamo še: *visoka turistična sezona* (iz neposrednega pomena 'visok' ne moremo izpeljati pomena, da gre za tiste mesece v letu, ko je turistični obisk največji), *vinska pot* (kjer gre za metaforični/metonimični prenos, saj *pot* nima samo denotativnega, pač pa tudi konotativni pomen, tj. gre za pot, ob kateri se ponuja možnost ogledov naravnih in kulturnih znamenitosti, vinskih degustacij in pokušine tradicionalnih jedi), *poletni kamp* (metaforični prenos, saj *kamp* nima samo denotativnega, pač pa tudi konotativni pomen, tj. ni nujno, da gre za nastanitev v kampu, pač pa je to posebna oblika preživljanja poletnih počitnic).

Podobna analiza je potrebna, ko se iz dvobesednega termina oblikuje tribesedni termin (samostalniška besedna zveza) ali pa kolokacijska zveza določenega leksema z dvobesednim terminom (pridevnika s samostalniško besedno zvezo). Tako ločimo tribesedne termine (trigrame), kot so: *primarna turistična ponudba*, *sekundarna turistična ponudba*, *dopolnilna turistična ponudba*, *celotni turistični proizvod*, *delni turistični proizvod*, od kolokacijske zveze pridevnik ali samostalnik + bigram: *razvoj turistične ponudbe*, *sodobna turistična ponudba*, *slovenska turistična ponudba*, *turistična ponudba kraja/ mesta*, *razvoj turističnega proizvoda*, *sestavljenost turističnega proizvoda*, *občinska turistična pisarna*.

Pogosto se dvobesedni termini rabijo tudi v skrajšani obliki kot enobesedni, npr. *(turistična) agencija*, *(turistični) aranžma*, *(turistična) zmogljivost*, *(turistična) znamenitost*, *(turistični) spominek*, čeprav s pridevnikom *turistični* pride do določenega pomenskega premika v pomenu besedne zveze, ki ga enobesedni termin ne vsebuje. Tako je npr. definicija spominka »predmet v spomin na kraj, znamenitost, ki je pogosto del turistične ponudbe in promocije«, definicija turističnega spominka pa »predmet, namenjen turistom v spomin na obiskani kraj, znamenitost, ki je del turistične ponudbe in promocije« (TURS). V turističnem diskurzu pojavljata obe različici, a je dvobesedni termin ožji in bolj natančen, povezan izključno s turistično dejavnostjo (na razlike v pomenu kažejo krepko pisani deli besedila).

4 Sklep

Iz predstavljenega procesa urejanja turističnega terminološkega slovarja so jasno razvidne prednosti korpusnega pristopa, po drugi strani pa strojno luščenje terminologije ne more v celoti nadomestiti ročnega pristopa, pač pa mora teči vzporedno z opisom pojmovnega sistema neke stroke oz. področja. Še posebej je to izrazito pri identifikaciji terminoloških stalnih besednih zvez, kjer je potrebna kvalitativna analiza medbesednih razmerij za razmejitev frazeoloških in nefrazeoloških stalnih besednih zvez od kolokacij, ki so kot pomemben del strokovnega jezika sicer dragoceno ilustrativno gradivo, jih pa ne uvrščamo med iztočnice terminološkega slovarja. Terminološke stalne besedne zveze na področju turizma so večinoma frazeološke v širšem pomenu besede, frazemi/idiomi so redki. Za prepoznavanje terminoloških besednih zvez je tako nedvomno najbolj merodajna opredelitev rabe termina v določenem sporazumevalnem okolju, razvidnem iz področnega korpusa, pri čemer rabo in pomen še dodatno opišejo področni strokovnjaki.

Literatura

ERJAVEC, Tomaž, 2006: *Multilingual tokenisation, tagging, and lemmatisation with totale*. 9th INTEX/NOOJ Conference, Belgrade, Serbia, June 1–3, 2006. Belgrade: Faculty of Mathematics, University of Belgrade.

HUMAR, Marjeta, 2009: Položaj in prihodnost slovenske terminologie in terminografije. Ledinek, Nina/Žagar Karer, Mojca/Humar, Marjeta (ur.): *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU, 41–46.

KOŠIR, Mateja, 2010: *Slovenska filmska terminologija v korpusu filmskih kritik. Magistrsko delo*. Nova Gorica: Univerza v Novi Gorici: Fakulteta za humanistiko.

KRŽIŠNIK, Erika, 1994: Frazeologija kot izražanje v »podobah«. Križaj Ortar, Martina/Bešter, Marja/Kržišnik, Erika: *Pouk slovenščine malo drugače*. Trzin: Different. 91–103.

LOGAR, Nataša, 2009: Korpusi v terminografiji: umik potrebe po introspektivni presoji. Ledinek, Nina/Žagar Karer, Mojca/Humar, Marjeta (ur.): *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU. 319–328.

LOGAR, Nataša/VINTAR, Špela, 2008: Korpusni pristop k izdelavi terminoloških slovarjev: od besednih seznamov in konkordanc do samodejnega luščenja izrazja. *Jezik in slovstvo* 53/5, 3–17.

MIKOLIČ, Vesna/BEGUŠ, Ana, 2010: Identifikacija terminov za turistični terminološki slovar. *Ann, Ser. hist. sociol.* 20/1, 233–240.

MIKOLIČ, Vesna / BEGUŠ, Ana / DUKIČ, Davorin / KODERMAN, Miha, 2008: Vpliv namembnosti korpusa na označevanje besedilnega gradiva za »Večjezični korpus turističnih besedil«. Erjavec, Tomaž / Žganec Gros, Jerneja (ur.): *Zbornik Šeste konference Jezikovne tehnologije, 16. do 17. oktober 2008, zbornik 11. mednarodne multikonference Informacijska družba – IS 2008, zvezek C (Informacijska družba)*. Ljubljana: Institut Jožef Stefan. 60–64.

STRAMLJIČ BREZNIK, Irena, 2003: Besedotvorna tipologija novonastalega besedja s področja mobilne telefonije. *Slavistična revija* 51 (Kongresna številka), 105–118.

VIDOVIČ MUHA, Ada, 1988: Nekatere jezikovnosistemske lastnosti strokovnih besednih zvez. Pogorelec, Breda / Sajovic, Tomaž / Počaj-Rus, Darinka (ur.): *XXIV Seminar slovenskega jezika, literature in kulture. Zbornik predavanj*. Ljubljana: Oddelek za slovanske jezike in književnosti Filozofske fakultete. 83–91.

VIDOVIČ MUHA, Ada, 2000: *Slovensko leksikalno pomenoslovje. Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Graph-based Analysis of Collocational Profiles

Piotr Pęzik (Łódź)

Abstract

The availability of large linguistic corpora has opened up new possibilities of studying the distributional characteristics of phraseological units. As corpus-based tools and resources such as Automatic Collocation Dictionaries (ACDs) become available, new methodological questions arise about their applications in phraseological research. This paper discusses selected aspects of generating and using ACDs in phraseological studies. In particular, graph-based methods of exploring and visualizing ACD entries are proposed to deal with the high-recall and sub-optimal precision of such resources. First, the HASK dictionaries of frequent English and Polish word combinations are described as specific examples of ACDs. Both the advantages and limitations of using these resources for phraseological studies are explained in terms of basic evaluation measures used for language processing systems. Next, the process of generating 'collocational graphs' for arbitrary sets of node words is explained. Methods of comparing and clustering large sets of words on the basis of their estimated collocational coincidence are proposed and applied to Polish and English nationality adjectives. Finally, a method of cross-linguistic comparison of collocational profiles is introduced and applied to a selection of wartime-related collocations. It is claimed that these exploratory techniques can be used to aid quantitative and qualitative phraseological studies based on collocational profiles recorded in ACDs.

1 Introduction

The formation of lexical collocations is a highly productive phraseological process resulting in the emergence of hundreds of thousands of textually recurrent combinations which remain largely unaccounted even in dedicated phraseological dictionaries. It has been observed that among the wide variety of functional and structural types of phraseological units lexical collocations present one of the biggest challenges for descriptive phraseology (Mel'čuk 2001). A single word may enter hundreds of combinations which are part and parcel of the ambient formulaic inventory of a given language. Although traditional dictionaries have recently improved in their recognition of phraseology, they can offer only a limited coverage of the vast richness of formulaicity in language due to obvious space limitations and the fact that they are compiled manually and with high quality requirements, which often leads to the exclusion of some of the more spurious word combinations. For one thing, there is a growing realization that investigations of culture- and language-specific formulae, conventional metaphors, idioms, memes and stereotypes as they are reflected in phraseological patterning should not be limited to dictionaries, personal introspection or anecdotal evidence only. Large language corpora are now recognized as essential sources of empirical data for lexicographic and phraseological research. Paradoxically, as corpora grow in size, thus solving the problem of data scarcity, they may create the opposite problem of data deluge in that standard corpus queries for frequent lexical items result in thousands of concordance lines, which linguists may find impossible to analyze consistently. For instance, let us assume we are interested in studying the typical collocations, multiword terms and other types of phraseological, formulaic and onomasiological units formed by a small set of three nationality adjectives: *Greek*, *French* and *German*. We might expect these adjectives to tend to pre-modify some nouns more frequently than others in large reference language corpora. Many such combinations might owe their statistical significance mainly to a distinctive range of language- and culture-specific meanings that they are used and re-used to denote. *Greek philosophers*, *French window*, *German reunification*, *Greek sculpture*, *French Crown* and *German troops* are all combinations which bear traces of the long history of the geographic, cultural, literary, political and linguistic contacts which speakers of English have made both directly and indirectly with other European nations and cultures. To use Wierzbicka's paraphrase of Wittgenstein's famous observation, each of these simple phrases may turn out to be "a whole cloud of culture condensed in a drop of phraseology" (Wierzbicka 2007). Condensing large clouds of corpus evidence into single drops of phraseology can be problematic, however. A simple query for combinations of the adjective *French* preceding a noun in the British National Corpus (BNC) results in more than 1400 concordance lines. *German* as an adjectival pre-modifier of nouns occurs over a thousand times, while *Greek* pre-modifies nouns at least 200 times in the BNC. This means that a comprehensive BNC-based study of the nominal collocates of those three adjectives would require the manual scrutiny of nearly 3000 concordances. Given that we are now well into the era of "giga-corpora" (corpora containing billions of word units – an order of magnitude larger than the BNC), it becomes clear that new methods of aggregating and exploring large corpus datasets are becoming necessary in those aspects of phraseological research which focus on the investigation of "ambient" combinatorial characteristics of lexical and grammatical units.

2 HASK AUTOMATIC COLLOCATION DICTIONARIES

Automatic collocations dictionaries (ACDs) are corpus-derived electronic databases of recurrent word combinations (Kigariff/Rychly 2010). They can be generated with positional or relational collocation extraction methods (Evert 2005). Positional ("knowledge-poor") methods generally rely on the computation of the statistical significance of word frequencies within a predefined window. Relational approaches apply morphological and syntactic filters in the process of collocation extraction. It has to be noted that the term automatic *collocations* dictionary is a bit of a misnomer in that ACDs may contain a wide variety of phraseological units (PUs), which most functional and structural phraseological typologies would not classify as lexical or grammatical collocations, cf. Burger (1998).

The following discussion of ACDs is based on two dictionaries of Polish and English word combinations (HASK) extracted automatically from the *National Corpus of Polish* (Przepiórkowski et al. 2012) and the *British National Corpus* (BNC 2001) respectively. The HASK dictionaries were compiled using a hybrid positional-relational collocation extraction method. First, a number of syntactic patterns were defined for combinations of nouns, adjectives verbs and adverbs co-occurring in predefined positional contexts. Examples of these patterns include adjectives immediately preceding nouns (potential adjectival modifiers) and verbs co-occurring with nouns within a window of 4 words (covering a variety of verb argument structure elements). For each such combination a number of association scores are computed, including *t*-score, mutual information, chi-square, log-log, Dunning's log-likelihood (Dunning 1993) and Dice score. The distribution of each combination is also measured for commonness using Julliand's dispersion (Juilland/Brodin/Davidovitch 1971) and Average Waiting Time coefficients (Savický/Hlavácová 2002). Users of the dictionary can dynamically define a conjunction of threshold values for collocations of a given entry to be regarded significant. The English version of the HASK dictionary currently contains over 150 000 entries with a total of 2.8 million recurrent word combinations. The Polish version contains more than 92 000 entries and 5,3 million word combinations. Both dictionaries are available through online user interfaces hosted at pelcra.pl/hask. Users can browse, download and visualize phraseological profiles as collocational graphs for single and multiple entries. Apart from the statistical descriptors, a complete list of concordances is available for each combination included in the dictionary, thus enabling the manual validation of the dictionary.

3 Evaluation of ACDs

One clear benefit of ACDs over traditional dictionaries is their hyper-textual access structure (Svensén 2009); entries can be cross-linked not only with each other but also with the underlying usage examples and concordances. A less obvious advantage of using ACDs in phraseological studies lies in their recall of specific types of phraseological and formulaic units. The higher coverage of ACDs usually comes at the expense of their precision; in contrast traditional phraseological dictionaries ACDs rely on simple distributional and low-level linguistic criteria in extracting PUs and they tend to contain more erroneously identified items.

The exact evaluation of the precision and coverage of ACDs vis-à-vis traditional dictionaries is problematic. Formulaicity and idiomaticity can be viewed as continuous functions with intensity levels ranging from pure idioms and fixed terminological compounds to weak collocations and creative variants of otherwise established phraseological units. This continuity is reflected in the different terms used to describe degrees of idiomaticity. For example, (Bolinger 1979) distinguishes between different degrees of "automatic" vs. "propositional" language, (Van Lancker/Kempler 1987) describe the "reflexive" vs. "novel" continuum of formulaicity, while Sinclair (1996) defines the "idiom" vs. "open-choice" principles as continuous tendencies. It is therefore impossible to put an exact figure on the number of recurrent collocations and other flexible types of phraseological units tending towards the open-choice end of the phraseological continuum.

The inherent fuzziness of formulaicity as a categorical feature does not prevent us from defining a formal model of evaluating the coverage and accuracy of phraseological dictionaries. In fact such models have been defined for the task of automatic multiword unit extraction, e. g. (Daille/Gaussier/Langé 1994). Similarly to many other areas of natural language processing and information retrieval (Manning/Raghavan/Schütze 2008) the evaluation of phraseology extraction systems can be measured in terms of "precision" and "recall". Per-entry "recall" of a phraseological dictionary can be defined as the number of relevant PUs of a given type identified in the entry divided by the number of all combinations of a given type found in a language or language variety. The phraseological "precision" of a dictionary entry is the number of phraseological units correctly identified divided by the total number of PUs of a given type included in the entry. A correctly identified PU constitutes a true positive (TP). A spurious phrase found in a corpus and correctly excluded from the dictionary entry is considered to be a true negative (TN). PUs which in some respect are relevant and yet were not included in the dictionary entry can be regarded as false negatives (FN), whereas combinations erroneously included in the entry are false positives (FP). This means that we can define precision (P) and recall (R) more formally as: P = TP/(TP+FP) and R = TP/(TP+FN).

Item	Collocate	f	t	Item	Collocate	f	t
1a	bright	34	4.66	5f	sardonic	26	5.04
1b	broad	46	5.85	5g	wry	121	10.95
1c	wide	30	2.81	6a	sad	20	3.49
2a	faint	85	9.01	7a	shy	14	3.37
2b	thin	29	4.19	8a	apologetic	15	3.8
2c	wan	15	3.83	8b	sheepish	4	1.95
2d	weak	24	3.96	9c	encouraging	12	2.91
3a	beatific	6	2.44	9d	indulgent	5	2.16
3b	cheerful	13	3.21	9e	reassuring	18	4.12
3c	dazzling	21	4.45	10a	polite	20	4.15
3d	happy	21	1.3	11a	beguiling	3	1.68
3e	radiant	16	3.92	12a	ready	18	1.19
3f	sunny	6	1.92	13a	fixed	9	2.15
3g	warm	44	5.41	13b	forced	6	2.18
4a	charming	33	5.44	14a	supercilious	2	–
4b	gentle	27	4.49	15a	conspiratorial	6	2.4
4c	sweet	28	4.5	15b	knowing	21	4.26
4d	winning	11	3.01	16a	grim	26	4.84
5a	arch	3	1.55	17a	humourless	9	2.97
5b	disarming	13	3.59	17b	mirthless	7	2.63
5c	enigmatic	12	3.36	18c	crooked	23	4.71
5d	mocking	11	3.28	18c	lopsided	4	1.97
5e	rueful	33	5.71	19a	toothless	4	1.94

Table 1: Coverage od adjectival collocates od the noun *smile* in the ODCSE.

As already stated, exact numbers of valid collocations in a language are impossible to estimate and therefore, in practice, the true recall of any dictionary remains unknown. Nevertheless, the concepts of recall and precision are helpful in explaining the motivation behind compiling ACDs. In general, ACDs are characterized by a higher recall rate for specific types of collocations they were designed to cover, with the typical trade-off being the relatively lower rate of precision. Let us illustrate this general observation with a specific example. Table 1: Coverage od adjectival collocates od the noun smile in the ODCSE.shows the coverage of adjectival collocates of the noun *smile* in the Oxford Collocations Dictionary for Students of English (ODCSE) (Lea/Crowther/Dignen 2003). Forty-six such combinations are identified and grouped into 19 semantic classes in the dictionary entry for *smile* as a noun. Columns f and T in the table list the plain BNC counts and t-scores for each collocate included in the dictionary.

Since the dictionary entry arguably contains few or no false positives, its precision can be said to be close to 100 %. However, the recall of the entry is far from complete. If we simply extract all sequences of adjectives immediately preceding the noun ‘smile’ from the BNC and keep those that occur 3 or more times, we will get a superset of the adjectival collocates identified in ODCSE. Sorting this superset in descending order of *t*-score values and trimming it at the score level of t=1 results in a set of 190 adjectival collocates of the noun *smile*. [1] Arguably, many of those collocates could have been included in the entry of a learners’ dictionary, even if the dictionary was not meant to provide an exhaustive coverage of lexical collocations. For example, combinations such as *big smile* (43 occurrences), *small smile* (83), *warm smile* (42) and *slow smile* (54) are all much more recurrent in the BNC than most of the collocations included in the ODCSE entry for the word *smile*. Given their relatively high frequency, it would make sense to include these items in a dictionary targeted at learners of English, which is why we can regard them as false negatives. On closer inspection, the recall of this particular entry in ODCSE proves to be much lower than that of the HASK dictionary. On the other hand, the precision of HASK and any other ACD for that matter is less than perfect due to occasional false positives resulting from the use of automatic collocation extraction methods. This increase in recall may lead to the abovementioned problem of “data deluge”, especially when simultaneously analyzing collocations of different lexemes. In the remaining part of this paper, I present data exploration methods for large sets of collocations developed specifically to address the problem of phraseological data deluge. The use of these methods

[1] See <http://pelcra.pl/hask_en/browser.do?eh=8222b58fe249a1d06e7413f7719cde87>.

is illustrated with examples of profiles extracted for a set of Polish and English nationality adjectives.

4 Analyzing collocational graphs of nationality adjectives

In many languages, the nominal collocates of nationality adjectives reflect the long history of contacts and transfers between nations, cultures and languages. Parallel to their denotative functions, such word combinations are often loaded with pragmatic and attitudinal meaning. For example, the textual recurrence of phrases such as *niemiecka solidność* (a Polish cliché for ‘high German quality’) or *French farmers*[2] is conducive to the development and enforcement of positive and negative stereotypes about representatives of the respective nations. Even if seemingly direct equivalents of such collocations, terms, idioms and multiword naming units can be found in two or more languages, their stability, sentiment and relative share in the entire set of collocates for a given nationality adjective varies considerably across languages.

Colosaurus is one of the tools available through the HASK web application, which can be used to compare sets of words of the same morphological category in terms of their collocates. For example, by entering the following five nationality adjectives *Greek*, *French*, *German*, *Russian*, *Polish*, we can obtain a collocational graph similar to the one shown in Figure 1.

The input words serve as “seed vertices” $\{v_{s1} \ldots v_{sn}\}$ around which a collocational graph is built. Each collocate of a given type is represented as a “collocate vertex” $\{v_{c1} \ldots v_{cm}\}$ and has at least one edge connecting it with a seed vertex. The degree (number of edges) of a seed vertex may be zero. The edges of a collocational graph are weighted in that edge e_1 in v_{c1}-e_1-v_{s1} can have a set of weights representing the strength of the association of v_{s1} and v_{s2}, which is different from the set of weights of e_2 in v_{c1}-e_2-v_{s2}. For example, in the graph shown in Figure 1 the path $v_{s\text{-German}}$-$v_{c\text{-invasion}}$ over edge e_1 has weights {f=35, t-score=25, R=27,..}, whereas the path between $v_{s\text{-Russian}}$-e_2-$v_{c\text{-invasion}}$ has weights {f=17, t-score=3.6, R=14,…}. Values of the weights are represented in visualizations of collocational graphs by the width of the respective edges. Table 2 shows a second-order collocational adjacency matrix of five nationality adjectives in which each path between

[2] There seems to be a tendency in the British press to describe French farmers as troublemakers, which leads to the attachment of a certain non-compositional, attitudinal effect to this phrase.

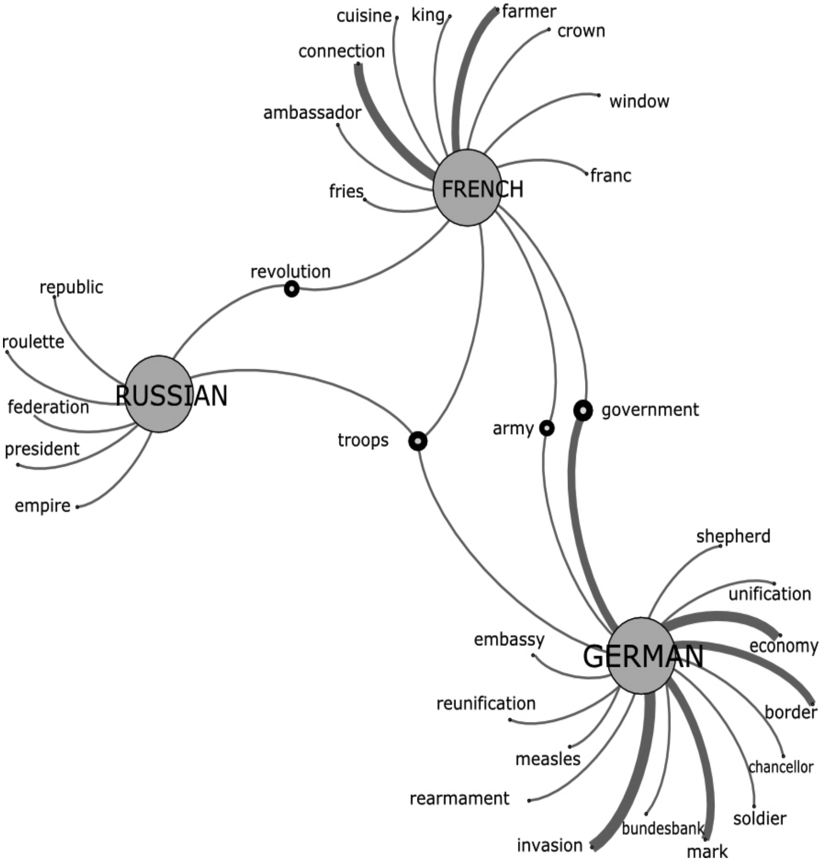


Figure 1: A simplified collocational graph for a set of English nationality adjectives extracted from the BNC.

two seed vertices such that v_{s1}-e_1-v_{c1}-e_2-v_{s2} is counted as 1. The number of nominal collocates shared by the adjectives *German* and *Polish* is 72, while the number of nominal collocates shared by the adjectives *Polish* and *Russian* is 60. The total number of nominal collocates for each adjective is given in the principal diagonal of the matrix.

$$\left(\begin{array}{cccccc}
 & German & Greek & Russian & French & Polish \\
German & 564 & 79 & 183 & 321 & 72 \\
Greek & 79 & 197 & 56 & 108 & 29 \\
Russian & 183 & 56 & 310 & 190 & 60 \\
French & 321 & 108 & 190 & 716 & 71 \\
Polish & 72 & 20 & 60 & 71 & 115
\end{array}\right)$$

Table 2: Second order collocational adjacency matrix for five nationality adjectives.

The number of collocational vertices included in the graph depends on the minimum criteria of collocability used in the user query. The graph matrix shown in Table 2 was generated using the cut-off parameter of at least three occurrences of an adjective immediately preceding a noun. Thus, in the resulting combinations the seed adjectives are usually adjectival pre-modifiers of their nominal collocates. Such a matrix of plain counts of shared collocates can be analyzed with a view to identifying significant differences between collocates sets of the respective seed nodes. For example, we may want to know the significance of the differences in the counts obtained for the nominal collocate sets of the adjectives *Russian*, *German* and *Polish*.

	German	**Russian**
Polish	72	60
Other	492	250

Table 3: 2×2 Contingency table for shared collocate counts.

Applying both the Fisher's exact and chi-square tests for the contingency table shown in Table 3 yields a very low (p=0.01) probability of these path counts being randomly generated, assuming that a nominal collocate of *Polish* is equally likely to be shared by the adjectives *German* and *Russian*. In other words, the relatively greater number of nominal collocates shared by the adjectives *Russian* and *Polish* seems to be non-random. For one thing, there are 9 nominal collocates in the BNC shared by the adjectives *Russian* and *Polish*, which occur at least 3 times and are not used with the adjective *German*. A closer look at the underlying concordances is enough to suspect that at least three of them, i. e. *peasantry*, *noble* (as in *Polish nobles*) and *nobility* may be indicative of the relatively extensive treatment of these social groups in British political and historical discourse about Russia and Poland due to the the role played by gentry and peasantry in the history of both of these countries. Otherwise, the significance of the difference in the number of collocational paths linking these three seed vertices may be attributable to the high number of nominal collocates of *German*, which are not shared by the adjective *Polish*. Many of them are used to describe Germany as an important military power, e. g. *German aggression*, *German bombardment*, *German invaders*, *German machine-guns*, *German raid*, etc. Clearly, some of the World War II image of Germany has been conserved in these combinations. Also interesting are those nominal collocations, multiword terms and naming units which, in a given collocational graph, are unique for one of the seed vertices. For *Polish* they include *Polish anti-semitism*, *Polish Corridor*, *Polish Solidarity*, *Polish Szlachta* or *Polish Squadron*. All of these phrases ultimately reflect some of the regular aspects the image of Poland created in British historical discourse as far as it is represented in the BNC.

4.1 Weighted path matrix and graph clustering

The counts of paths shown in the collocational path matrix can be weighted to better represent the significance of the collocates they share. One possible formula of computing such weights is given below:

$$w_{v_i v_j} = \frac{\sum_m^n log(w(e_i) + w(e_j))}{deg(v_{si}) + deg(v_{sj})}$$

where $w(e_i)$ and $w(e_i)$ are weights on edges in m to n paths connecting a given pair of seed vertices v_{si} and v_{sj}: v_{si}-e_i-v_{cn}-e_j-$v_{sj.}$ The sum of such weights is divided by the sum of the degrees (i. e. number of collocational edges) of the seed vertices. The intuition here is that high association values for the individual collocates shared between seed vertices increase the overall score. The resulting degree-normalized matrix is shown below:

$$\left(\begin{array}{cccccc}
 & German & Greek & Russian & French & Polish \\
German & 1.31 & 0.32 & 0.58 & 0.73 & 0.32 \\
Greek & 0.32 & 1.37 & 0.31 & 0.38 & 0.21 \\
Russian & 0.58 & 0.31 & 1.28 & 0.55 & 0.4 \\
French & 0.73 & 0.38 & 0.55 & 1.38 & 0.28 \\
Polish & 0.32 & 0.21 & 0.4 & 0.28 & 1.09
\end{array}\right)$$

Table 4: Weighted normalized collcoational matrix.

The adjectives *German* and *French* have the score of 0.73, indicating a relatively high number of shared collocates with significant association scores. The weight computed for the collocates shared by the adjectives *Polish* and *Greek* is lower (0.21). By inverting such values, we can obtain a distance matrix, which can be used to visualize complex collocational graphs using clustering techniques. Figure 2 shows a hierarchical dendrogram generated from a distance matrix of seven nationality adjectives (*German*, *French*, *Greek*, *American*, *Spanish*, *Russian* and *Polish*. A relatively low number of shared collocates in the distance matrix (shown in Table 5) results in a high distance value. This is the case for the adjectives *Polish* and *American*, which only share 12 out the total of more than 400 collocates meeting the criteria used to generate the collocational graph.

$$\left(\begin{array}{cccccccc}
 & German & French & Greek & American & Spanish & Russian & Polish \\
German & 0 & 2.42 & 4.60 & 2.83 & 3.12 & 2.19 & 3.81 \\
French & 2.42 & 0 & 4.36 & 2.77 & 2.83 & 2.75 & 6.65 \\
Greek & 4.60 & 4.36 & 0 & 5.80 & 4.46 & 5.07 & 7.07 \\
American & 2.83 & 2.77 & 5.80 & 0 & 6.31 & 3.07 & 11.35 \\
Spanish & 3.12 & 2.83 & 4.46 & 6.31 & 0 & 3.17 & 5.08 \\
Russian & 2.19 & 2.75 & 5.07 & 3.07 & 3.17 & 0 & 3.08 \\
Polish & 3.81 & 6.65 & 7.07 & 11.35 & 5.08 & 3.08 & 0
\end{array}\right)$$

Table 5: Distance matrix for seven nationality adjectives.

So far we have only looked at collocational profiles extracted from the English version of the HASK dictionary. In the remaining part of this paper I discuss possible techniques of comparing Polish and English collocational profiles.

The adjectives *German*, *French* and *Russian* are closely clustered, while *Greek* and *Polish* are visualized as outliers in the dendrogram. This data exploration method is more suitable for large sets of seed vertices which could be difficult to intelligibly visualize in a collocational graph. Obviously, such aggregated visualizations can only serve as pointers to further, qualitative analysis. Nevertheless, they are helpful in exploring the phraseology of large sets of lexemes.

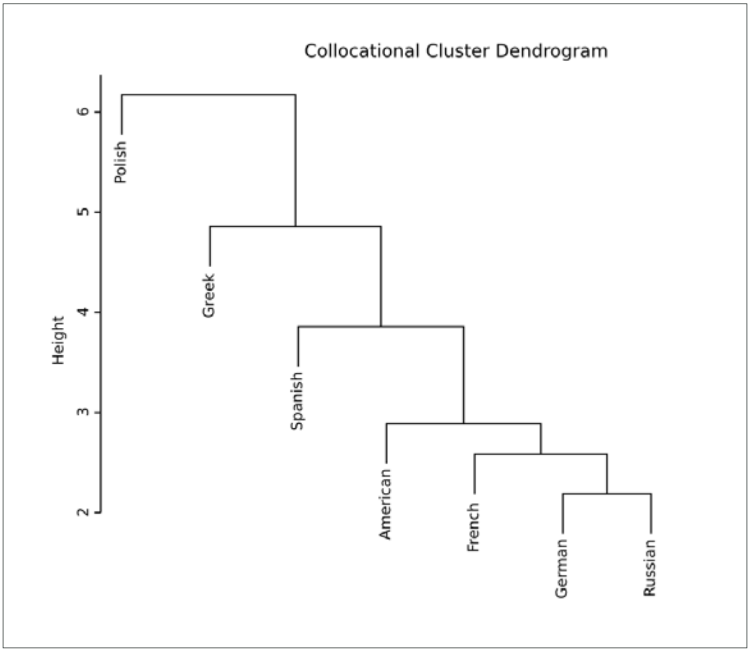


Figure 2: Collocational graph for seven nationality adjectives.

5 Cross-linguistic analysis of collocational profiles

The comparative analysis of collocational graphs extracted from ACDs yields interesting results about the topology of phraseological networks across different languages. Let us assume that there is a simple lexical function which maps each seed vertex in one language (L1) with a corresponding seed vertex in L2. For example, the seed adjective *Greek* in English may corresponds to the adjective *grecki* in Polish in that they are close lexical equivalents with similar sense distributions. Let us also assume that a similar mapping function exists for all the collocate vertices of the seed vertices, so that for example the Polish noun *nacjonalizm* is the lexical and distributional equivalent of the English adjective *nationalism*. In such a case we could expect the collocational graph generated for such sets of vertices from sufficiently large dictionaries to be isomorphic across languages. Quite obviously, this is rarely the case. Both *grecki* and *Greek* have their own "local grammars" (Sinclair 2010) and complex syntagmatic restrictions, which significantly reduce the scope of equivalence. For example, even though both *grecki Cypryjczyk* and *Greek Cypriot* are recurrent phrases in Polish and British texts, the latter is much more salient in British political and historical discourse thus reflecting the historical links Cyprus has with the UK. *Ryba po grecku* is a Polish fish dish name, which is idiomatic in that it has no obvious, extra-linguistic links to Greece. Needless to say, no direct equivalent of this phrase can be found in the collocational graph generated for the English adjective *Greek*. The English term *French window* is an analogous example of an idiomatic phrase, which has no direct Polish equivalent. The cross-linguistic analysis of collocational graphs can thus reveal certain phraseological tendencies which reflect language- and culture-specific priming of individual lexemes.

In the remaining part of this paper I take a closer look at the profiles of the English adjective *German* and the Polish adjective *niemiecki* with respect to the distribution of their nominal collocates. Table 6 shows 10 top-ranked collocates of the two adjectives extracted automatically from the BNC and NKJP corpora.

German			Niemiecki		
Collocate	**f**	**t-score**	**Collocate**	**f**	**t-score**
unification	89	9.29	język	1455	36.21
chancellor	87	8.59	**żołnierz**	711	24.96
army	103	7.84	**okupacja**	570	23.54
unity	68	7.58	**wojsko**	606	23.01
reunification	56	7.41	firma	905	22.3
soldier	72	7.31	mniejszość	492	21.42
embassy	58	7.24	**armia**	455	19.92
ideology	58	7.0	**oficer**	368	17.97
troops	63	6.79	marka	326	16.92
government	230	6.7	owczarek	247	15.61

Table 6: Top-scored nominal collocates of the adjectives *German* and *niemiecki*.

Both lists contain combinations which owe their significant recurrence to frequent references to the role of Germany in World War II. A closer look at the concordances underlying terms and collocations such as *German army*, *German soldiers*, *German ideology*, *German troops*, *niemiecki żołnierz*, *niemiecka okupacja*, *niemieckie wojsko*, or *niemiecka armia*, show that they are used predominantly in a WW2 context. On the other hand, there are quite a few recurrent multiword terms used to refer to post-war events, entities and institutions in German history. Some highly significant combinations found in the BNC refer to German reunification and German political institutions in a post-war context. Similarly, the Polish concordances obtained for phrases such as *mniejszość niemiecka* (*German minority*)[3], *niemiecka firma* or *niemiecka marka*[4] *show that they are used mainly in post-war contexts. The question one might ask is whether the differences in the proportions of WW2 vis-à-vis post-war phraseological and terminological units involving the adjectives niemiecki* and *German* are significant.

	PostWar	WW2
German	20	15
niemiecki	17	20

Table 7: Nominal collocate counts for *German* and *niemiecki*.

Table 7 shows the distribution of the nominal collocates of the two adjectives in question. Concordances for each collocate were manually inspected with a view to identifying their predominant context of use and they were labeled as referring predominantly to WW2, post-war or other contexts. Out of the 50 top-scored nominal collocates of *niemiecki* 20 are used to refer to post-war aspects of German culture with 15 denoting mainly WW2-related concepts. *German*, on the other hand, forms 15 recurrent WW2 combinations with 15 post-war ones. It has to be emphasized that some of the more frequent combinations such as *German army* were used both to refer to WW2, post-war and pre-war contexts, as examples b and c below show:

(a) With *German unification* fresh in the memories of local people, some fear that the eastern territories might one day revert to Germany as well. (BNC ABJ)

(b) The following month a *German army* advanced into the Balkans and in three weeks occupied Yugoslavia and Greece. (BNC CCS)

(c) The *East German army* last night intervened in the crisis gripping the country for the first time. (BNC A8X)

[3] This term also happens to be the Polish name of *Deutsche Minderheit* – a political entity representing the German minority in the Polish parliamentary elections.

[4] An ambiguous phrase in Polish which usually refers to the German currency, but occasionally it is also used to mean ‘a German brand’.

In such cases, the more frequent usage was used to categorize a given combination as primarily pre- or post-war related. It might be tempting to conclude that one reason for the relatively higher number of war-related collocations and terms in Polish is the greater intensity of the Polish war-time experience. However, both Fisher's exact and chi-square p-values for the values shown in Table 7 are insignificant ($p=0.36$, $p=0.48$ respectively). This could be partly attributable to the differences in the diachronic coverage of the BNC and NKJP corpora from which the counts were obtained. The BNC edition used to extract the HASK dictionary was published in 2001, whereas most of the texts included in the National Corpus of Polish were published in the years 2000–2010. The use of war-related phraseology in Polish public discourse may have diminished over the last two decades and so the differences observed could have been different if the two corpora were diachronically comparable. Regardless of the reason, the differences between the observed counts seem to be insignificant.

6 Summary

The sheer number of potentially salient word combinations contained in the automatic collocation dictionaries presented in this paper justifies the development of new methods of exploring and visualizing corpus data. It has been shown that ACDs can be generated to improve the recall of PUs in phraseological studies. Different types of phraseological combinations captured in ACDs can be aggregated, compared and clustered using the graph-based methodology outlined in this paper. The examples of such methods presented in the paper were focused around the idea of analyzing the reflections of cultural, political and historical contacts and transfers in recurrent word combinations. Both Polish and English nationality adjectives have been found to form hundreds of combinations including idiomatic terms, restricted collocations and seemingly regular phrases, which, on closer inspection turn out to contribute to the overall latent referential and attitudinal effects triggered by the use of a certain nationality adjective. Finally, it has been suggested that words falling under a specific semantic category, such as nationality adjectives, can be clustered on the basis of their phraseology. Obviously the analysis of collocational graphs can be extended beyond the category of nationality adjectives and the HASK dictionaries described in this paper can be used for similar phraseological studies.

Acknowledgments

This development of the dictionaries used in this research was supported in the years 2011–2012 by grants for the CESAR project: CIP-ICT-PSP CESAR 271022, MNiSW 2139/CIP2007-2011/2.

References

BNC, 2001: *The British National Corpus*. Version 2 (BNC World). Distributed by Oxford University Computing Services.

BOLINGER, Dwight, 1979: Meaning and Memory. *Experience Forms: Their Cultural and Individual Place and Function*. 95–111.

BURGER, Harald, 1998: *Phraseologie: Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt Verlag.

DAILLE, Béatrice / GAUSSIER, Éric / LANGÉ, Jean-Marc, 1994: Towards Automatic Extraction of Monolingual and Bilingual Terminology. *Proceedings of the 15th Conference on Computational Linguistics – Volume 1*, 515–521. <http://dx.doi.org/10.3115/991886.991975>.

DUNNING, Ted, 1993: Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19/1, 61–74.

EVERT, Stefan, 2005: *The Statistics of Word Cooccurrences. Word Pairs and Collocations. Phil. Diss.* Stuttgart: Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

JUILLAND, Alphonse G. / BRODIN, Dorothy R. / DAVIDOVITCH, Catherine, 1971: *Frequency Dictionary of French Words*. The Hague: Mouton de Gruyter.

KILGARIFF, Adam / RYCHLY, Pavel, 2010: Defining the Definiendum. De Schryver, Gilles-Maurice (ed.): *A way with words: recent advances in lexical theory and analysis: a festschrift for Patrick Hanks*.Kampala: Menha Publishers. 299–312.

VAN LANCKER, Diana R. / KEMPLER, Daniel, 1987: Comprehension of Familiar Phrases by Left-but Not by Right-hemisphere Damaged Patients. *Brain and Language* 32/2, 265–277.

LEA, Diana / CROWTHER, Jonathan / DIGNEN, Sheila, 2003: *Oxford Collocations Dictionary: For Students of English*. Oxford: Oxford Univ. Press.

MANNING, Christopher D. / RAGHAVAN, Prabhakar / SCHÜTZE, Hinrich, 2008: *Introduction to Information Retrieval*. Vol. 1. Cambridge: Cambridge Univ. Press.

MEL'ČUK, Igor, 2001: Collocations and lexical functions. Cowie, Anthony Paul (ed.): *Phraseology: theory, analysis, and applications*. Oxford [etc.]: Oxford Univ. Press. 23–54.

PRZEPIÓRKOWSKI, Adam et al. (ed.), 2012: *Narodowy Korpus Języka Polskiego*. Warsaw: Wydawnictwo Naukowe PWN. <http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf>.

SAVICKỲ, Petr / HLAVÁCOVÁ, Jaroslava, 2002: Measures of Word Commonness. *Journal of Quantitative Linguistics* 9/3, 215–231.

SINCLAIR, John, 1996: The Search for Units of Meaning. *Textus* 9/1, 75–106.

SVENSÉN, Bo, 2009: *A handbook of lexicography: the theory and practice of dictionary-making*. New York: Cambridge University Press.

WIERZBICKA, Anna, 2007: Reasonably well: Natural Semantic Metalanguage as a tool for the study of phraseology and its cultural underpinnings. Skandera, Paul (ed.): *Phraseology and culture in English*. Berlin/New York: Mouton de Gruyter. 49–78.

Idioms in Dictionaries for Translators[1]

LIEZL POTGIETER (Stellenbosch)

Abstract

Dictionaries are important and necessary resources for translators, but translators nevertheless have mixed feelings about bilingual dictionaries. While on the one hand they cannot really work without them, bilingual dictionaries on the other hand often give translators little or no help. It appears that bilingual dictionaries are an insufficient resource for professional translators when translating idioms. The first thing that will be looked at is the presentation of multiwords (which include idioms). A number of suggestions are made as to how this can be improved. Another problem with regard to the treatment of idioms is the translation equivalents that are provided. There are a number of improvements that can be made in this regard. There will also be looked at the current treatment of variable units and language units. Lastly, a possible model will be presented to help improve the treatment of idioms in bilingual dictionaries and make it more user-friendly for translators.

1 INTRODUCTION

> *In every language, human beings use idioms. In every language, those idioms share a common property. Decompose any idiom into its constituent words, look at the meaning of those words, and there is no way to reconstruct the idiom. In other words, the word-for-word interpretation of any idiom makes no sense whatsoever. Unless you have a dictionary specifically for translating English-language idioms, give up any hope of translating them.*
>
> Weiss 2004

The translation of idioms causes many problems for translators and the current treatment of idioms in bilingual dictionaries doesn't make the translators' job any easier.

Although the ideal solution for the translator's problem would be a complete list of fully treated idioms as outer text of a dictionary, or otherwise the inclusion of idioms as full main lemmas in the vertical ordering of the central list, there are several restrictions and peripheral conditions as discussed in lexicographic theory which also have to be kept in mind.

[1] This article is an excerpt from an unpublished Master's Thesis *Die bewerking van idiome in tweetalige woordeboeke* that was completed at the University of Stellenbosch, Stellenbosch, Republic of South Africa in April 2006.

A list of idioms is a handy way to present and treat idioms, but it often leads to problems. The alphabetical listing of idioms can be very complicated because it is difficult to determine how the idioms should be listed and the word order of idioms also often changes when used in certain texts and/or contexts. These things make it difficult for the lexicographer to present and treat idioms in an outer text in a successful and user-friendly way. As a result alternative ways have to be found to present and treat idioms in bilingual dictionaries. In the following paper several suggestions will be made.

2 Presentation

The inadequate way in which idioms are presented in bilingual dictionaries is one of the reasons why these dictionaries do not help translators sufficiently during the translation of idioms.

Currently idioms and their translation equivalents are presented in many bilingual dictionaries as part of the semantic commentary of the articles of the lemmas under which they are included and they are often found among the items giving example material. Although an idiom is in itself a lexical item which actually should have lemma status, in many existing bilingual dictionaries it only forms part of the microstructure. This method does not only deny idioms their status as lexical items, but often also implies that there exists a semantic relationship between an idiom and the preceding main lemma. There are however many other ways in which idioms can be presented in the central text which will not only make the treatment more user-friendly, but will also be able to help the translator to quickly and easily retrieve the correct idiom and a suitable translation equivalent.

One of the first changes lexicographers can make in order to improve the treatment of idioms in bilingual dictionaries is to promote the idioms to macro-structural treatment units.

A distinction can be made between a macrostructure with a vertical ordering of lemmas and one with a horizontal ordering of lemmas. Traditionally only vertically ordered lemmas are seen as main lemmas, while horizontally ordered lemmas are usually macrostructural items with sublemma status.

One of the main reasons why lexicographers make use of horizontal ordering is in an attempt to save space. They also make use of further textual condensation and space-saving by leaving out the lemmas' mutual initial component and replacing it with a space-saving marker, for example a tilde (~).

Example 1 – Tilde as space-saving marker (from MD):

> **bac´on**, spek, varkspek; spekvleis, *bring HOME the* ~, die paal haal; die draai kry; *SAVE one's* ~, daar heelhuids van afkom; *STREAKY* ~, streepspekvleis; **~ beetle**, spektor; **~er**, spekvark; **~ rind**, swoer(d); **~ roll**, spekrolletjie; **~y**, spekagtig.

In cases such as this example where lexicographers make use of textual condensation as well as horizontal ordering, the user can only reach the lemma via the vertically ordered main lemma.

In cases where the horizontally ordered lemmas have not been condensed, the previous vertically ordered lemma is not the only way to reach the horizontally ordered lemma and in such cases it is therefore much easier for the dictionary user to reach the relevant horizontally ordered lemma and/or idiom. In these cases where horizontally ordered lemmas have been written out in full and the lexicographers have made use of strict alphabetic ordering, the horizontally ordered lemmas are not just sublemmas but are fully-fledged main lemmas.

Example 2 – Horizontally ordered main lemmas (from VAW):

> **asyn**´. Suur vloeistof deur gisting verkry, bestaande uit water en asynsuur; **asynagtig; asynbottel; asynekstrak; asynerig; asynflessie; asynlug; asynmakery; asynmoer; asynsmaak; asynsout; asynstandertjie; asynsuur; asynvaatjie.**

The listing of idioms as vertically ordered main lemmas are problematic, because it is difficult to determine where in the central alphabetical list the idioms should be listed. It is however important that idioms should form part of the dictionary's macrostructure and should not be presented as part of the illustrative material. It should be treated in a more complete way. By including idioms in the dictionary as horizontally ordered lemmas without making use of textual condensation, the problem is eliminated that would have existed if the idioms had been ordered vertically, but it is still easy for the dictionary user to find the translation equivalent for the idiom he or she is looking for. In that way the vertically ordered lemma is no longer the only way to get to the idiom.

Example 3 – Idioms without textual condensation and with nest lemmatisation (adapted from MD):

> **bac´on,** spek, varkspek; spekvleis, *STREAKY* ~, streepspekvleis; ~ **beetle,** spektor; **~er,** spekvark; ~ **rind,** swoer(d); ~ **roll,** spekrolletjie; **~y,** spekagtig; **bring HOME the bacon**, die paal haal; die draai kry; **SAVE one's bacon**, daar heelhuids van afkom.

Example 4 – Idioms without textual condensation and with nest lemmatization (adapted from PEAD):

> **ba̲con** vark(spek), spekvleis; rookspek, ontbytspek; ~ *and eggs* eiers met spek; ~ **beetle** spektor. ~ **rind** swoerd. ~ **roll** spekrolletjie, spekvleisrol;
>
> **bring home the bacon**, (*infml.*) die broodwinner wees; **save one's bacon** die situasie (*of* jou bas) red, heelhuids daarvan afkom.

Lexicographers can improve the treatment of idioms in bilingual dictionaries even more by making use of a new type of nest lemmatisation. Where lexicographers make use of nest lemmatisation text blocks are formed in which the different types of data are found, but it still forms a whole, or a nest. If the lexicographers further want to improve the treatment of idioms in bilingual dictionaries after they have given the idioms sublemma status, they can make use of nest lemmatisation to form text blocks within the article trajectory. Text blocks comprise of similar data types being grouped together. These search zones within the article can for example include the following: blocks for each of the translation equivalents with their subcomments on semantics with sufficient contextual guidance and/or example material and a block for collocations. Within the article stretch it can include a partial article stretch for sublemmas. Separate text blocks for idioms can also be attached to the article of the main lemma. By making use of these text blocks in the treatment of idioms, it can be ensured that the idioms are not presented among the example material, but are easy to see and to use because the idioms now have a prominent position as macro-structural items. By making use of a quick access structure a text block that contains a partial article stretch with idioms as lemmata can be further highlighted with the help of structural markers.

Example 7 – Text blocks with structural markers (adapted from MD):

> **bac´on,** spek, varkspek; spekvleis, streepspekvleis; *STREAKY* ~; ~ **beetle,** spektor; **~er,** spekvark; ~ **rind,** swoer(d); ~ **roll,** spekrolletjie; **~y,** spekagtig.
>
> **Idm.:** *bring HOME the* ~, die paal haal; die draai kry; *SAVE one's* ~, daar heelhuids van afkom.

Example 8 – Text blocks with structural markers (adapted from PEAD):

> **ba̲con** vark(spek), spekvleis; rookspek, ontbytspek; ~ *and eggs* eiers met spek; ~ **beetle** spektor. ~ **rind** swoerd. ~ **roll** spekrolletjie, spekvleisrol.
>
> **Idm.: bring home the bacon**, (*infml.*) die broodwinner wees; **save one's bacon** die situasie (*of* jou bas) red, heelhuids daarvan afkom.

By creating separate microstructural text blocks for the cotext entries and macrostructural text blocks for the sublemmas, complex words and idioms in separate text blocks, the microstructure of the article and the presentation of idioms can be sufficiently improved. Because idioms can also be presented as horizontally ordered sublemmas in a separate text block, it helps translators to quickly and easily find what they are looking for. It can also contribute to simplifying the translators' search for an idiom in the target language with which to translate a source-language idiom.

3 Translation equivalents

The presentation of idioms in bilingual dictionaries is definitely not the only problem and the translation equivalents supplied, as well as the treatment is often an even bigger problem for translators when it comes to the correct translation of idioms.

The following section offers several possible solutions with regard to the problem of idioms, their translation equivalents and their treatment in bilingual dictionaries.

3.1 Equivalence

Dictionary users usually consult bilingual dictionaries in search of a translation equivalent in order to replace a specific source language word, expression or idiom in the target language. Translators therefore do not only consult dictionaries in order to help them with text reception and comprehension, but also to help them with text production. But it was found that dictionary users are seldom provided with contextual and cotextual guidance to help them to know which translation equivalent to use in which context or situation. This lack of contextual and cotextual guidance makes it very difficult for users to achieve communicative equivalence.

When translators make use of bilingual dictionaries during the translation of idioms, it is because they are looking for a correct translation equivalent for the idiom in question. It is therefore important that lexicographers will not only pay attention to the provision of translation equivalents, but also the provision of sufficient cotextual and contextual guidance to help the translator to quickly and easily find the correct translation equivalent and use it correctly. Acknowledgement of idioms as sublemmas should also lead to a complete treatment of the idiom. These contextual and cotextual entries can be presented as example sentences and additional data (for example glosses).

In cases where a relationship exists between the source language idiom and the target language idiom, there exists semantic as well as communicative equivalence between the source and target language and the idiom can usually be translated easily and without any problems. It is however the cases where there doesn't exist a relationship of absolute equivalence and congruence, where translators often struggle with the translation of idioms.

3.2 Divergence

It often happens that a source language idiom has more than one translation equivalent in the target language, but where the various translation equivalents are only partial synonyms or even represent different polysemous values of the source language idiom. In these cases it is especially important that the lexicographer sufficiently treat the idiom in order ensure that the correct translation equivalent will be used in the correct context.

In the matter of lexical divergence where a relationship of partial synonymy exists between the different translation equivalents, it is firstly important that it will be indicated to the dictionary user that the translation equivalent(s) are not absolute synonyms, but only partial synonyms. If the source and target language idioms are only partially equivalent then not only is it necessary that the lexicographer will indicate the differences to the user, but he or she must also provide the necessary cotextual and contextual guidance so that the user will know which translation equivalent to use in which context or not. It is especially important that lexicographers should give enough guidance to dictionary users in the case of semantic divergence (in other words, in cases where the source language idioms are polysemous). As can be seen in example 9 the polysemous values are only indicated with the help of a semi-colon and the dictionary users are provided with little or no additional information to choose the correct translation equivalent for the specific text or context.

Example 9 – Semantic divergence without additional information (from PEAD):

> **kant** … *kant en* ***klaar*** *wees* be all set; be cut and dried; be signed, sealed and delivered …

The article in example 9 is an example of a case of semantic divergence. Not only doesn't it indicate to the user what the differences in meaning between the different translation equivalents are, but the user also doesn't know which translation equivalent to use in which type of text and/or context. It is therefore important that the lexicographer will provide the dictionary user with the necessary contextual and cotextual guidance and will make the user aware of the different senses as can be seen in example 10.

Example 10 – Semantic divergence with additional information (adapted from PEAD):

> **kant** … *kant en* ***klaar*** *wees* be all set (ready for something); be cut and dried (cannot be changed); be signed, sealed and delivered (with all the necessary legal documents signed) …

This can be done with the help of glosses or example material. By adding a gloss with supplementary information the treatment becomes more user-friendly. It is however important that the lexicographer will be consistent with the inclusion of glosses and that he or she will not insert them haphazardly. It is important that all translation equivalents of which the context or meaning can cause problems be provided with glosses. Only in cases where there are no obscurity with regard to meaning or context can the gloss with contextual guidance be left out. From the perspective of the translator the inclusion of additional information is especially important because translators are usually pressured for time and have to find the correct translation equivalent as quickly as possible. By including the glosses the translators' search is limited to just one dictionary and then they don't have to consult other sources as well in order to determine in which text or context they should use the different translation equivalents.

Also in cases where there are polydivergence in the article as can be seen in example 11, it is important that the user will get sufficient guidance so he or she can quickly and easily identify the correct translation equivalent (as can be seen in example 12).

Example 11 – Polydivergence without contextual guidance (from MD):

> **break** … ~ *DOWN*, inmekaarsak, beswyk; in trane uitbars; bly steek; teëspoed kry; onklaar raak; ontleed …

Example 12 – Polydivergence with contextual guidance (adapted from MD):

> **break** … ~ *DOWN*, inmekaarsak, beswyk (neerval of sterf); in trane uitbars (huil); bly steek (bly staan); teëspoed kry (voertuigprobleme); onklaar raak (breek of gaan staan); ontleed (in dele skei of analiseer) …

Since translators and other dictionary users usually first look for a target language idiom with which to translate the source language idiom, it is therefore important that the translation equivalents be listed as such (example 13), in other words by first listing the translation equivalents for the idiom which are idioms and then listing those that are only single words or paraphrases (example 14).

Example 13 – Before – Ordering of translation equivalents: first idioms, then single words and/or paraphrases (from PEAD):

oog … *oë **knip** vir iemand* wink at someone; make eyes at someone …

Example 14 – After – Ordering of translation equivalents: first idioms, then single words and/or paraphrases (adapted from PEAD):

oog … *oë **knip** vir iemand* make eyes at someone; wink at someone …

3.3 Zero equivalence and surrogate equivalence

As in the case with single words, there also are cases of zero equivalence with idioms, in other words, idioms for which there aren't any translation equivalents in the target language. It is however important that lexicographers will still include these idioms in the dictionary and provide them with an explanation of meaning or a surrogate equivalent.

Because the dictionary user works with the assumption that if the source language item is an idiom, the target language item will also be an idiom, it is very important that the lexicographer will indicate to the dictionary user if an idiom has a surrogate equivalent and there doesn't exist an idiom with which to translate it in the target language so that the dictionary user will know in which cases there is only a single word or a paraphrase of meaning to be used in the target language (as can be seen in example 15 and 16). This can be done by adding simple structural markers to the surrogate equivalents.

Example 15 – Surrogate equivalents with structural markers (adapted from MD):

length … *GO to great (all) ~s*, *alles in jou vermoë doen …

Example 16 – Surrogate equivalents with structural markers (adapted from PEAD):

> **peace** … ~ *of* ***mind*** ¤gemoedsrus, ¤gerustheid …

3.4 False friends

> *Confusion arises because word A (which belongs to the foreign language ...) looks or sounds exactly or nearly like word B, which belongs to the ... mother tongue. The user then establishes an unwarranted interlingual equivalence on the basis of this total or partial similarity.*
>
> Hayward and Moulin (in Gouws 1986: 179)

Another possible problem, of which lexicographers have to make dictionary users aware, is the occurrence of false friends. False friends prevail when it looks as if there exists a relationship of equivalence between two words or idioms based on certain similarities, while in fact the two words or idioms are not equivalent.[2]

One such an example of false friends is the English idiom *to pepper someone* and the Afrikaans idiom *om iemand te peper* which can also be translated as *to pepper someone*. The meaning of the English idiom is to bombard someone with questions while the Afrikaans idiom means to hit someone (Prinsloo 2004: 274). It is important that the lexicographer makes the user aware that the English idiom cannot be translated with the Afrikaans idiom. It can for example be done by making use of a gloss or additional note after the idiom as can be seen in the following example.

Example 17 – False friends (adapted from PEAD):

> **pepper** … ~ *s.o./s.t. with* … iem./iets met … bestook (*vrae ens.*) [nie *iemand peper met* … nie] …

The same also applies to the idiom *make out with someone* and the expression *met iemand uitmaak* (translated as *make out with someone*). While the English idiom means to kiss someone, the Afrikaans expression means to end a relationship. Although the two idioms look similar, their meanings are very different and it is therefore very important that the lexicographer will make the users aware of this difference.

[2] For more on false friends, see Gouws et al. 2004: 797–806.

Example 18 – False friends (adapted from PEAD):

> **make** … ~ *out with s. o.*, (*Am.*, *infml.*) ’n vryery met iemand hê [Let wel: Hierdie idioom kan nie vertaal word as *met iemand uitmaak* nie] …

In this way the lexicographer can prevent the translator from translating the idiom incorrectly due to false friends in the target language.

4 Direct translations

Something else that lexicographers must pay heed to when treating idioms in bilingual dictionaries is to make dictionary users (and translators) aware of the fact that although some idioms in the colloquial language can be translated directly, the direct translations are not always the correct ones. An example of such a case is the English idiom *born with a silver spoon in one’s mouth* although it are often directly translated into Afrikaans, the correct Afrikaans translation is *met ’n goue lepel in die mond gebore wees* [born with a *golden* spoon in one’s mouth].

Example 19 – Direct translations with notes (adapted from MD):

> **spoon** … *born with a SILVER ~ in one’s mouth*, met ’n goue lepel gebore wees [Nie ’n *silwer* lepel nie]; ’n gelukskind wees …

Once again, as in the case of false friends, the lexicographers can make use of glosses or notes to make the dictionary user aware of the fact that these idioms are often translated incorrectly.

4.1 Optional language units and variants

The last aspect that translators can pay attention to in the treatment of idioms in bilingual dictionaries, is the treatment of different types of language units and variants. It is however important that the lexicographers will have a system according to which these types of units are treated.

In the case of optional language units the use of brackets work well. It is important that lexicographers indicate when the idiom has variant forms so the dictionary users won’t think they are essential parts of the idiom (see example 20). By placing the optional language units in brackets, the translators who have to translate the idioms can clearly see which parts of the idiom are essential and which are optional.

Example 21 – Optional language units in brackets (from MD):

> **end** … *make* (*both*) *~s MEET*, die tering na die nering sit …

Lexicographers should also pay attention to the treatment of variants. Currently variants are indicated to the dictionary user, but the system is very inconsistent – in some cases it is indicated with brackets, in other cases with a forward slash or by using words like *or*. These different presentations can be very confusing for users because they don't necessarily realise that it is the same type of language unit being treated in all the cases. It is therefore important that the lexicographer will present the data in a way that the dictionary user (or translator) can easily deduce what information the dictionary user is providing and how he or she can or should use the information in his or her text. It is therefore necessary that the lexicographers will have a system for the treatment of variants and that it will be applied consistently throughout the dictionary. It will be good if the lexicographers make use of the forward slash to indicate the different variants since the use of *or* can take up too much space.

Example 22 – Variants separated with forward slashes (adapted from PEAD):

> **eye** … *have an ~* ***on*** *s.t.* 'n ogie/die oog op iets hê …

Example 23 – Variants separated with forward slashes (adapted from MD):

> **length** … *GO to great/all ~s*, alles in jou vermoë doen …

The last aspect of variants that lexicographers must pay attention to is to ensure that the alternative variants are always clearly indicated as such. Unlike the idiom *to see something in a certain light* where its treatment looks as if the word *different* is an essential part of the idiom while it is actually only one of several alternatives that can be used in the idiom. The presentation of this (as well as other similar) idioms can be improved a lot if not only one of the possibilities are listed and if the alternatives are indicated as alternatives and not essential parts.

Example 24 – An unclear indication of alternative variants (from PEAD):

> **light** … ***see*** *s.t. in a different ~* iets in 'n ander lig beskou …

Example 25 – Clear indication of alternative variants (adapted from PEAD):

> **light** … ***see*** *s.t. in a different/bad/good* ~ iets in ’n ander/slegte/goeie lig beskou …

In this way the variants are listed and the treatment of the idiom is improved. A translator confronted with the idiom *to see something in a bad light* and who have to translate it, can now easily deduce what the correct translation equivalent for the idiom is.

5 In closing

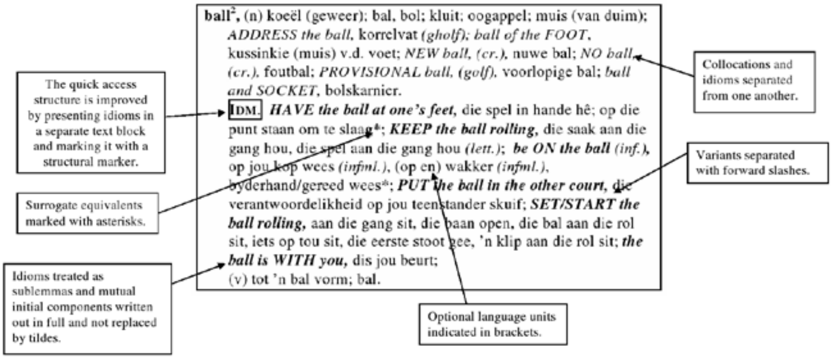
In the next two articles there can be seen how the treatment of idioms can be improved by making use of the suggestions made in this paper. In example 26 you can see the original article as found in *Major Dictionary*.

Example 26 – *ball* (from MD):

> **ball2,** (n) koeël (geweer); bal, bol; kluit; oogappel; muis (van duim); *ADDRESS the* ~, korrelvat (*gholf*); ~ *of the FOOT*, kussinkie (muis) v.d. voet; *HAVE the* ~ *at one’s feet*, die spel in hande hê; op die punt staan om te slaag; *KEEP the* ~ *rolling*, die spel (die saak) aan die gang hou; *NEW* ~, (*cr.*), nuwe bal; *NO* ~, (*cr.*), foutbal; *be ON the* ~, wakker, byderhand (gereed) wees; *PROVISIONAL* ~, (*golf*), voorlopige bal; *PUT the* ~ *in the other court*, die verantwoordelikheid op jou teenstander skuif; *SET* (*START*) *the* ~ *rolling*, aan die gang sit, die baan open, die bal aan die rol sit, iets op tou sit, die eerste stoot gee, ’n klip aan die rol sit; ~ *and SOCKET*, bolskarnier; *the* ~ *is WITH you*, dis jou beurt; (v) tot ’n bal vorm; bal.

The next article (example 27) is the improved article with arrows indicating the different improvements.

Example 27 – Improved article:



ball^2, (n) koeël (geweer); bal, bol; kluit; oogappel; muis (van duim);
ADDRESS the ball, korrelvat *(gholf); ball of the FOOT*, kussinkie (muis) v.d. voet; *NEW ball, (cr.)*, nuwe bal; *NO ball, (cr.)*, foutbal; *PROVISIONAL ball, (golf)*, voorlopige bal; *ball and SOCKET*, bolskarnier.
IDM. ***HAVE the ball at one’s feet,*** die spel in hande hê; op die punt staan om te slaag*; ***KEEP the ball rolling,*** die saak aan die gang hou, die spel aan die gang hou *(lett.)*; ***be ON the ball*** *(inf.)*, op jou kop wees *(infml.)*, (op en) wakker *(infml.)*, byderhand/gereed wees*; ***PUT the ball in the other court,*** die verantwoordelikheid op jou teenstander skuif; ***SET/START the ball rolling,*** aan die gang sit, die baan open, die bal aan die rol sit, iets op tou sit, die eerste stoot gee, ’n klip aan die rol sit; ***the ball is WITH you,*** dis jou beurt;
(v) tot ’n bal vorm; bal.

The quick access structure is improved by presenting idioms in a separate text block and marking it with a structural marker.

Surrogate equivalents marked with asterisks.

Idioms treated as sublemmas and mutual initial components written out in full and not replaced by tildes.

Collocations and idioms separated from one another.

Variants separated with forward slashes.

Optional language units indicated in brackets.

As can be seen there are several possibilities and ways in which the current treatment of idioms in bilingual dictionaries can be adapted to be more user-friendly and can also help translators to quickly and easily find the correct translation equivalent for an idiom within a specific text and/or context.

It is however also important that lexicographers will pay attention to the front matter of the dictionary and that they will provide dictionary users with a comprehensive and useful users' guidelines text which will explain exactly how the lemmas, sublemmas as well as idioms are treated in the dictionary and what the meaning of the different structural markers, brackets and forward slashes are. In this way lexicographers can help translators and other dictionary users to translate idioms more successfully.

Bibliography

EKSTEEN, Louis C., 1997: *Major Dictionary*. Kaapstad: Pharos. (MD)

GOUWS, Rufus H., 1996: Idioms and collocations in bilingual dictionaries and their Afrikaans translation equivalents. *Lexicographica* 12, 54–88.

GOUWS, Rufus H. / PRINSLOO, Danie J. / De SCHRYVER, Gilles-Maurice, 2004: Friends will be Friends – True of False. Lexicographic approaches to the treatment of false friends. *Proceedings of the eleventh EURALEX congress*. Lorient: Université de Bretagne-Sud. 797–806.

LABUSCHAGNE, Frans J. / EKSTEEN, Louis C., 1992: *Verklarende Afrikaanse woordeboek*. Pretoria: JL van Schaik. (VAW)

PEAD, 2005: *Pharos English-Afrikaans Dictionary*. Kaapstad: Pharos Dictionaries.

PRINSLOO, Anton F., 2004: *Spreekwoorde en waar hulle vandaan kom*. Cape Town: Pharos.

WEISS, John, 2004: *The tricky business of translation: Idiom's delight*. <www.devel.lyx.org/translation_hints.php3>. Access: 24. 02. 2005.

Valency Patterns in Dictionaries[1]

Makoto Sumiyoshi (Osaka)

Abstract

Much emphasis has recently been placed on the importance of dictionaries being “phrase-centred” (Béjoint 2010) or “pattern-driven” (Hanks 2008), which implies that valency patterns are an essential part of dictionaries (Herbst/Klotz 2009; BBI2). Comparing valency patterns in monolingual learners’ dictionaries (MLDs), valency pattern dictionaries (VPDs), and authentic data, this paper argues that (i) sometimes, differences exist in the descriptions of valency patterns between MLDs and VPDs; (ii) MLDs are sometimes under the influence of prescriptivism, which leads to the exclusion of some valency patterns that reflect language changes that have been happening in contemporary English; (iii) MLDs should pay more attention to authentic data to provide learners of English with more comprehensive phraseological information; and (iv) phraseologists and lexicographers should be more sensitive to language changes to be more precise about the valency patterns of words.

1 Introduction

Lexicographers are now trying to integrate phraseological aspects of language into dictionaries (Granger/Paquot 2008a: 1345). Some researchers have recently argued that dictionaries should be “much more phrasal than they currently are” (Granger/Paquot 2008a: 1353), “pattern-driven” (Hanks 2008: 103) or “phrase-centred” (Béjoint 2010: 318). These arguments reflect the growing perception that multi-word expressions and patterns play a pivotal role in first and second language acquisition, and language teaching.

Grammatical patterns, that is, valency patterns, should be fully and successfully incorporated into descriptions in dictionaries, especially, monolingual and bilingual learners’ dictionaries, because such dictionaries are designed for learners who “need to become acquainted with the patterns of” the language they are learning for encoding purposes (Hornby 1975: v). They are “all the constructions which a speaker of the language must know in order to use the word flexibly and fluently” (Atkins/Rundell 2008: 219f.). However, valency

[1] The present work was supported by JSPS KAKENHI Grant Numbers 23720256 and 24652097.

patterns, given their sometimes elusive nature, pose complicated challenges for lexicographers in dictionary-making.

The objectives of this paper are to (i) compare valency patterns in representative monolingual learners' dictionaries (MLDs) and valency pattern dictionaries (VPDs) to show that sometimes discrepancies exist between them; (ii) point out that MLDs are sometimes under prescriptive influence, and as a result of this, descriptions of valency patterns may not be systematic; (iii) show that lexicographers should pay more attention to authentic data to provide learners of English with more comprehensive phraseological information, including less frequently used patterns; and (iv) argue that lexicographers and phraseologists should be more sensitive to language changes to be more precise about valency patterns.

The authentic data in the following discussion was obtained from the *Corpus of Contemporary American English* (COCA), the *Corpus of Historical American English* (COHA), and the *British National Corpus* (BNC), all of which are accessible at <http://corpus.byu.edu/>.

2 VALENCY PATTERNS

2.1 Valency patterns as phrases

Herbst et al.'s *Valency Dictionary of English* rewords the term *valency* as *complementation* (xxv), and argues that it is "an important area of the description of English, one which is on the boundaries of lexis and grammar" (vii). Faulhaber (2011: 6) defines a verb valency pattern as "the simultaneous choice of one or a number of complements in combination with a verb functioning as valency carrier". Herbst (2009) argues that valency is a lexical phenomenon and that Sinclair's idiom principle can be depended upon to account for valency phenomena. In this paper, I would like to define a valency pattern as a phrase consisting of a verb, adjective, or noun (i. e. valency carrier) plus a syntactic complementation (including words before the valency carrier and after it) chosen along with the valency carrier as a prefabricated sequence of words, that is, as a phraseological unit (see also Aarts 1999: 19; Hunston/Francis 2000).

Some phraseologists argue that, with the exception of prepositional phrases, grammatical patterns or complementations should not be dealt with in the realm of phraseology (Granger/Paquot 2008b: 43). However, given the arguments by phraseologists that no clear distinction exists between lexis and grammar, it may not be reasonable to divide collocations into lexical and grammatical ones and to maintain that the latter does not come under the rubric of phraseology. If no distinction exists between grammar and lexis, “the traditional domain of syntax will be invaded by lexical hordes” (Sinclair 1991: 29). Given the blurring of the distinction between grammar and lexis, both lexical collocations and grammatical collocations can be justifiably regarded as kinds of prefabricated phrases. Béjoint (2010: 342) argues that “*pattern* is a general word” and that “it includes collocations, syntactic patterns, à-la-Hornby and co-occurrences over large spans”.

2.2 Valency pattern dictionaries of English

According to Faulhaber (2011: 3), “because of its orientation towards item-specific properties, valency theory seems to be predestined to a close cooperation with lexicography”. Corpus-based or corpus-driven research on English valency patterns led to the publication of paper- or web-based phraseological dictionaries:[2]

– English valency pattern dictionaries

Herbst et al. (2004): *A Valency Dictonary of English* (VDE)

Benson et al. (2009): *The BBI Combinatory Dictionary of English* (3rd edition) (BBI3)

Francis et al. (1996): *Collins COBUILD Grammar Patterns 1: Verbs* (Patterns 1)

Francis et al. (1998): *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives* (Patterns 2)

The following discussion will revolve around such findings about English valency patterns. Even if the term *valency* is not used or dictionaries are not compiled in the framework of valency theory, the pattern descriptions contained in them can be equated with valency patterns, because they all describe formal realizations of the complements valency carriers take.

[2] Some valency pattern projects are in progress and open to the public on the Internet: *The Contragram Dutch-French-English Contrastive Valency Dictionary* <http://www.contragram.ugent.be/cvvd/cvvd.htm>; *Pattern Dictionary of English Verbs* <http://deb.fi.muni.cz/pdev/>; *Erlangen Valency Patternbank* <http://www.patternbank.uni-erlangen.de>.

2.3 Valency pattern descriptions in dictionaries: How challenging are they for lexicographers?

Sinclair (2008: xvii) implies that lexicographers have to overcome considerable problems to leave the well-trodden path of traditional dictionaries. The difficulty of making dictionaries more phrasal has been reiterated by some researchers (Hunston/Francis 2000: 8; Moon 2008: 314; Hanks 2006: 121; Béjoint 2010: 319 and 343).

Then, how challenging is it for lexicographers to describe valency patterns in dictionaries? Hornby (1975: v) says *intend*, *hope*, *want*, and *propose* are compatible with *to*-infinitive clauses, while *suggest* is not. However, valency pattern identification is not always so easy. Lexicographers searching corpora for valency patterns are likely to be at a loss when faced with inconsistencies shown by data.

2.3.1 *demand* + X + *to be* V-*ed*

It is not easy to identify what kind of syntactic structure follows the verb *demand*. VDE (219f., s. v. *demand* [*verb*]) includes the [N to passive-INF] pattern in the valency patterns of the verb, exemplifying it with the following illustrations:

(1) a. As you do more work or exercise, your body will *demand* more blood to be supplied to working muscles to supply them with oxygen and nutrients.

b. They *demanded* more planes to be made available and regular flights.

A question arises as to whether this use of *demand* is in normal currency. Givón (1980: 358) judges the verb used with the *to*-infinitive complement to be unacceptable:

(2) *She *demanded* him *to* do it.

In the example, the verb in the *to*-infinitive clause is not passivized, so it may be possible to claim that the passivized variant of the pattern is acceptable, while the unpassivized one is not. The wildcard search of [*demand**] [n*] *to be* to find that valency pattern and its variants in COCA yielded only two examples, with irrelevant examples excluded from the count:

(3) a. Officers, to him, were just *demanding* guests *to be placated* with good service.

b. He *demanded* $20,000 *to be left* on this cellar staircase near Mrs. Silverman's apartment.

VDE gives the frequency label *rare* to a particular valency pattern in entries when it occurs in very few instances in the corpus used for the research. With this label, the pattern is "not recommended to non-native speakers for active use but it should not be considered wrong" (xx). The label *rare* is not given to the valency pattern *demand* + NP to-INF, which means that this pattern is regarded as a typical one that *demand* is associated with. However, given the scarcity of instances of *demand* + X + *to be* V-*ed* in COCA along with Givón's intuitive judgment, it does not seem justifiable to list this valency pattern in the entry for *demand*. This shows the difficulty of identifying the particular valency pattern of a particular verb even with corpus evidence.

2.3.2 *apologize* + *that*-clause

According to the MLDs available today, the verb *apologize* in its canonical valency pattern is used as an intransitive verb followed by optional prepositional phrases (i. e. *apologize* (*to* somebody) (*for* something)). However, as I discussed elsewhere (Sumiyoshi 2009; Sumiyoshi 2011), in present-day English, the verb has started to occur with another valency pattern, that is, a *that*-clause, which some prescriptive or theoretical grammarians judge to be unacceptable.

Cambridge International Dictionary of English (CIDE) lists the valency pattern in the entry for *apologize* with the grammar pattern code [+ *that* clause] illustrated by *He apologized that the statistics had been inaccurate*. The revised editions of CIDE, that is, CALD1, CALD2 and CALD3, omitted this pattern description. No monolingual learners' dictionaries now are loaded with such information in the entry for *apologize*. VDE does not have an entry for *apologize*.

In *Oxford Learner's Thesaurus*, which can be characterized by its macrostructure as a synonym dictionary integrated with information on patterns and collocations, the valency pattern is given in the entry for *apologize*. One of the online valency patterns dictionaries, the *Contragram Dutch-French-English Contrastive Valency Dictionary* (see note 1) also includes the *that*-clause pattern (NP _ (*to* NP) *that* Pfin) in the list of the valency patterns of *apologize*, along with some illustrations.

Facing these changes in dictionary descriptions, we realize how difficult it is to identify valency patterns of a particular word and, if identified, to describe them in dictionaries. Even if a particular pattern is identified through corpus research, the identified pattern may not be presented for some, at times prescriptive, reasons.

2.3.3 Verbs of manner of speaking + *that*-clause

Some verbs of manner of speaking occur more easily with *that*-clauses than others do (Quirk et al. 1985: 1182). Tables 1 and 2 show whether *that*-clause patterns are listed in examples or with grammar codes in entries for verbs of manner of speaking in monolingual learners' and English-Japanese learners' dictionaries, respectively. Seven verbs of manner of speaking are chosen here. As is clear from the tables, MLDs tend to be selective in choosing what to give to the users, while English-Japanese learners' dictionaries tend to be more comprehensive in their descriptions of patterns. The fact that no mention is made of a particular pattern in the entry for a particular word does not mean the pattern is unavailable for the word; however, the user tends to read the entry as suggesting it is.

	LDCE$_5$	*LAAD*$_2$	*OALD*$_7$	*Cobuild*$_6$	*CALD*$_3$	*MED*$_2$
bellow						
grumble	+		+	+		+
mumble			+			
murmur			+	+		
mutter			+			
shout			+		+	
whisper			+	+		

Table 1: Verbs of Manner of Speaking and *That*-clause Patterns in MLDs (Sumiyoshi 2009).

	Genius$_4$	*Wisdom*$_2$	*Youth Progressive*	*O-Lex*	*Luminous*$_2$	*Royal*
bellow			+			
grumble	+	+	+	+		
mumble	+	+	+	+	+	+
murmur	+	+	+	+	+	
mutter	+	+	+	+	+	+
shout	+	+	+	+	+	+
whisper	+	+	+	+	+	+

Table 2: Verbs of Manner of Speaking and *That*-clause Patterns in EJLDs (Sumiyoshi 2009).

Siepmann (2008) argues that the reason that dictionaries fail to record multi-word expressions is that teams of lexicographers, consisting only of native speakers of English, often do not "notice the idiomaticity of the multi-word expressions because of their semantic transparency" (193). Thus, multi-word expressions escape the attention of lexicographers, because they are too familiar to native speakers of English. Lexicographers from non-English speaking countries may often more easily identify valency patterns in a data source. See Hunston (2010) for further discussion about the difficulty of identifying patterns in general.

3 Case studies

Valency Patterns	Monolingual Learners' Dictionaries					Dictionaries of Valency Patterns		
	LDOCE5	MED2	OALD8	CALD3	CIDE	VDE	BBI3	Patterns 1 & Patterns 2
accept + *to* V		***2**		***2**	**+**	+		
convince + X + *to* V	**+*1**	+	**+*1**	**+**	**+**	+	+	+
advise + V-*ing*			+	+	+	+	+	+
It is ***big*** *of* X *to* V	**+**	**+**	**+*3**	**+*3**	**+**	+	+	+
It is ***responsible*** *of* X *to* V						+		+
X *is* ***right*** + *to* V	+		+	+	+	+	+	+
X *is* ***right*** + *that*-clause						+		+
It is ***right*** *of* X *to* V***4**		+			+		+	+
X *is* ***cool*** + *with…*	+		+	+		+		
X *is* ***cool*** + *about…*	+				+	+		+
It is ***cool*** (*for* X) *to* V						**+*5**		+

Notes:

*1 The dictionaries provide the valency pattern in the entry for *convince*, but they include a usage note in the entry for *persuade* stating some British speakers believe this use is incorrect.

*2 The dictionaries include a usage note saying it is wrong to use *accept* instead of *agree* with *to*-infinitives.

*3 The valency pattern itself is not given in the entry for *big*, although the idiomatic phrases *That's big of you* (OALD8) and *That's really big of you!* (CALD) are given instead.

*4 The valency pattern *'it is right* (*for* X) *to* V' should be considered different from this valency pattern. In *VDE*, for example, the valency patterns **[it] + for N + to-INF/[it] + for N to-INF** are given with the illustration *I felt Barry didn't think it was right for me to go away on my own.*

*5 The frequency label *rare* is given.

Table 3: Valency Patterns in Monolingual Learners' Dictionaries and Valency Dictionaries.

In this section, some words and their valency patterns are chosen for case studies in which comparisons are made between valency descriptions of MLDs and those of valency/phraseological dictionaries. Although it was published about 20 years ago, CIDE was also used for the studies because the dictionary is best suited to see changes in lexicographical description. COBUILD6 is excluded from the investigation, because *COBUILD Grammar Patterns 1* and *2* are included in the phraseological dictionaries used for this investigation.

3.1 Overview

Table 3 shows whether the valency patterns listed are specified with grammar pattern codes or with illustrations in the dictionaries used for this investigation. While VDE is fully exhaustive in its coverage of the valency patterns examined here, the patterns are not fully treated in MLDs. It is surprising to find that *advise* + V-*ing* goes unrecorded in LDOCE5 and MED2 despite the fact that it is fully treated in the three valency pattern dictionaries. Although CIDE is more comprehensive, the subsequent revised edition, CALD3, is less so. MED2's patchy coverage is somewhat striking.

Overall, the adjective valency patterns under study do not receive adequate treatment in the dictionaries examined. This is probably because it is more difficult to identify which adjective goes in which adjective valency patterns than to identify verb valency patterns (Yagi 1999: 139). The meaning of an adjective is determined by the patterns it occurs in, so just examining the sense of an adjective itself does not lead to a successful understanding of the valency patterns associated with it. For example, in the sentence *it is big of you to admit that you were wrong*, the phraseology *it is big of* X *to* V, and not the adjective *big* itself, determines the meaning of the sentence "you are generous or kind". The adjective *big* has that meaning only in this phraseology and the sentence *you are big* means something entirely different (Hunston/Francis 2000: 105). In addition, this *big*-valency pattern has an ironical implication, as many dictionaries point out. In this configuration, the speaker does not assemble each word one by one. Rather, the speaker chooses this sequence of words, or these multi-word expressions, at one go to express the meaning intended.

The adjective *responsible* is also less likely to be associated with the valency pattern *It is* ADJ *of* X *to* V. Although examples can be found in COCA, it's rare.

1	And so I think	**it was responsible of me to,**	you know, talk through my
2	But	**it is very responsible of you to**	to explain it's not about timing the market
3	But I think	**it's not responsible of us to**	just take them at their word.
4	I think	**it would not be responsible of me to**	say that I anticipate that

Figure 1: Selected and abridged concordances for *It is responsible of* X *to* V in COCA.

The valency pattern X *is right to* V is statistically justified to be included in dictionaries. Among the three valency patterns of *right* given here, this is the most frequently used. It is interesting to learn that, contrary to the pattern descriptions found in MED2, and the two VPDs, no examples of the valency pattern *it is right of* X *to* V can be found in COCA at the time of writing. The valency pattern *it is right for* X *to* V does exist. Note that this valency pattern is usually associated with another phrase *I don't think*, as shown by Figure 2. Thus, it is better to treat this valency pattern in a larger syntactic environment and regard the whole part *I don't think it is right for* X *to* V as a phrase. The valency pattern X *is right* + *that*-clause more frequently occurs, so it would be more useful to add this valency pattern to the entry for *right* in MLDs.

1	Because I don't think	**it's right for him to**	have to answer every personal question.
2	I don't think	**it's right for me to**	try to be something that I'm not.
3	I don't think	**it's right for you to**	stay together.
4	I don't think	**it's right for you to**	say that Pat sounds like Bonior and Gephardt.
5	nobody will know where we are,	**it's right for us to**	be together.

Figure 2: Selected and abridged concordances for *It is right for* X *to* V in COCA.

Learners' dictionaries do not have to be as exhaustive in pattern description as valency/pattern dictionaries, because atypical patterns, if included in learners' dictionaries, are confusing and tend to harm students rather than help them. However, this does not mean that dictionaries do not have to aim at more comprehensive coverage. Adjective pattern descriptions in monolingual learners' dictionaries should be greatly improved, and phraseologists can greatly contribute to this.

3.2 *accept* + *to* V / *convince* + X + *to* V

Some prescriptive grammarians do not allow *that*-clauses to follow the verbs *accept* and *convince* (Follett 1998: 12 and 84; Turton/Heaton 1996: 5 and 83). All the dictionaries included in Table 3 show that the verb *convince* is used in the *convince* + X + *to* V pattern. In this valency pattern, the verb is used in the sense of *persuade*, which, in the past, was not acceptable for many people. The canonical valency patterns of this verb were *convince* + X + *of* N and *convince* + X + *that*-clause. The *to*-infinitive pattern of *convince* first occurred in the 1950s in the US after semantic confusion arose between *persuade* and *convince*. Since then it has been gaining currency. Such semantic confusion should not have occurred under the terms of prescriptivism. The usage note found in some dictionaries and prescriptivists' remarks are a residue of this prescriptivism. On the other hand, all the dictionaries record the valency pattern in the entry for *convince* without mentioning this prescriptivism, which is probably a reflection of descriptivism whereby lexicographers judge *convince* + X + *to* V to be acceptable in present-day English. Iyeiri (2012) undertakes a detailed quantitative and historical investigation of *convince* used in this valency pattern, and she clarifies some intriguing facts about it, saying the active voice of *convince* is used in this valency pattern, while the passive voice, *be convinced*, is a fixed idiomatic phrase that is less likely to occur with *to*-infinitives.

VDE illustrates the *accept* + *to* V pattern with the sentence *I would be delighted if you accept to come with me* (5). In CIDE, you can find the example *She's accepted to give the opening speech at our conference* in the entry for *accept*, which has been omitted from the subsequent revised editions (see Table 3). Here are some selected concordance lines for *accept* + *to* V from COCA.

1	because of this, we have	**accepted to**	be as part of a joint delegation with Jordan
2	and the Soviet Union never	**accepted to**	go otherwise.
3	Hussein was more or less cornered into	**accepting to**	have George Bush on television.
4	among the children of families who	**accepted to**	participate in the research.

Figure 3: Selected and abridged concordances for *accept* + *to* V in COCA.

This valency pattern is not frequently used in the corpus, which may give lexicographers good reason to exclude it from a list of the valency patterns of *accept*. However, undeniably, this pattern does exist. Algeo (2006: 247) argues that this valency pattern of the verb seems to be characteristic of British English; however, it occurs also in American English. Just as *convince* has acquired a new valency pattern under the influence of *persuade*, so has *accept* begun to be used in the valency pattern *accept* + *to* V by analogy with *agree* + *to* V in both British and American English. Judging from the descriptions in MLDs, *accept* is now under a stronger influence of prescriptivism, while *convince* has been emancipated from such a rule.

The discussion above suggests the difficulty of identifying a valency pattern of a particular word. Additionally, it reminds us to consider to what extent we should be descriptive or prescriptive in dictionary-making. CIDE tended to be more exhaustive, that is, descriptive in its pattern description, but overall, the current MLDs are inclined toward prescriptivism. This is unavoidable, given that learners have to learn more typical, uncontroversial, valency patterns. As far as *to*-infinitive clauses go, *accept* rejects them and *convince* accepts them, in spite of the fact that neither of them is acceptable in the context of prescriptivism. This fact suggests that it is important for lexicographers to be more sensitive to language changes, because valency patterns that a particular word takes are always in a state of flux.

3.3 *Cool* and its valency patterns

1	and all of that was what I was, I **was**	**cool with**	that until I learned that I can be more than
2	If you'**re** not	**cool with**	letting him listen in on your ritual, no actual
3	**Are** you	**cool with**	that?
4	"Agree." "Just tell her it'**s**	**cool with**	you.
5	Whatever she does **is**	**cool with**	me.

Figure 4: Selected and abridged concordance for *cool* + *with*… in COCA.

1	I'm happy, confident, and even	**cool about**	being stared at.
2	Kim was pretty	**cool about**	all the honors she's received
3	As it turned out, she was very	**cool about**	things, easygoing and
4	children. But she is	**cool about**	the angry letters she gets

Figure 5: Selected and abridged concordance lines for *cool* + *about* X in COCA.

The slang use of *cool* to mean *good* and *fine* originates from Black English; the use started in the 1930s (*Chambers Slang Dictionary*, s. v. *cool*). When the adjective is followed by the preposition *with*, the phraseology means "be fine or OK with…", but when it is followed by the preposition *about*, the phraseology means "don't care about…" (Figure 4 and 5). The senses of the adjective *cool* in this slang use can be distinguished by the following prepositions. These two prepositional valency patterns are recorded only in LDOCE5 and VDE, both of which fail to explain this difference in meaning that results from differences in phraseology. The fact that the adjective *cool* can be used to mean *good* opens up the possibility of it being used in the valency pattern *it is* ADJ *for* X *to* V (Figure 6).

Although VDE says this valency pattern of *cool* is rare, this is probably because the usage of *cool* in this sense has been in the process of changing, from slang use to informal use. Ephemeral slang use of a word usually takes time to establish itself in the language. Even after it has made headway into informal use, the word needs some more time to take the valency pattern compatible with the newly acquired meaning. As Table 4 suggests, that valency pattern of *cool* has been slightly but steadily on the rise in frequency. The valency pattern change started in the 1970s, probably because *good* and other semantically similar adjectives take the valency pattern *it is* Adj (*for* X) *to* V, and in the 21st century the pattern has been establishing itself as one of the valency patterns of the informal use of the adjective.

1	a casual smile on his face.	**It's cool to**	be part of this tribe.
2	own celebrity go to his head.	**It's cool to**	be recognized, but that's not why I do this,
3	they've done it!	**It's cool to**	hear my voice coming out of Stripes' mouth.
4	advertising into thinking	**it's cool to**	smoke. If you want to
5	Anyway, I guess	**it's cool for us to**	stay till he gets back.

Figure 6: Selected and abridged concordances for *It is cool* (*for* X) *to* V in COCA.

	1940s	1950s	1960s	1970s	1980s	1990s	2000s
It is *cool* (for X) to V	–	–	–	3	1	2	16

Table 4: Frequencies of *it is cool* (*for* X) *to* V in COHA.

4 Conclusions

Valency patterns have received relatively favourable treatment since the embryonic stage of MLDs, but it seems that with regard to accuracy about valency pattern descriptions in dictionaries, much remains to be desired. This is because identifying valency patterns, which are a type of multi-word expression, is not easy, even when corpora are easily available; as a result, "all dictionaries can still be found wanting in their treatment of phrases, in exhaustiveness, in accuracy or both" (Béjoint 2010: 318).

A reason for this difficulty is that valency patterns are in part related to language change. Language is always in constant flux, and so are valency patterns. Barlow (1999: 328) adduces historical evidence to show that while the verb *claim* never occurred with a *that*-complement clause early in the 18th century, in the 19th century, its valency pattern (though he does not use this term) changed under the influence of its synonymous verb *assert*, and began to fit well with the V + *that*-clause pattern. This is a phenomenon parallel to the case of valency pattern expansion observed in the case of *accept* + *to* V, which was triggered by *agree* + *to* V and has been changing to become a verb semantically compatible with *to*-infinitives in present-day English. The reason that *accept* + *to* V is judged to be anomalous is that the valency pattern is still in its embryonic stage, and it will be some more time before the pattern establishes itself.

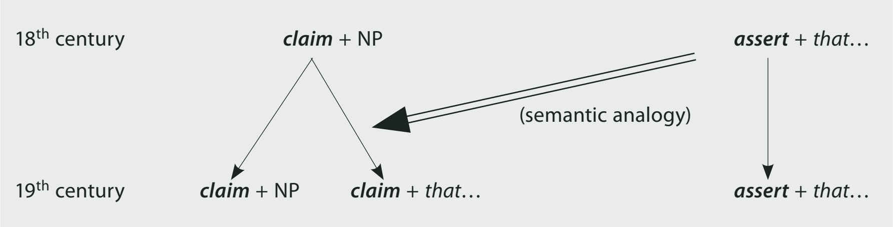


Figure 7: Valency pattern expansion of *claim*.

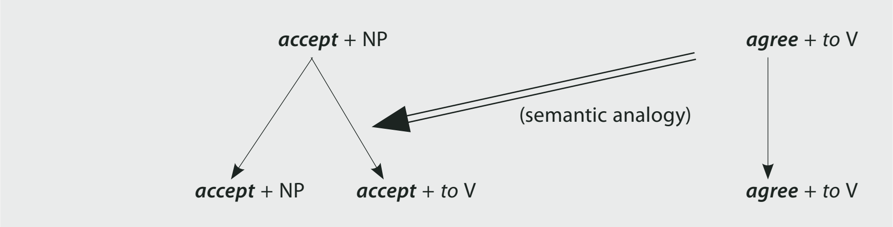


Figure 8: Valency pattern expansion of *accept*.

It is often difficult to draw a sharp line between what is possible and what is not, because, in some cases, the possibility of a word taking a particular valency pattern is related to language changes occurring in present-day English. Low frequency may indicate that a particular valency pattern has just come into use as a result of language changes. MLDs tend to be strictly selective in their choice of valency patterns to be described, while VPDs have a proclivity to be as exhaustive as possible in deciding what to include. Both types of dictionaries provide us with snap-shots of English in a particular age along with corpus evidence. However, the "pictures" taken by these dictionaries sometimes provide some clues to detect what is happening in a language. By comparing the descriptions in these dictionaries, phraseologists, in collaboration with lexicographers, can contribute to clarifying language change. This opens up the possibility of phraseology expanding beyond the territory it has concentrated on. In other words, phraseology can play a more important role in language research than Gries (2008) argues.

References

Dictionaries

BBI3, 2009: *The BBI Combinatory Dictionary of English. 3rd edition*. Benson, Morton et al. (eds.) Amsterdam/Philadelphia: John Benjamins.

CALD1, 2003: *Cambridge Advanced Learner's Dictionary 1st edition*. Cambridge: Cambridge Univ. Press.

CALD2, 2005: *Cambridge Advanced Learner's Dictionary 2nd edition*. Cambridge: Cambridge Univ. Press.

CALD3, 2008: *Cambridge Advanced Learner's Dictionary 3rd edition*. Cambridge: Cambridge Univ. Press.

Chambers Slang Dictionary, 2008: *Chambers Slang Dictionary.* Edinburgh: Chambers.

CIDE, 1995: *Cambridge International Dictionary of English*. Cambridge: Cambridge Univ. Press.

LDOCE5, 2009: *Longman Dictionary of Contemporary English 5th edition*. London: Longman.

MED2, 2007: *Macmillan English Dictionary for Advanced Learners. 2nd edition*. Oxford: Macmillan.

OALD8, 2010: *Oxford Advanced Learners Dictionary of English 8th edition*. Oxford: Oxford Univ. Press.

Oxford Learners' Thesaurus, 2008: *Oxford Learners' Thesaurus: A dictionary of synonyms*. Oxford: Oxford Univ. Press.

Patterns 1, 1996: *Collins COBUILD Grammar Patterns 1: Verbs*. Francis, Gill et al. (eds.). London: Harper Collins.

Patterns 2, 1998: *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. Francis, Gill et al. (eds.). London: Harper Collins.

VDE, 2004: *A Valency Dictionary of English: A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Herbst, Thomas et al. (eds.). Berlin: Mouton.

Online corpora

British National Corpus: < http://corpus.byu.edu/>. (Access 10. 8. 2012)

Corpus of Contemporary American English: < http://corpus.byu.edu/>. (Access 10. 8. 2012)

Corpus of Historical American English: <http://corpus.byu.edu/>. (Access 10. 8. 2012)

Other references

AARTS, Flor, 1999: Syntactic information in OALD5, LDOCE3, COUBUILD2, CIDE. Herbst, Thomas / Popp, Kerstin (eds.): *The Perfect Dictionaries?* Berlin/New York: Mouton. 15–32.

ALGEO, John, 2006: *British or American English?* Cambridge: Cambridge Univ. Press.

ATKINS, Sue B. T. / RUNDELL, Michael, 2008: *The Oxford Guide to Practical Lexicography*. Oxford: Oxford Univ. Press.

BARLOW, Michael, 1999: Usage, blends, and grammar. Barlow, Michael / Kemmer, Suzanne (eds.): *Usage Based Models of Language*. Stanford, California: CSLI.

BÉJOINT, Henri, 2010: *The Lexicography of English*. Oxford: Oxford Univ. Press.

FAULHABER, Susen, 2011: *Verb Valency Patterns*. Berlin/New York: Mouton.

FOLLETT, Willson, 1998: *Modern American Usage*. New York: Hill and Wang.

GIVÓN, Talmy, 1980: The binding hierarchy and the typology of complements. *Studies in Language* 4.3, 333–377.

GRANGER, Sylviane / PAQUOT, Magali, 2008a: From dictionary to phrasebook? Bernal, Elisenda / DeCesaries, Janet (eds.): *Proceedings of the XIII Euralex International Congress*. Barcelona: Universitat Pompeu Fabra. 1345–1355.

GRANGER, Sylviane / PAQUOT, Magali, 2008b: Disentangling the phraseological web. Granger, Sylviane / Meunier, Fanny (eds.): *Phraseology*. Amsterdam/Philadelphia: John Benjamins. 27–49.

GRIES, Stefan T., 2008: Phraseology and linguistic theory. Granger, Sylviane / Meunier, Fanny (eds.): *Phraseology*. Amsterdam/Philadelphia: John Benjamins. 3–25.

HANKS, Patrick, 2006: Lexicography: overview. Brown, Keith (ed.): *The Encyclopedia of Language and Linguistics*. 2nd edition. Oxford: Oxford University Press. 579–584.

HANKS, Patrick, 2008: Lexical patterns: From Hornby to Hunston and beyond. Bernal, Elisenda / DeCesaris, Janet (eds.): *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra. 89–129.

HERBST, Thomas, 2009: Valency – item-specificity and idiom principle. Römer, Uta / Schulze, Rainer (eds.): *Exploring the Lexis-Grammar Interface*. Amstermdams: John Benjamins. 49–68.

HERBST, Thomas / KLOTS, Michael, 2009: Syntagmatic and phraseological dictionaries. Cowie, A. P. (ed.): *The Oxford History of English Lexicography*. Vol. II. Oxford: Oxford Univ. Press. 219–244.

HORNBY, Albert S., 1975: *Guide to Patterns and Usage in English 2nd edition*. Oxford: Oxford Univ. Press.

HUNSTON, Susan, 2010: How can a corpus be used to explore patterns? O'Keeffe, Anne / McCarthy, Michael (eds.): *The Routledge Handbook of Corpus Linguistics*. London: Routledge. 152–166.

HUNSTON, Susan / FRANCIS, Gill, 2000: *Pattern Grammar*. Amsterdam/Philadelphia: John Benjamins.

IYEIRI, Yoko, 2011: *To Convince Someone To Do Something* in Present-Day American English Ohashi, Hiroshi et al. (eds.). *Kotoba to Kokoro no Tankyu*. Tokyo: Kaitakusha. 363–376.

MOON, Rosamund, 2008: Dictionaries and collocation. Granger, Sylviane / Meunier, Fanny (eds.): *Phraseology*. Amsterdam/Philadelphia: John Benjamins. 313–336.

QUIRK, Randolph / GREENBAUM, Sydney / LEECH, Geoffrey / SVARTVIK, Jan, 1985: *A Comprehensive Grammar of the English Language*. London: Longman.

SIEPMANN, Dirk, 2008: Phraseology in learners' dictionaries. Meunier, Fanny / Granger, Sylviane. (eds.): *Phraseology in Foreign Language Learning and Teaching*. Amsterdam/ Philadelphia: John Benjamins. 185–202.

SINCLAIR, John, 1991: *Corpus Concordance Collocation*. Oxford: Oxford Univ. Press.

SINCLAIR, John, 2008: Introduction. Granger, Sylviane / Meunier, Fanny (eds.): *Phraseology*. Amsterdam/Philadelphia: John Benjamins. xv–xxvii.

SUMIYOSHI, Makoto, 2009: A comparative study of pattern descriptions of *that*-taking verbs between English-Japanese learners' dictionaries and monolingual learners' dictionaries. *Proceedings of Asialex 2009 on CD-ROM*. Bangkok, Thailand: Asialex.

SUMIYOSHI, Makoto, 2011: *That*-taking predicates in English: A phraseological approach. Yagi, Katsumasa et al. (eds.). *Phraseology, Corpus Linguistics and Lexicography*. Hyogo: Kwansei Univ. Press. 167–184.

TURTON, Nigel, D. / HEATON, J. B., 1987: *Longman Dictionary of Common Errors*. London: Longman.

YAGI, Katsumasa, 1999: *Eigo no Bunpo to Goho*. Tokyo: Kenkyusha Publisher.

Un projet phraséographique: critères et choix

Claudia Maria Xatara (São Paulo)

Abstract

Every dictionary must come from a judicious lexicographic project preparation. The same principle applies to a phraseological dictionary, whose project will guide the editor's choices. We proposed an online dictionary of idioms where each entry is always in Portuguese of Brazil, with its definition, additional information, an illustrative example, indications of synonymy, if any, and its equivalents in Portuguese of Portugal and in the three variants of the French (France, Belgium and Canada). We will therefore approach all types of constraints with respect to the phraseographic choices we made, regarding: 1. nomenclature (sources, frequency, lemmatization); 2. definitions and polysemous items and homonyms; 3. information with respect to registers, expressiveness of use, origin and syntactic restrictions; 4. the examples; 5. synonymy relating to the entry; 6. references; 7. approximate equivalents; 8. the entries layout; 9. the construction of a software for the online dictionary.

1 Introduction

Nous avons élaboré un dictionnaire en ligne, *Dictionnaire d'expressions idiomatiques portugais du Brésil et du Portugal – français de France, de la Belgique et du Canada*, dans le cadre d'une recherche menée à l'UNESP, en collaboration avec d'autres collègues et des étudiants en Master et en Doctorat et aussi en collaboration internationale avec des chercheurs du Portugal, du Canada, de la France et de la Belgique (cette dernière collaboration encore en cours). Cependant, pour avoir élaboré ce dictionnaire, nous avons eu besoin a priori d'un projet phraséographique, qui ont donné le profil et les caractéristiques particulières à cet ouvrage. C'est ce que nous voyons à la suite.

2 Le projet phraséographique

2.1 Les grandes lignes

Tout d'abord nous avons défini l'objet de ce dictionnaire, c'est-à-dire, le type d'unité phraséologique dont nous irions nous occuper. Et nous avons choisi traiter exclusivement des *expressions idiomatiques* (EI). Deuxièmement nous nous sommes basés sur des études qui essayaient à leur tour de définir le concept d'EI (surtout Gross, 1993; Xatara, 1995; Mejri, 2002; Petit, 2003) et nous avons adoptés une définition bien délimitée, à savoir: *séquence polylexicale, figurée et figée par la tradition culturelle d'une communauté linguistique*.

Ensuite nous avons pensé au type de dictionnaire prétendu et avons décidé de proposer un dictionnaire bilingue qui élirait le paire de langues portugais-français dans quelques-unes de leurs variantes: le portugais du Brésil (PB), le portugais du Portugal (PP), le français de la France (FF), le français de la Belgique (FB) et le français du Canada (FC).

Puis nous avons déterminé, par rapport à la nature de ce dictionnaire, qu'il aurait l'approche sémasiologique, parce qu'il partirait des EI (les unités signifiantes) comme entrées et présenterait ensuite leurs sens, mais aussi l'approche analogique parce qu'il rassemblerait des expressions synonymiques.

Comme objectif, notre intention avec ce dictionnaire était de construire un ouvrage qui témoigne en décodage ou pour l'encodage, l'usage de ces unités phraséologiques, dont le figement est un phénomène scalaire, avec justement des éléments variables. Tel est qu'il s'agit à vrai dire d'un semi-figement.

Le format prévu a été l'électronique, ce qui permettra sa périodique extension et mise à jour, ainsi que plusieurs interfaces de recherche. Pour cela il serait nécessaire de créer, sous la base du logiciel XML, les balises pour rendre compte de toute la macrostructure et microstructure du dictionnaire, à l'aide des linguistes informaticiens. Ensuite ces fichiers seraient traités pour sa mise en ligne.

2.2 La nomenclature en PB et ses équivalents

Nous avons envisagé partir d'un recueil ellaboré depuis 1990 et y ajouter de nouveaux éléments. Les El en portugais ou en français ont alors été collectées de sources secondaires (terme proposé par Haensch et al. (1982), à la différence de sources primaires qui seraient des documents authentiques et des corpus): vers 30 dictionnaires de langue, monolingues ou bilingues.

Toutes les expressions dans les deux langues et leurs variantes devraient respecter un seuil minimum de *fréquence* d'après leur occurrence sur le *Web* utilisé comme corpus linguistique (Kilgarriff/Grefenstette, 2003). Le seuil pris en compte est d'une occurrence par million de mots mesurée par page trouvée sur le Web (Colson 2003, 2006, 2007).

En partant de ces principes et selon les données fournies par Grefenstette et Nioche (2000), Evans et al. (2004) et l'Union Latine (2007), pour qu'une EI soit considérée comme usuelle dans les langues en question, elle devrait satisfaire aux seuils suivants: 56 occurrences pour le PB, dans des site: br; 14 pour le PP, dans des site: pt; 157 pour le FF, dans des site: fr; et 07 pour le FC, dans des site: ca.

3 Les choix microstructuraux

3.1 Quant aux éléments concernant les définitions, les abréviations, les symboles et les renvois, nous avons déterminé le PB comme la métalangue de description.

3.2 La classe grammaticale de chaque entrée ne serait pas explicitée, mais les équivalents devraient suivre strictement la classe grammaticale de l'entrée. Alors, à une séquence verbale (SV) comme entrée en PB correspondent donc des équivalents SV dans les autres langues; à une séquence nominale (SN) sont associés des équivalents appartenant à la classe de SN; si une séquence est en fonction adjective (SAdj) en PB, ses équivalents appartiennent à la classe SAdj; et pour les entrées qui sont des séquences adverbiales (SAdv), les équivalents PP, FF et FC appartiennent à cette même classe. Par exemple, l'entrée dont le noyau est agulha no palheiro, un SN, devra correspondre en français au SN *aiguille dans une botte de foin*, et non à des SV comme *être (comme)*, *(re)chercher*, *paraître*, *sembler*, etc + *une aiguille dans une botte de foin*.

3.3 A chaque EI en PB nous devrions proposer une définition paraphrastique, simple et courte. Dans le cas des entrées polysémiques, chaque définition ou sens serait numéroté. Par exemple, *abrir o (seu) coração* est une entrée polysémique, avec 2 sens: 1. faire confiance à ses sentiments et parler franchement, sincèrement = *ouvrir son coeur*, *parler à coeur ouvert*, *vider son coeur*; et 2. être receptif, sensible aux bons sentiments = *ouvrir son coeur.*

3.4 Sous la définition, nous ajouterions des informations complémentaires – des remarques qui nous paraissent utiles pour l'usager du dictionnaire, soit une marque de niveau de langue ou de valeur de signification, soit une observation par rapport à l'origine ou à la motivation créatrice de l'expression, soit encore quelque contrainte syntaxique. Alors nous signalerons:

- Les niveaux de langues: nous considérerons les niveau *familier* (qui implique une proximité entre locuteurs les autorisant à s'affranchir de certaines règles de contrôle dans l'interlocution) et *standard* comme spécifiques des EI. A ce titre, ils ne seront pas renseignés dans les entrées. En revanche, les registres *soutenu* (recherche d'un niveau de pureté, d'élégance) et *vulgaire* (emploi de termes grossiers ou présentant des connotations grossières) seront indiqués, ainsi que *très familier* (intermédiaire entre familier et vulgaire). De la même façon chaque équivalent recevra une indication de niveau quand celui-ci est différent du niveau de l'entrée.

  Ex.: *andar com a cabeça na lua = avoir la tête dans le cul* [vulg.]

- Les marques de valeur: les acceptions *intensives* (sens dont les effets sont plus forts), *ironiques* (sens qui procèdent du dénigrement, en utilisant souvent l'antiphrase), *euphémistiques* (sens qui présentent une certaine minoration), *mélioratives* (sens qui impliquent une perception favorable ou laudative) et *péjoratives* (sens qui impliquent, au contraire, une perception défavorable, dénigrante ou dépréciative) seront indiquées notamment parce qu'elles sont pertinentes à la fonction d'encodage ou de production textuelle que ce dictionnaire veut aussi assurer.

  Ex.: *a preço de banana* [pej.] = *pour une bouchée de pain*

- L'origine: nous avons remarqué que les origines hypothétiques ou supposées abondent dans l'univers phraséologique, pour cela nous devrons restreindre l'explicitation aux origines supposées les plus fiables (surtout quand il y a un rapport historique auquel l'EI s'attache) ou en ce qui concerne les domaines qui a motivé la création de l'expression (comme Biologie, Histoire, Religion, etc.).

  Ex.: *ano de vacas magras* [orig.: Bíblia; vem do Antigo Testamento]

- Les contraintes syntaxiques: les restrictions de sujet ou compléments ou de combinatoires verbales, pour les SN, SAdj ou SAdv, composeront aussi un type de données insérées dans les informations complémentaires.

  Ex.: *não caber em si* [intens.; refere-se ao ato de estar em um estado de felicidade tão extremo que a pessoa sente como se estivesse transcendendo; sujeito da expressão: pessoa]

3.5 Les exemples seraient tous extraits du Web avec indication de l'adresse et date de consultation. Alors, pour illustrer l'usage de l'entrée *à beira do abismo*, on prendra des contextes comme:

> */…/ pior é que o Governo entregou todo o nosso patrimônio público, cortou os gastos sociais, seguiu à risca a cartilha do FMI e o Brasil está* ***a beira do abismo…***
> (mst.org.br/informativos/JST/221/editorial.html, 02/04/04)

De la même façon, pour illustrer l'usage de son équivalent en français de France, *au bord de l'abîme*, c'est le Web qui nous fournira le contexte:

> *Quand on se voit* ***au bord de l'abîme*** *et qu'il semble que Dieu vous ait abandonné, on n'hésite plus à attendre de lui un miracle.*
> (mots-auteurs.fr/mot.php?mot=ab%EEme, 18/02/01)

3.6 Par rapport à la synonymie, si l'EI de l'entrée avait une ou plusieurs autres expressions de sens similaires, celles-ci seraient indiquées par ordre alphabétique (par exemple: *acertar na mosca* et *acertar no alvo* sont des synonymes de l'entrée *acertar em cheio*).

3.7 Pour les abréviations, nous devrions les utiliser afin d'indiquer:

– le niveau très familier ou le niveau vulgaire de langage (le niveau *culto* (cultivé) figurerait dans le dictionnaire, mais pas abrégé);
– le sens euphémique, ironique, mélioratif ou péjoratif;
– l'origine (en général l'indication d'un domaine) ou l'origine supposée (en général indication d'une croyance populaire).

Alors nous auront:

coloq. dist. (*coloquial distenso*) = niveau très familier
euf. (*eufêmico*) = sens euphémique
iron. (*irônico*) = sens ironique
mel. (*melhorativo*) = sens mélioratif
orig. (*origem*) = origine (en général l'indication d'un domaine)
orig. sup. (*origem suposta*) = origine supposée (en général indication d'une croyance populaire)
pej. (*pejorativo*) = sens péjoratif
vulg. (*vulgar*) = niveau vulgaire de langage

3.8 Quelques symboles ont été aussi prévus comme:

– les parenthèses pour indiquer un emploi facultatif

Ex.: *a dar com* (*o*, *um*) *pau* – un mot qui peut être introduit
dar mole (*pro azar*) – plus d'une possibilité d'emploi – en général des variantes au bord droit de l'expression

– le [se] pour accompagner un verbe pronominal

Ex.: *serrer [se] la ceinture*

– les crochets pour les explications ou informations complémentaires

Ex.: pour l'entrée *amigo da onça*, il suit les informations [pej.; referência a esse animal, considerado selvagem e perigoso];

– et le vide mathématique (Ø) pour indiquer l'inexistence d'équivalent aussi idiomatique. Dans ce cas, quand nous n'aurons pas trouvé une EI équivalente, nous devrons donner une suggestion de traduction non-idiomatique.

Ex.: *bom de bico = Ø charognard, profiteur*

3.9 Pour les renvois, nous avons pensé aux cas des synonymes dont les équivalents figureraient dans l'EI qui commencerait par une lettre antérieure à celle de l'entrée:

(a) Sin. e equiv. em = Synonyme et équivalent dans (indication d'un seul synonyme qui commence par une lettre antérieure à celle de l'entrée)

Ex.: *a preço de ouro* → Synonymes et équivalents dans *a peso de ouro*

(b) Sin. = Synonyme (indication d'un ou plusieurs synonymes par ordre alphabétique)

Equiv. em = Équivalent en langue étrangère dans (indication du synonyme qui commence par une lettre antérieure à celle de l'entrée et dont figurent tous les équivalents)

Ex.: *acertar na mosca* → Synonymes *acertar em cheio, acertar no alvo*
→ Équivalents dans *acertar em cheio*

Ainsi, l'usager trouvera les équivalents dans l'entrée *acertar em cheio*, à savoir: *être (mettre*, *taper*, *tirer) dans le mille*, *faire mouche*, *mettre en plein dedans*, *toucher sa cible.*

3.10 Après avoir décidé, au niveau du projet phraséographique, tous les éléments exposés antérieurement, nous sommes enfin arrivés à proposer le plan des articles. Leur structure présenterait l'entrée, toujours en PB, avec définition, informations complémentaires, exemples, indication de synonymie, et ses équivalents en PP, FF, FB et FC, comme le tableau ci-dessous:

<table>
<tr><td colspan="4">**EI en PB :** paraphrase définitionnelle en PB
[informations complémentaires]
Exemple = contexte extrait du Web
(source, date)
Sin.: **synonymes en PB** (quand il y en a)
Equiv. dans : **EI synonyme en PB** (= 1ère expression par ordre alphabétique : renvoi)</td></tr>
<tr><td>**EI en PP** = équiv. à l’entrée du PB **:** *Contexte*
(source, date)</td><td>**EI en FF** = équiv. à l’entrée du PB **:** *Contexte*
(source, date)</td><td>**EI en FB:** équiv. à l’entrée du PB **:** *Contexte*
(source, date)</td><td>**EI en FC** = équiv. à l’entrée du PB **:** *Contexte*
(source, date)</td></tr>
</table>

A titre d’illustration, l’article *armar o (um) barraco* (EI verbale), *balaio de gatos* (EI nominale) et *na maciota* (EI adverbiale):

<table>
<tr><td colspan="4">**armar o (um) barraco:** fazer um escândalo
[referência à forma irregular como são construídos os barracos nas favelas]
*A garota entrou em pânico e começou a **armar o barraco**, dizendo que pagou caríssimo pelo abadá e que ela tinha o direito de subir aonde a mesma desejasse.*
(jcorreio.com.br/ver_noticia.asp?CodNoticia=1762&Secao=, 22/05/04)
Sin.: **armar o circo**</td></tr>
<tr><td>**armar a (uma) barraca:** *Acho que vai ter de ser à moda antiga… **armar a barraca** e não sair de lá enquanto não me derem respostas!*
(mazdapt.com/forum/f124/consultorio-tecnico-13493/page12.html, 23/04/10)

fazer uma peixeirada: *O único objectivo da entrevista era tirar o homem do sério e **fazer uma peixeirada** que eles sabem que as pessoas adoram ver e aumentar assim as audiências.*
(tdtonline.org/viewtopic.php?t=4186&f=10, 23/04/10)</td><td>**crier comme un putois** [coloq. dist.]**:** *D’ailleurs j’arrête pas de **crier comme un putois** et je vois bien que ça la rend malheureuse /…/*
(forum.doctissimo.fr/…/nocturne-epuisant-cododo-sujet_17184_1.htm, 02/05/11)

faire du chambard: *Bien sûr l’ex est revenu à la charge plusieurs fois pour **faire du ‘‘chambard’’** mais il s’est bien vite essoufflé et s’est très vite consolé.*
(forum.femmeactuelle.fr/PostsList.aspx?ForumPostID, 23/02/11)

faire du foin: *Il faut dire que **nous avions fait du foin** dans le refuge et que le quelqu’un qui avait fait le coup devait se sentir passablement morveux.*
(actuados.fr/article,1203,0.html, 06/10/05)

gueuler comme un putois: *Par contre c’est affolant ces gosses, il fut **gueuler comme un putois** pour ce faire entendre /…/*
(jumeaux-et-plus.fr/component/option…/wap2,wap2, 02/05/11)</td><td>**faire du chambard:** *Je pense qu’il a bien vite remarqué que nous n’étions pas venus pour **faire du chambard**, mais bien pour «marquer le coup» et nous rappeler à son bon souvenir suite à sa visite d’hier sur «notre» chantier …*
(ruehaute.skynetblogs.be/voeux/, 19/01/12)

faire du foin: *La déception était personnelle mais je ne suis pas du style à **faire du foin** et à polémiquer.*
(lavenir.net/article/detail.aspx?articleid=39085038, 19/01/12)

gueuler comme un putois: *La colère monte en moi et commence à **gueuler comme un putois**, je crois l’instant d’une seconde que je vais changer de sport: lâcher la course pour enfiler des gants de boxe.*
(bestofverviers.be/index.php?option=com_content&view=article&id=548:olivier-beaumont-nous-raconte-sa-course-a-pied-son-ultra-trail-du-mont-blanc&catid=23:sportives&Itemid=32, 19/01/12)</td><td>**brasser de la marde:** *Je ne réponds plus aux journalistes, surtout quand c’est dans le but de **brasser de la marde** pour aucune raison.*
(mikeward.ca/blogue/?m=wspvpicajsnydiiu&paged=12, 23/04/10)

faire de la marde: *Lui, là, je l’ai démardé trois ou quatre fois. Pas capable de le placer: il se brûle partout. Partout où il va, il **fait de la marde**.*
(m.ledevoir.com/economie/emploi/284741/le-rambo-de-la-ftq-construction-se-defend, 23/04/10)</td></tr>
</table>

<table>
<tr><td colspan="4"><b>balaio de gatos:</b> confusão, desordem
[pej.; alusão à imagem denotativa dos movimentos desatinados que vários gatos fariam num balaio]

<i>Nas décadas seguintes, a new age ganhou tantas divisões que acabou virando <b>um balaio de gatos</b>.</i>
(veja.abril.com.br/090102/p_106.html, 08/03/11)

Sin.: <b>saco de gatos</b></td></tr>
<tr>
<td><b>balaio de gatos:</b> A oposicao é um <b>balaio de gatos</b>! Por isso o AJJ e sua turma se perpetuaram na Madeira! (dnoticias.pt/comment/reply/248554/69658, 08/03/11)

<b>saco de gatos:</b> Dividido em grupos e grupinhos, o PSD não é mais do que <b>um saco de gatos</b> onde já ninguém se entende. (cmjornal.xl.pt/detalhe/noticias/…/um-saco-de-gatos, 08/03/11)</td>
<td><b>sac d'embrouilles:</b> <i>La vie de Kevin devient vite un dangereux <b>sac d'embrouilles</b>.</i> (gallimard.fr/gallimard-cgi/AppliV1/affied.pl?ouvrage=210520012177200, 07/04/05)

<b>sac de noeuds:</b> <i>Enfin, la candidature turque ajoute son propre noeud de difficultés au <b>sac de nœuds</b> européen.</i> (epoint.fr/edito/document.html?did=145262, 21/04/05)</td>
<td><b>sac d'embrouilles:</b> <i>Les quatre experts chargés de vider le <b>sac d'embrouilles</b> sont clairs: il sera difficile d'auditionner les magistrats, sous peine d'entacher de nullité les procédures judiciaires en cours de l'Affaire Fortis.</i> (richard3.com/2009/02/les_gaietes_du_parlement.html, 07/02/12)

<b>sac de noeuds:</b> <i>Chacun retrouvera dans ces personnages des traits de personnes connues peu ou prou, et appréciera, à la mesure de l'écho qu'elle fait chez lui, la résolution cathartique de ce <b>sac de noeuds</b> familial.</i> (martinrou.be/site/index.php?node_id=742, 07/02/12)</td>
<td><b>sac de noeuds:</b> <i>Si nous commençons à nous emmêler nous-mêmes dans les fils de nos festivals de cinéma, imaginez le <b>sac de noeuds</b> qu'y voient nos amis étrangers. (moncinema.cyberpresse.ca '…'</i> Chroniqueurs, <i>08/03/11</i>)</td>
</tr>
</table>

<table>
<tr><td colspan="4"><b>na maciota:</b> tranquilamente, sem muito esforço ou dificuldades
[essa expressão vem da denotação popular de «maciota»: folga, repouso, suavidade]

<i>Claro que é mais fácil conseguir as coisas <b>na maciota</b>.</i>
(diatribe.com.br/thule/hs06.shtml, 01/11/04)</td></tr>
<tr>
<td><b>nas calmas:</b> <i>Sporting joga nas calmas, apresenta um futebol sereno (coisa rara por aqueles lados) e ganha o jogo <b>nas calmas</b>, e sem casos.</i> (ruimoura.net/blog/2009/01/12/futebol-em-grande/, 02/03/10)</td>
<td><b>en (père) peinard:</b> <i>On espère que la qualité du travail déjà accompli et celle des produits réalisés l'aideront à passer le cap <b>en père peinard</b>.</i> (manoamano.free.fr/visages/inde/inde_nappe_ambiance.htm, 14/07/05)

<b>sans s'en faire:</b> <i>On s'ballade <b>sans s'en faire</b>. On respire l'air de la mer. /…/</i> (animezvous.fr/parole/…/mouette-et-le-chat-la – França, 13/05/11)</td>
<td><b>sans s'en faire:</b> <i>Comment jardiner <b>sans s'en faire</b>, planter en bonne connaissance des plantes et s'offrir un petit paradis fleuri sans y passer tout son temps?</i> (editions-tondeur.be/extraitarticle.php?id=568, 02/06/12)</td>
<td><b>en pépère:</b> <i>Sur 96 km, j'ai hâte que ça éclate; que le peloton se rebelle… et qu'on se repose un peu à rouler <b>en pépère</b> aussi.</i> (velocia.ca/forums/le…/9648-macho-le-velo-4.html, 29/11/10)</td>
</tr>
</table>

4 Conclusion

Inédit pour présenter les EI de ces deux langues et variantes et inédit par la structuration visuelle proposée, l'ouvrage décrit pourra répondre aux attentes d'apprenants, professeurs, traducteurs et chercheurs intéressés par l'usage des expressions et leurs équivalents en portugais et en français. Nous croyons aussi que sa mise en ligne favorisera d'une façon plus efficace sa divulgation au service de l'ensemble de la communauté francophone et lusophone. Mais la consistance et cohérence scientifique de ce dictionnaire a découlé des choix préalables, pris à partir des critères bien fondés dans son projet phraséographique.

Bref, avant de mettre en place l'élaboration de n'importe quel dictionnaire phraséologique, nous trouvons indispensable d'établir le projet en décidant: objet, type, nature, objectif, format, nomenclature et éléments de la microstructure (ceux-ci en détails) de ce dictionnaire.

Références Bibliographiques

CAMACHO, Beatriz, 2008: *Estudo comparativo de expressões idiomáticas do português do Brasil e de Portugal e do francês da França e do Canadá*. Dissertação (Mestrado em Linguística). São José do Rio Preto: IBILCE, Universidade Estadual Paulista.

COLSON, Jean-Pierre, 2003: Corpus linguistics and phraseological statistics: a few hypotheses and examples. Burger, H./Hächi Buhofer, A./Gréciano, G. (eds.). *Flut von Texten – Vielfalt der Kulturen*. Ascona 2001 zu Methodologie und Kulturspezifik der Phraseologie. Baltmannsweiler: Schneider Verlag Hohengehren. 47–59.

COLSON, Jean-Pierre, 2007: The World Wide Web as a corpus for set phrases. Burger, H. et al. (eds.). *Phraseologie / Phraseology.* Berlin/New York: de Gruyter, 1071–1077.

COLSON, Jean-Pierre, 2006: Towards computational phraseology. The project of an idiom concordancer. Häcki-Buhofer, Annelies/Burger, Harald (Hrsg.): *Phraseology in Motion* 1. *Methoden und Kritik*. Baltmannsweiler: Schneider Verlag Hohengehren. 21–32.

EVANS, D. et al. *The multi-lingual web*: <http://www.infonortics.com/searchengines/sh04/04/pro.html>. Access 01. 01. 2004.

GREFENSTETTE, Gregory/NIOCHE, Julien, 2000: Estimation of English and non-English language use on the www. *Proceedings of RIAO'2000, Content-Based Multimedia Information Access, Paris, April 12–14, 2000*. 237–246

GROSS, Maurice, 1993: Les phrases figées en français. *L'information grammaticale* 59, 36–41.

HAENSCH, Günther et al., 1982: *La Lexicografia: de la Lingüística teórica a la Lexicografia práctica*. Madrid: Gredos.

KILGARRIFF, Adam / GREFENSTETTE, Gregory, 2003: Introduction to the special issue on the Web as Corpus. *Computational Linguistics* 29/3, 333–347.

MEJRI, Salah, 2002: Le figement lexical: nouvelles tendances. *Cahiers de lexicologie* 80, 213–225.

PETIT, Gérard, 2003: Figement lexical et lemmatisation: les locutions de type SV. *Cahiers de Lexicologie* 82, 1.

RIVA, Huélinton Cassiano, 2009: *Dicionário onomasiológico de expressões idiomáticas usuais na língua portuguesa do Brasil*. São José do Rio Preto. Tese (Doutorado em Linguística). São José do Rio Preto: IBILCE, Universidade Estadual Paulista.

UNION LATINE. *La place du français dans l'internet*: <http://dtil.unilat.org/LI/2002/fr/index.htm>. Acess 01. 05. 2007.

XATARA, Claudia, 1995: O resgate das expressões idiomáticas. *Alfa: Revista de Linguística* 38, 195–210.

Autoren / Authors

Torben Arboe
Aarhus University
Peter Skautrup Centre of Jutlandic
Dialect Research
Jens Chr. Skous Vej 4, bygning 1483, 6
8000 Aarhus C, Denmark
jysta@hum.au.dk

Elena Berthemet
Université de Bretagne Occidentale
20 rue Duquesne
29200 Brest, France
elena.berthemet@univ-brest.fr

Jean-Pierre Colson
Université catholique de Louvain
Faculté de Philosophie, Arts et Lettres
Place Blaise Pascal 1, bte L3.03.33
1348 Louvain-la-Neuve, Belgium
jean-pierre.colson@uclouvain.be

Cosimo De Giovanni
University of Cagliari
Foreign Languages
and Literatures faculty
Via San Giorgio 14, Campus Aresu
09128 Cagliari, Italy
cosimodegiovanni@gmail.com

Marcel Dräger
Universität Basel
Deutsches Seminar
Am Nadelberg 4
4051 Basel, Schweiz
marcel.draeger@unibas.ch

Claire Ducarme
Université de Liège
Linguistique du français
et dialectologie wallonne
Place Cockerill, 3-5/40
4000 Liège, Belgium
claire.ducarme@ulg.ac.be

René Frauchiger
Universität Basel
Deutsches Seminar
Am Nadelberg 4
4051 Basel, Schweiz

Peter Grzybek
Universität Graz
Institut für Slawistik
Merangasse 70/I
8010 Graz, Österreich
peter.grzybek@uni-graz.at

Milena Hnátková
Charles University Prague
Faculty of Philosophy,
Institute of Theoretical and
Computational Linguistics
Nam. J. Palacha 2
116 38 Praha 1, Czech republic
milena.hnatkova@ff.cuni.cz

Ai Inoue
National Defense Academy of Japan
1-10-20 Hashirimizu
Yokosuka City, Kanagawa Prefecture,
Japan
narudo24@hotmail.com

Emmerich Kelih
Universität Wien
Institut für Slawistik
Spitalgasse 2, Hof 3
1090 Wien, Österreich
emmerich.kelih@univie.ac.at

Marie Kopřivová
Charles University Prague
Faculty of Philosophy,
Institute of The Czech National Corpus
Nam. J. Palacha 2
116 38 Praha 1, Czech republic
marie.koprivova@ff.cuni.cz

Nataša Kralj
Srednja elektro-računalniška
šola Maribor
Smetanova ul. 6
2000 Maribor, Slovenija
natasa.kralj@guest.arnes.si

Marlène Linsmayer
Universität Basel
Deutsches Seminar
Am Nadelberg 4
4051 Basel, Schweiz

Claudia Lückert, geb. **Aurich**
Westfälische Wilhelms-Universität
Münster
Englisches Seminar
Johannisstr. 12-20
48143 Münster, Deutschland
claudia.aurich@uni-muenster.de

Jasmina Markič
Univerza v Ljubljani
Filozofska fakulteta
Aškerčeva 2
1000 Ljubljana, Slovenija
jasmina.markic@ff.uni-lj.si

Matej Meterc
Ulica Cirila Tavčarja 8
4270 Jesenice, Slovenija
matej.meterc@gmail.com

Vesna Mikolič
Univerza na Primorskem
Fakulteta za humanistične študije,
Znanstveno-raziskovalno središče
Titov trg 5
6000 Koper, Slovenija
vesna.mikolic@fhs.upr.si

Piotr Pęzik
University of Łódź
Narutowicza 65
90-131 Łódź, Poland
piotr.pezik@gmail.com

Liezl Potgieter
PO Box 367
Paarl, 7620, South Africa
liezlpotgieter@gmail.com
liezlp@vodamail.co.za

Makoto Sumiyoshi
Setsunan University
17-8, Ikeda-nakamachi,
Neyagawa-shi
Osaka, 572-8508, Japan
sumiyosi@ilc.setsunan.ac.jp

Darinka Verdonik
Univerza v Mariboru
Fakulteta za elektrotehniko,
računalništvo in informatiko
Smetanova ul. 17
2000 Maribor, Slovenija
darinka.verdonik@um.si

Alessandra Widmer
Universität Basel
Deutsches Seminar
Am Nadelberg 4
4051 Basel, Schweiz

Claudia Maria Xatara
UNESP, Université
de l'État de São Paulo
R. Cristóvão Colombo, 2265
15054-000 São José
do Rio Preto, Brazil
xatara@sjrp.unesp.br

Phraseology *in* the dictionary, phraseology *in* the corpus – this sounds less complex than it is: upon closer examination, we are faced with a multi-faceted matter: not only can we understand 'phraseology' as a scientific discipline (i. e., in terms of phraseological research), but we can also think of it as an object of study (i. e., the treasure of phraseological units occurring in clearly defined linguistic material or even in a language as a whole). After all, phraseology is not contained per se in a dictionary or a corpus, and thus "simply placed", at our disposal for other purposes – starting from general interests for private use, through instructional and educational purposes, right up to phraseological research. Rather, the dictionary and corpus (or, more correctly: different kinds of dictionaries and corpora) are different in this and other respects and, additionally, stand in multiple complex interrelations.

Presentations based on these contributions were held at the EUROPHRAS conference (a traditional biannual conference organized by the European Society of Phraseology – EUROPHRAS) hosted by the University of Maribor between the 27th and the 31th of August 2012.

9 789616 930055

Cena: 22 EUR